DCA: Dividing and Conquering Amnesia in Incremental Object Detection

Aoting Zhang[1,3], Dongbao Yang[1,3*], Chang Liu[5], Xiaopeng Hong[4*], Miao Shang[4], Yu Zhou[2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[3]School of Cyber Security, University of Chinese Academy of Sciences
[4]Harbin Institute of Technology
[5]Tsinghua University
{zhangaoting, yangdongbao}@iie.ac.cn, liuchang2022@tsinghua.edu.cn
hongxiaopeng@ieee.org, miaos0522@gmail.com, yzhou@nankai.edu.cn

Abstract

Incremental object detection (IOD) aims to cultivate an object detector that can continuously localize and recognize novel classes while preserving its performance on previous classes. Existing methods achieve certain success by improving knowledge distillation and exemplar replay for transformer-based detection frameworks, but the intrinsic forgetting mechanisms remain underexplored. In this paper, we dive into the cause of forgetting and discover forgetting imbalance between localization and recognition in transformer-based IOD, which means that localization is less-forgetting and can generalize to future classes, whereas catastrophic forgetting occurs primarily on recognition. Based on these insights, we propose a Divide-and-Conquer Amnesia (DCA) strategy, which redesigns the transformer-based IOD into a localization-then-recognition process. DCA can well maintain and transfer the localization ability, leaving decoupled fragile recognition to be specially conquered. To reduce feature drift in recognition, we leverage semantic knowledge encoded in pre-trained language models to anchor class representations within a unified feature space across incremental tasks. This involves designing a duplex classifier fusion and embedding class semantic features into the recognition decoding process in the form of queries. Extensive experiments validate that our approach achieves state-of-the-art performance, especially for long-term incremental scenarios. For example, under the four-step setting on MS-COCO, our DCA strategy significantly improves the final AP by 6.9%.

Introduction

Object detection has experienced great advancements in recent years (He et al. 2017; Carion et al. 2020). Nevertheless, prevailing models rely on fixed data, rendering them inadequate to adapt to dynamic data in real-world scenarios. To equip object detection with the ability of lifelong learning and knowledge integration like humans, incremental object detection (IOD) is proposed, which can continuously localize and recognize object instances of new concepts while maintaining old knowledge. It is a pivotal stride toward attaining artificial general intelligence.

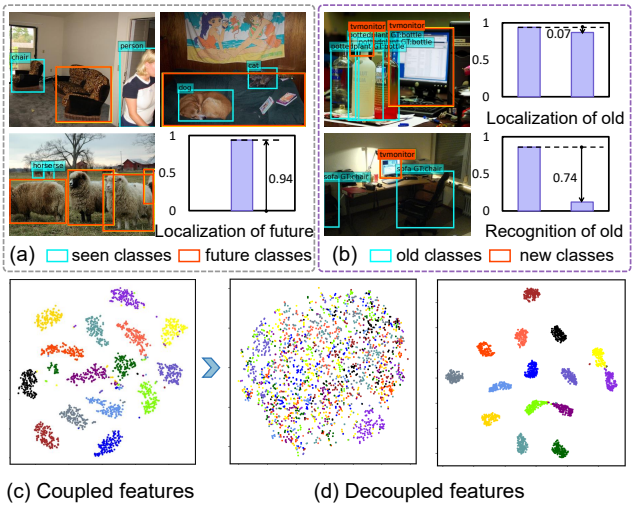
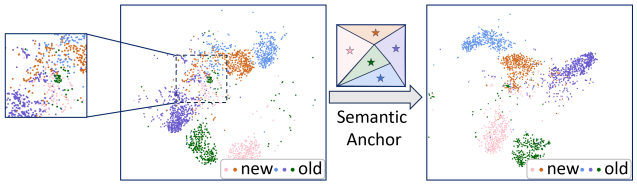


Figure 1: Forgetting imbalance within DETR-based IOD. (a) Localization is class-agnostic and can generalize predicted boxes on future classes with a high recall of 94%. (b) After fine-tuning on new data, localization is less-forgetting with recall of old classes slightly dropping from 94% to 87%, while average accuracy of recognition drops from 86% to 12%. (c) In original DETR, localization features are clearly influenced by recognition features and become category-specific. (d) After decoupling features, localization is obviously class-agnostic and we can focus on solving catastrophic forgetting on recognition.

IOD suffers from catastrophic forgetting (Kirkpatrick et al. 2017), where models tend to erase prior knowledge when fine-tuning on new data. Due to privacy and security concerns, limited access to historical data makes training from scratch difficult. To address this, most methods preserve old knowledge through exemplar replay (ER) (Liu et al. 2020b; Joseph et al. 2021a) and knowledge distillation (KD) (Zhou et al. 2020; Yang et al. 2022a), which adopt CNN-based detectors containing numerous hand-crafted components as the basic framework (Ren et al. 2015).

Within transformer-based incremental detection (DETR) frameworks, established strategies like KD and ER are popu-

*Corresponding Author.



(a) w/o semantic (b) w semantic

Figure 2: (a) Feature ambiguity arises between new and old tasks due to feature drift. (b) With semantic that guides the optimization direction of each class, clear boundaries are effectively maintained.

larly used. For example, ACF (Kang et al. 2023) distills distance matrix and interactive features to attain class consistency. This inertial reuse has limited DETR-based IOD techniques to their full potential. *Could we identify the underlying causes of catastrophic forgetting and suit the remedy to the case?* To answer this, we analyze *forgetting* of classification and localization in DETR-based IOD, two primary tasks in object detection. We find that localization in DETR-based IOD is class-agnostic and less-forgetting while recognition suffers from severe forgetting, which we call the *forgetting imbalance* between localization and recognition of DETR-based IOD. As shown in Figure 1(a), although the model is trained exclusively on seen categories (*chair, person, horse, dog, cat*), it is capable of generating accurate localization for future categories (*sofa, sheep, pottedplant*) with a high recall of 94%. After finetuning on new data (Figure 1(b)), localization still provides accurate boxes for old categories while recognition has serious forgetting, such as *bottle* being mistakenly identified as *pottedplant*. The *localization* recall of old classes drops slightly from 94% to 87% while average accuracy of *recognition* drop from 86% to 12%.

To address *forgetting imbalance*, we break away from the mutual entanglement of forgetting and provide a reliable method for DETR-based IOD. We propose a Divide-and-Conquer Amnesia (DCA) strategy which decouples incremental detection into less-forgetting localization and fragile recognition. DCA first decodes class-agnostic location information. The recognition decoding process further determines the specific category of an object based on the predicted position. Figure 1(c&d) compare localization features and recognition features in original DETR and after decoupling features, which illustrates that our "dividing" strategy frees the class-agnostic localization features from coupled features, and we can focus on solving recognition forgetting.

In the incremental recognition process, due to limited data access to past and future tasks, the semantic objectives for each task are optimized independently, focusing solely on the current task. The features of old classes experience significant drift or are completely overwritten, leading to feature ambiguity between new and old tasks, as shown in Figure 2(a), which impedes the model's compatibility with past, present, and future learning. Class semantic features from pre-trained language models (PLMs) anchor class representations, which constitute the semantic space, providing a unified optimization direction across tasks. Moreover, semantic understanding enables the model to comprehend relationships between new and old classes, promoting knowledge transfer from old to new tasks. We develop two mechanisms to embed semantic information. The first is duplex classifier fusion, which newly introduces a semantic classification head to calculate the similarities between recognition features and derived semantics.The second is query-role embedding, where semantic features of known tasks are incorporated into the recognition decoding process in the form of queries, implicitly introducing inter-class relationships through the attention mechanism.

To sum up, our contributions are as follows:

- We discover *forgetting imbalance* in DETR-based IOD and further propose a Divide-and-Conquer Amnesia (DCA) strategy, which decouples incremental object detection into less-forgetting localization and class recognition to mitigate mutual interference.
- To conquer severe forgetting in recognition, we effectively embed semantic knowledge from pre-trained language models to promote unified optimization across tasks by designing duplex classifier fusion and integrating semantic features into the recognition decoder in the form of queries, reducing feature shift in recognition.
- Extensive experiments on two datasets demonstrate our proposed DCA achieves state-of-the-art performance compared to other exemplar-free methods and is particularly suitable for long-term incremental scenarios.

Related Works

Incremental Learning. Various approaches to mitigate catastrophic forgetting can be broadly divided into three categories: i) Rehearsal-based methods (Rebuffi et al. 2017; Huang et al. 2024a; Zhu et al. 2025) involve storing a limited subset of old exemplars in memory buffers or utilizing a supplementary generator to synthesize pseudo samples for old data, which are then incorporated into training along with new data. ii) Regularization-based methods (Hou et al. 2019; Douillard et al. 2020; Huang et al. 2024b) endeavor to design a loss function that penalizes changes in pivotal parameters during learning new tasks or utilize knowledge distillation to retain effective information acquired by previous models, such as output logits and intermediate features. iii) Architecture-based methods (Yan, Xie, and He 2021; Wang et al. 2022; Douillard et al. 2022) modify the network architecture by adding sub-networks or experts when new tasks arrive while maintaining the previous network frozen.

Incremental Object Detection. Previous methods focus on distilling knowledge such as outputs (Shmelkov, Schmid, and Alahari 2017), feature maps (Hao et al. 2019) and various relations (Peng et al. 2021; Yang et al. 2022b), or replaying exemplars and intermediate features (Acharya, Hayes, and Kanan 2020; Joseph et al. 2021b; Yang et al. 2023a; Joseph et al. 2021a; Yang et al. 2023b) to mitigate forgetting. Moreover, IncDet (Liu et al. 2020a) is based on parameter isolation, which mines and freezes important parameters. Built upon the DETR-based detection framework, CL-DETR distills knowledge from the most informative predictions of the old model and preserves the label distribution

of the training set. ACF distills the distance matrix to preserve the inter-class discrimination and interactive features to maintain intra-class consistency. In this work, we uncover the causes of catastrophic forgetting in DETR-based IOD and propose a divide-and-conquer strategy to focus on the forgetting of class recognition. Moreover, our method has the advantage of exemplar-free overhead.

Incremental Learning with Foundation Models. Recent trends in continual learning involve integrating pre-trained vision transformers with parameter-efficient fine-tuning techniques to adapt the model to downstream tasks. These techniques include prompt tuning (Jia et al. 2022), adapters (Chen et al. 2022), LoRA (Hu et al. 2022), etc. The core is to build additional learnable parameters or modules to guide the pre-trained representation and select appropriate prompts during reference. Another work leverages pre-trained vision-language models (Radford et al. 2021) as foundation models (Wang et al. 2023; Smith et al. 2023; Yang et al. 2024), which can learn complex patterns between images and language. These methods learn lightweight adapter modules for both textual and visual pathways (Gao et al. 2024) or soft prompts that serve as inputs to frozen visual and text encoders (Zhou et al. 2022) to capture task-specific information. However, directly integrating foundation models with DETR-based IOD incurs a lot of storage and computational overhead. In this work, we explore the rich inherent semantics mediated by pre-trained language models (PLMs) (either unimodal (Kenton and Toutanova 2019) or multimodal (Radford et al. 2021)) to alleviate forgetting, which flexibly bridges IOD and foundation models.

Preliminary

Problem Definition. The objective of IOD is to develop a unified detector that can adapt to newly encountered classes. Formally, given the dataset $\mathcal{D}=\{(x,y)\}$, where x is the image with corresponding object annotations y, the total category set of objects is $\mathcal{C}$. We partition the category set $\mathcal{C}$ into T subsets, with $\mathcal{C}=\mathcal{C}_1\cup\cdots\cup\mathcal{C}_T$, and corresponding sub-datasets constitute the training set for each phase. All label sets are mutually exclusive. At phase t, a group of classes $\mathcal{C}_t$ are exposed to the detector. For an image in the current phase, it may contain multiple objects from new classes $\mathcal{C}_t$ and other classes (old $\mathcal{C}_{1:t-1}$ and future $\mathcal{C}_{t+1:T}$). Only annotations belonging to $\mathcal{C}_t$ are preserved. After training phase t, the model is evaluated on all seen classes $\mathcal{C}_{1:t}=\mathcal{C}_1\cup\cdots\cup\mathcal{C}_t$.

Transformer-based Detectors. Following CL-DETR, we adopt Deformable DETR (short for D-DETR) as the baseline architecture, which consists of a CNN backbone, a transformer encoder-decoder, and predictors of class and bounding box. The CNN backbone and transformer encoder aim to extract enhanced feature sequences $\mathcal{V}_\mathrm{e}$. The decoder takes features and a set of learnable object queries $\mathcal{Q}=\left\{q_i\in\mathbb{R}^d|i=1,...,N\right\}$ as input and outputs the object embeddings $\mathcal{E}$, followed by predictors to parse out bounding boxes $\mathcal{B}$ and classes $\mathcal{P}$. The process can be expressed as:

$$\mathcal{E}=\mathrm{Decoder}(\mathcal{V}_e,\mathcal{Q}), \tag{1}$$

$$\mathcal{B}=\mathrm{Reg}(\mathcal{E}),\mathcal{P}=\mathrm{Cls}(\mathcal{E}). \tag{2}$$

Hungarian algorithm is used to find a bipartite matching between ground truths and predictions of object queries.

Methodology

Overview

Considering the above observations, we redesign the decoding process of DETR-based IOD into a localization-then-recognition process to maintain the less-forgetting localization well and focus on conquering fragile recognition. Figure 3 shows the overall framework of DCA, which retains the backbone and transformer encoder of D-DETR for extracting feature sequences $\mathcal{V}_e$ while modifying the transformer decoder. First, randomly initialized learnable location queries $\mathcal{Q}_{local}$ along with feature sequences $\mathcal{V}_e$ are sent to *Decoupled Localization Decoder* to obtain location embeddings $\mathcal{E}_{local}$, followed by a regression head for predicting wide-coverage boxes. Next, for prior locations and one-to-one bipartite matching between localization and recognition decoding, class queries $\mathcal{Q}_{cls}$ are initialized with location embeddings. To integrate inter-class relationships, semantic features $\mathcal{Q}_{se}$, obtained by sending class names into PLMs, enter decoder in the form of queries. $\mathcal{Q}_{cls}$, concatenated with $\mathcal{Q}_{se}$, are updated through *Decoupled Semantic-guided Recognition Decoder* to get class embeddings $\mathcal{E}_{cls}$. *Duplex Classifier Fusion* predicts final recognition scores, which adds the semantic-based head to build a unified feature optimization space within and across tasks.

Decoupled Localization and Recognition

As illustrated in Eq.(1)(2), the original coupled architecture of DETR shares localization and recognition features $\mathcal{E}$ which are sent to Reg. head and Cls. head for predicted boxes and scores. As observed in Figure 1(a)(b), since localization is less-forgetting and can also generalize to future classes similar to seen classes, directly put constraints on these coupled features to preserve old knowledge, such as knowledge distillation, will affect the generalization of localization, thereby reducing plasticity. To avoid mutual interference, we split the original transformer decoder into decoupled localization decoder and recognition decoder.

As shown in Figure 3, the feature extractor retained in D-DETR extract enhanced image feature sequences $\mathcal{V}_e$. For localization, N randomly initialized location queries $\mathcal{Q}_{local}\in\mathbb{R}^{N\times d}$ and visual features $\mathcal{V}_e$ are input to decoupled localization decoder to obtain location embeddings $\mathcal{E}_{local}\in\mathbb{R}^{N\times d}$ of foreground objects, followed by Reg. head for box coordinates $\mathcal{B}=\left\{\bar{b}_1,\ldots,\bar{b}_N\right\}\in\mathbb{R}^{N\times 4}$. The localization decoder is composed of L blocks of transformer layers, following D-DETR. For recognition, it aims to classify the objects in predicted boxes. We use all location embeddings and image feature tokens as input for decoupled recognition decoder. To integrate inter-class relationships, semantic features $\mathcal{Q}_{se}$ from pre-trained language models are also fed forward to class recognition decoder (see the next section for details). For each location embedding, we get its class embedding $\mathcal{E}_{cls}\in\mathbb{R}^{N\times d}$. Then, duplex classifier fusion head is applied to $\mathcal{E}_{cls}$ to get class probabilities $\mathcal{P}=\{\bar{p}_1,\ldots,\bar{p}_N\}\in\mathbb{R}^{N\times K}$ after $\mathrm{sigmoid}(\cdot)$, where $K=|\mathcal{C}_{1:t}|$ is the number of

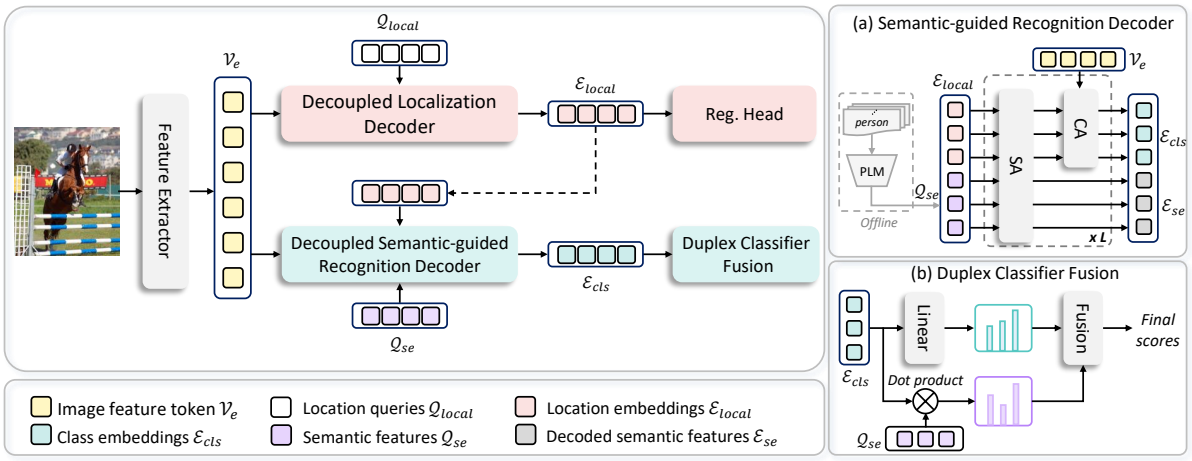


Figure 3: Pipeline of DCA. The extracted feature sequences $\mathcal{V}_e$ are first fed to decoupled localization decoder for object location embeddings $\mathcal{E}_{local}$, which are then sent to decoupled Semantic-guided Recognition Decoder to probe features to get class embeddings $\mathcal{E}_{cls}$. To integrate inter-class relationships, we embed semantic features $\mathcal{Q}_{se}$ from PLMs into recognition decoder in the form of queries and perform self-attention (SA) with location embeddings. To promote unified optimization across tasks, Duplex Classifier Fusion adds a semantic head to calculate similarities between $\mathcal{E}_{cls}$ and $\mathcal{Q}_{se}$ which are combined with the standard linear head to generate final recognition scores.

all seen classes. Compared with original coupled decoding in Eq.(1)(2), our decoupled decoding can be expressed as:

$$\mathcal{E}_{local} = \text{Decoder\_Local}(\mathcal{V}_e, \mathcal{Q}_{local}), \quad (3)$$

$$\mathcal{Q}_{cat} = \text{Concat}(\mathcal{E}_{local}, \mathcal{Q}_{se}), \quad (4)$$

$$\mathcal{E}_{cls} = \text{Decoder\_Cls}(\mathcal{V}_e, \mathcal{Q}_{cat}), \quad (5)$$

$$\mathcal{B} = \text{Reg}(\mathcal{E}_{local}), \mathcal{P} = \text{DCF}(\mathcal{E}_{cls}), \quad (6)$$

where Decoder_Local, Decoder_Cls denote the localization decoder and recognition decoder respectively, which share parameters to have a comparable parameters count to other baselines. DCF means duplex classifier fusion. After obtaining the optimal assignment $\bar{\sigma}$ between prediction results and ground-truth annotations $y = \{(c_i, b_i)\}_{i=1}^{N}$, we optimize the model by minimizing the following detection loss:

$$\mathcal{L}_{det} = \sum_{i=1}^{N} \left[-\log \bar{p}_{\bar{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \phi\}} \mathcal{L}_{box}(b_i, \bar{b}_{\bar{\sigma}(i)}) \right], \quad (7)$$

where $\mathcal{L}_{box} = \lambda_{iou}\mathcal{L}_{iou}(b_i, \bar{b}_{\bar{\sigma}(i)}) + \lambda_{\mathcal{L}1} \left\| b_i - \bar{b}_{\bar{\sigma}(i)} \right\|_1$ and $\mathbb{1}(\cdot)$ is the indicator function. Comparison in Figure 1(c)(d) illustrates that our decoupling operation isolates class-agnostic localization features and recognition features.

Class Recognition with Semantic Guidance

To conquer recognition forgetting, we explore the semantic knowledge from pre-trained language models (PLMs) to guide class recognition, which helps provide a unified optimization target for the entire incremental process. We first integrate semantic features into the recognition decoder in the form of queries to consider high-level semantic relationships between objects, forming decoupled semantic-guided recognition decoder. Then, we leverage the well-structured feature space spanned by these semantic features to design the duplex classifier fusion, reducing feature drift.

Semantic-guided Recognition Decoder. It consists of L attention blocks, each comprising a self-attention (SA) and a cross-attention (CA) layer. In the self-attention module, each query (here also referred to as location embeddings from the localization decoder) updates its representation by aggregating information from all other queries, which allows the model to reason about the relationships between different queries. It helps the decoder refine the predictions by considering the interaction among all predicted objects, such as reducing duplicate detections and adjusting bounding boxes based on the presence of other objects. We concatenate the semantic features with queries, and these enriched queries serve as the input to the decoder's self-attention layer. By integrating semantic features, each query becomes more robust in its representation, which is not limited to just physical attributes of objects (like spatial and appearance-based information) but also high-level semantic relationships between objects. It is beneficial to IOD for two reasons. Firstly, semantic features allow the model to generalize better to new classes. The model leverages shared semantic traits across different but related classes, facilitating quicker and more accurate adaptation to new data. Secondly, semantic queries ensure that the model maintains a consistent understanding and representation of old classes over multiple incremental steps by injecting the semantic features into class queries, which effectively mitigates the semantic drift of old classes.

As shown in Figure 3(a), semantic features $\mathcal{Q}_{se}$ are generated by filling known class names $\mathcal{C}_{1:t}$ into a template and feeding to the pre-trained language model. Class queries ($i.e.$, $\mathcal{E}_{local}$) and semantic features are input to self-attention to deduce relations among objects and simultaneously infuse object features with prior semantic information. The enriched queries interact with image features in the cross-attention. The l-th semantic-guided decoder block can be il-

lustrated as follows:

$$(\mathcal{E}_{cls}^{l'}, \mathcal{E}_{se}^{l}) = \mathrm{SA}(\mathcal{E}_{cls}^{l-1}, \mathcal{E}_{se}^{l-1}), \tag{8}$$

$$\mathcal{E}_{cls}^{l} = \mathrm{CA}(\mathcal{E}_{cls}^{l'}, \mathcal{V}_e), \tag{9}$$

where $\mathcal{E}_{cls}^{l'}$ is the intermediate output by SA, and $\mathcal{E}_{cls}^{l}$ and $\mathcal{E}_{se}^{l}$ are decoded class embeddings and semantic features of l-th block. To prevent semantic drift when decoded semantic features are fed into the next transformer block, we impose the semantic consistency loss to keep consistent with the original semantic features $\mathcal{Q}_{se}$ by cosine similarities $\cos(\cdot)$ across all semantic-guided recognition blocks:

$$\mathcal{L}_{cons} = \sum_{l=1}^{L} \left(1 - \cos(\mathcal{Q}_{se}, \mathcal{E}_{se}^{l})\right). \tag{10}$$

Duplex Classifier Fusion. Previous optimization of classifiers focuses narrowly on the current training data. Due to limited access to old data and the unpredictability of new data, new weights of the model overwrite previous knowledge, resulting in distortion and overlap of feature spaces between old and new tasks. It is difficult for the model to be compatible with both old and new classes. Semantic-based classifier offers a more robust solution, which is endowed with a well-constructed global feature space, spanned by semantic features derived from all known class names. Due to being pre-trained on a large dataset, this model gains a conceptual understanding of data and can learn universal semantic relationships both within and across tasks. When new classes arrive, pre-defined new semantic features can be registered without compromising the integrity of the old feature space, thereby mitigating catastrophic forgetting. Integrating the strengths of both, we propose duplex classifier fusion for the balance between plasticity and stability.

As illustrated in Figure 3(b), class embeddings $\mathcal{E}_{cls}$ are fed through the linear layer to get class probabilities $\mathcal{H} = \{\bar{h}_1, \ldots, \bar{h}_N\} \in \mathbb{R}^{N \times K}$ after $\mathrm{sigmoid}(\cdot)$ activation. On the other hand, a projection layer is used to map class embeddings to semantic space for projected class embeddings $\mathcal{E}_{proj}$ and then semantic class probabilities $\mathcal{S} = \{\bar{s}_1, \ldots, \bar{s}_N\} \in \mathbb{R}^{N \times K}$ are obtained by calculating similarities between projected class embeddings and semantic features of all known classes. Here, we directly employ a weighted approach (weight $\beta = 0.5$) to fuse these probabilities to get the overall classification probabilities $\mathcal{P} = \{\bar{p}_1, \ldots, \bar{p}_N\} \in \mathbb{R}^{N \times K}$ for training and inference:

$$\bar{p}_i = \beta \cdot \bar{h}_i + (1 - \beta) \cdot \bar{s}_i \tag{11}$$

Hybrid Knowledge Distillation

Given that the current training data contains unlabeled old objects, we add pseudo labels to foreground predictions of the old model as supplement supervision, solving background interference. However, the shortage of old data leads to the bias towards new ones. We employ hybrid knowledge distillation to retain a diverse range of information from various perspectives and levels. Firstly, we constrain the detection outputs including class probabilities and boxes:

$$\mathcal{L}_{out}^{kd} = \mathcal{L}_{mse}(\bar{p}^{new}, \bar{p}^{old}) + \mathcal{L}_{box}(\bar{b}^{new}, \bar{b}^{old}), \tag{12}$$

where $(\bar{p}^{new}, \bar{b}^{new})$, $(\bar{p}^{old}, \bar{b}^{old})$ represent prediction results of new and old model respectively. Meanwhile, significant drift in image features can adversely affect the recognition decoding, leading to a marked reduction in class distinction. Class embeddings are also susceptible to severe deviation that exacerbate forgetting. To counteract this, we preserve visual features learned by the old model, encompassing both the encoded features $\mathcal{V}_e$ and decoded class embeddings $\mathcal{E}_{cls}$. To improve the plasticity for new classes, we use pseudo-labels of old classes as instance-level masks for selection distillation:

$$\mathcal{L}_{vis}^{kd} = \mathcal{G}(\mathcal{V}_e) + \mathcal{G}(\mathcal{E}_{cls}), \tag{13}$$

where $\mathcal{G}(f) = \frac{1}{N^{old}} \sum_{j=1}^{N^{old}} A_{ij} \left\| f_{ij}^{new} - f_{ij}^{old} \right\|_1$, and N^{old} is the sum of old pseudo-boxes. To prevent the projected features $\mathcal{E}_{proj}$ shifting, we impose distillation:

$$\mathcal{L}_{proj}^{kd} = \mathcal{G}(\mathcal{E}_{proj}). \tag{14}$$

Then the hybrid knowledge distillation is $\mathcal{L}_{hkd} = \mathcal{L}_{out}^{kd} + \mathcal{L}_{vis}^{kd} + \mathcal{L}_{proj}^{kd}$. To sum up, the overall loss for training the new model is given by ($\mathcal{L}_{hkd} = 0$ when training base model):

$$\mathcal{L}_{all} = \mathcal{L}_{det} + \mathcal{L}_{cons} + \mathcal{L}_{hkd}. \tag{15}$$

Experiments

Experimental Setups

Datasets and Metrics. We evaluate on two widely used datasets: PASCAL VOC (Everingham 2007) and MS COCO (Lin et al. 2014). VOC contains 20 foreground classes, while COCO covers 80 object categories. AP_{50} and AP are reported for metrics. To measure the gap between an incremental model response and the ideal setting, we use the absolute gap (AbsGap) and relative gap (RelGap).

Incremental Protocols. We simulate diverse learning scenarios. For VOC, we consider three different settings, where a group of classes (10, 5 and last class) are introduced incrementally to the detector. For COCO, we conduct experiments under 70+10, 60+20, 50+30 and 40+40 settings. To increase the task difficulty, multi-step settings are evaluated where base model is trained with 40 classes and 20 or 10 classes are added in each of the following phases.

Implementation Details. The architecture of DCA detector is an adaptation of D-DETR, which leverages ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) as the backbone. Following CL-DETR, we use the standard configurations without iterative bounding box refinement and the two-stage mechanism. For the shared decoder, the number of layers is set to $L = 6$ and the number of location queries $N = 100$. During inference, top-50 high-scoring detections per image are used for evaluation. In DCA, we use CLIP text encoder as the language model to generate semantic features in an offline manner. Our codes are available at https://github.com/InfLoop111/DCA.

Comparison with State-Of-The-Art Methods

One-step incremental settings. As reported in Table 1, DCA outperforms all previous non-exemplar methods in different data split setting. In particular, in 10+10 setting, DCA

	Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
10+10 Setting	Faster ILOD	72.8	75.7	71.2	60.5	61.7	70.4	83.3	76.6	53.1	72.3	36.7	70.9	66.8	67.6	66.1	24.7	63.1	48.1	57.1	43.6	62.1
	MMA					69.3										63.9						66.6
	ORE-EBUI*	63.5	70.9	58.9	42.9	34.1	76.2	80.7	76.3	34.1	66.1	56.1	70.4	80.2	72.3	81.8	42.7	71.6	68.1	77	67.7	64.5
	OW-DETR*	61.8	69.1	67.8	45.8	47.3	78.3	78.4	78.6	36.2	71.5	57.5	75.3	76.2	77.4	79.5	40.1	66.8	66.3	75.6	64.1	65.7
	CL-DETR†	55.0	63.0	51.1	35.3	36.2	48.2	60.5	45.9	29.8	21.1	61.6	72.1	75.0	75.4	75.1	38.5	65.4	62.2	76.4	68.4	55.8
	PROB†	70.4	75.4	67.3	48.1	55.9	73.5	78.5	75.4	42.8	72.2	64.2	73.8	76.0	74.8	75.3	40.2	66.2	73.3	64.4	64.0	66.5
	ACF					67.0										70.1						68.6
	DCA	79.7	82.3	71.0	61.3	65.1	80.0	86.9	77.3	56.2	73.1	65.8	81.8	85.2	81.5	80.6	45.6	70.0	68.8	82.7	73.0	**73.4**
15+5 Setting	Faster ILOD	66.5	78.1	71.8	54.6	61.4	68.4	82.6	82.7	52.1	74.3	63.1	78.6	80.5	78.4	80.4	36.7	61.7	59.3	67.9	59.1	67.9
	MMA								73.0										60.5			69.9
	ORE-EBUI*	75.4	81	67.1	51.9	55.7	77.2	85.6	81.7	46.1	76.2	55.4	76.7	86.2	78.5	82.1	32.8	63.6	54.7	77.7	64.6	68.5
	OW-DETR*	77.1	76.5	69.2	51.3	61.3	79.8	84.2	81.0	49.7	79.6	58.1	79.0	83.1	67.8	85.4	33.2	65.1	62.0	73.9	65.0	69.1
	CL-DETR†	64.0	60.0	56.6	38.5	44.9	42.6	68.7	55.0	36.4	50.9	37.8	59.2	76.0	71.4	72.2	32.1	32.1	34.6	47.7	52.3	51.6
	PROB*	77.9	77.0	77.5	56.7	63.9	75.0	85.5	82.3	50.0	78.5	63.1	75.8	80.0	78.3	77.2	38.4	69.8	57.1	73.7	64.9	70.1
	ACF								71.6										65.9			70.2
	DCA	75.0	84.1	78.2	64.8	63.8	69.9	87.6	86.1	58.6	72.2	73.0	83.7	83.4	82.6	83.8	36.7	52.3	53.0	72.6	61.7	**71.2**
19+1 Setting	Faster ILOD	64.2	74.7	73.2	55.5	53.7	70.8	82.9	82.6	51.6	79.7	58.7	78.8	81.8	75.3	77.4	43.1	73.8	61.7	69.8	61.1	68.5
	MMA										71.1										63.4	70.7
	ORE-EBUI*	67.3	76.8	60	48.4	58.8	81.1	86.5	75.8	41.5	79.6	54.6	72.8	85.9	81.7	82.4	44.8	75.8	68.2	75.7	60.1	68.8
	OW-DETR*	70.5	77.2	73.8	54.0	55.6	79.0	80.8	80.6	43.2	80.4	53.5	77.5	89.5	82.0	74.7	43.3	71.9	66.6	79.4	62.0	69.8
	CL-DETR†	63.1	61.0	56.3	42.2	50.9	48.5	72.5	61.3	44.3	68.2	47.3	67.3	66.4	63.9	71.9	37.5	66.9	44.6	62.2	45.7	57.1
	PROB*	80.3	78.9	77.6	59.7	63.7	75.2	86.0	83.9	53.7	82.8	66.5	82.7	80.6	83.8	77.9	48.9	74.5	69.9	77.6	48.5	72.6
	ACF										71.9										66.9	70.6
	DCA	73.9	83.2	79.6	60.1	63.6	75.0	86.7	85.6	61.2	79.8	71.3	83.2	85.3	82.4	83.4	49.9	76.3	70.0	79.2	61.1	**74.5**

Table 1: Per-class average precision on VOC test dataset where 10, 5 or 1 classes are added at once. Best among rows **in bold** and second best are underlined. Methods with * store old-class data or use extra wild data and with † come from re-implementation.

Method	**70+10**		**60+20**		**50+30**		**40+40**	
	AP	AP50	AP	AP50	AP	AP50	AP	AP50
LwF	7.1	12.4	5.8	10.8	5.0	9.5	17.2	25.4
RILOD	24.5	37.9	25.4	38.8	28.5	43.2	29.9	45.0
SID	32.8	49.0	32.7	49.8	33.8	51.0	34.0	51.4
ERD	34.9	51.9	35.8	52.9	36.6	54.0	36.9	54.5
CL-DETR	35.8	53.5	-	-	-	-	39.2	56.1
CL-DETR*	40.4	58.0	-	-	-	-	42.0	**60.1**
ACF	37.6	-	38.3	-	38.8	-	39.8	-
DCA	**41.3**	**59.2**	**41.9**	**54.8**	**39.9**	**56.1**	**42.8**	58.4

Table 2: Results (AP/AP_{50}) on COCO two-step setting. - represents no corresponding results in the original paper.

Method	**(1-40)**	**+(40-60)**	**+(60-80)**	**AbsGap↓**	**RelGap↓**
CF	45.7/ 66.3	10.7/ 15.8	9.4/ 13.3	30.8/ 45.0	0.77/ 0.77
RILOD	45.7/ 66.3	27.8/ 42.8	15.8/ 4.0	24.4/ 54.3	0.61/ 0.93
SID	45.7/ 66.3	34.0/ 51.8	23.8/ 36.5	16.4/ 21.8	0.41/ 0.37
ERD	45.7/ 66.3	36.7/ 54.6	32.4/ 48.6	7.8/ 9.7	0.19/ 0.17
ACF	48.0/ -	39.3/ -	36.6/ -	3.7/ -	0.09/ -
DCA	48.0/ 68.9	**42.7/ 59.6**	**40.3/ 54.1**	**2.3/ 7.3**	**0.05/ 0.12**

Table 3: Results (AP/AP_{50}) under two-step setting on COCO where (a-b) is the base normal training for classes a-b and +(c-d) is the incremental training for classes c-d.

significantly improves the best non-exemplar method ACF by 4.8%. DCA even exceeds all exemplar-based methods which store samples. Take the result of 19+1 setting as an example, we surpass the second best method PROB by 1.9%, which validates the superiority of DCA. This phenomenon remains the same in COCO dataset with more categories. As shown in Table 2, in 70+10 setting, DCA achieves notable improvements of 0.9% in AP and 1.2% in AP50 even though the best method stores exemplars to match the training set distribution. When adding more new classes, DCA consistently obtains best in both AP and AP50. Notably, improvements become more apparent as the number of categories in the initial stage increases.

Multi-step incremental settings. Table 3 and Table 4 show the results under a more demanding setting, where multiple incremental steps are performed to learn new classes. Remarkably, without ER to mitigate forgetting, the performance of existing methods declines drastically. For example, the AP50 of ERD decreased from 66.3% in the first stage to 31.8% after learning all classes in four-step setting, nearing 50% drop. DCA consistently maintains high performance, with AP50 ranging from 68.9% to 54.1% in two-step setting and 68.9% to 49.6% in four-step setting. DCA exhibits significantly smaller AbsGap, which quantifies the absolute gap to joint training, and RelGap measures the relative

Method	+(40-50)	+(50-60)	+(60-70)	+(70-80)	AbsGap↓	RelGap↓
CF	5.8 8.5	5.7 8.3	6.3 8.5	3.3 4.8	36.9 53.5	0.92 0.92
RILOD	25.4 38.9	11.2 17.3	10.5 15.6	8.4 12.5	31.8 45.8	0.79 0.79
SID	34.6 52.1	24.1 38.0	14.6 23.0	12.6 23.3	27.6 35.0	0.69 0.60
ERD	36.4 53.9	30.8 46.7	26.2 39.9	20.7 31.8	19.5 26.5	0.49 0.46
ACF	39.1 -	35.4 -	32.0 -	30.3 -	10.0 -	0.25 -
DCA	**44.0** **61.2**	**41.1** **56.5**	**39.2** **53.8**	**37.2** **49.6**	**5.4** **11.8**	**0.13** **0.19**

Table 4: Results (AP/AP_{50}) under COCO four-step setting.

Row	DLR	SRD	DCF	HKD	All (gain △)	Old	New	Avg.
1					62.5 (+0.0)	64.8	55.4	60.1
2	✓				65.7 (+3.2)	68.9	56.1	62.5
3		✓			64.1 (+1.6)	67.3	54.7	61.0
4	✓	✓			67.9 (+5.4)	70.8	59.0	64.9
5	✓	✓	✓		68.6 (+6.1)	72.1	57.0	64.6
6	✓	✓	✓	✓	**71.2** (+8.7)	76.5	55.3	65.9
7				✓	66.0 (+3.5)	72.6	46.0	59.3

Table 5: Ablations on VOC 15+5 setting. DLR denotes Decoupled Localization and Recognition. SRD is Semantic-guided Recognition Decoder, DCF means Duplex Classification Fusion. HKD is Hybrid Knowledge Distillation.

gap, indicating that our performance is approaching the upper bound. All results confirm that semantics-guided decoding framework is a powerful and robust incremental detection pipeline. Meanwhile, unlike architecture-based methods, DCA does not increase many network parameters except for some projection layers for dimensional alignment.

Ablation Studies

Component ablations. As shown in Table 5, fine-tuning with pseudo-labeling gets the lowest result 62.5%. Progressive fusions of proposed modules enjoy consistent performance gains. DLR decouples localization and recognition features and reduces the nutual influence of forgetting, while semantic guidance can be precisely applied to the recognition features, maximizing its efficacy. Further combined with HKD, our complete framework surpasses the baseline by up to 8.7%, while using only HKD results in 66.0% mAP. These results validate that using semantic guidance during decoding to eliminate reliance on old data is effective and these components are complementary to each other.

Analysis of HKD. Results in Table 6 indicate the critical role of feature regularization $\mathcal{L}_{cls}^{kd}$, $\mathcal{L}_{vis}^{kd}$ and $\mathcal{L}_{proj}^{kd}$, verifying that recognition forgetting is significantly influenced by feature drift. Combining with other modules, the forgetting

Row	$\mathcal{L}_{cls}^{kd}$	$\mathcal{L}_{vis}^{kd}$	$\mathcal{L}_{proj}^{kd}$	All	Old	New	Avg.
1	✓			68.8	72.7	57.2	64.9
2	✓	✓		70.5	75.0	57.1	64.6
3	✓		✓	69.0	73.3	56.4	64.9
4	✓	✓	✓	**71.2**	76.5	55.3	65.9

Table 6: Ablation results on Hybrid Knowledge Distillation.

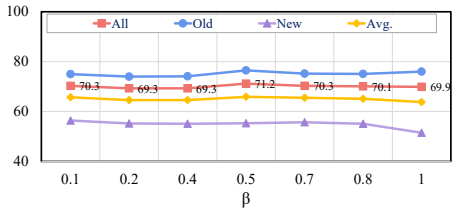


Figure 4: Impact analysis about the balance weight. β controls the importance of probabilities of the linear classifier.

issues in IOD are further mitigated.

Analysis of the balance weight. β controls the relative importance of the standard linear head and our introduced semantic head in duplex classifier. β is set to $\{0.1, 0.2, \ldots, 1.0\}$. All experiments are run in the 15+5 setting on VOC benchmark, and results are reported in Figure 4. When $\beta = 0.5$, DCA establishes the best performance on the All metric equally considering all classes, and the Avg. metric equally weighting new and old classes attains the highest. Although duplex classifier fusion increases inference time by 1.6% (from 0.0687 to 0.0698 s/sample), it is negligible compared to the performance improvement.

Robustness to different language models. We supplement results of using other language models to obtain class semantic features. In 15+5 setting, results on CLIP, BERT-M (Kenton and Toutanova 2019) and BERT-S are 71.2%, 71.0% and 70.5% respectively, both higher than 70.2% of exemplar-free IOD. Our method focuses on leveraging semantic space to boost IOD and is not restricted to the utilization of CLIP text encoder, which offers a viable route to pre-trained semantic spaces.

Conclusions

In this paper, we uncover the forgetting imbalance between localization and recognition in transformer-based IOD and propose a Divide-and-Conquer Amnesia (DCA) strategy to effectively mitigate catastrophic forgetting. DCA restructures transformer-based IOD into a localization-then-recognition process, which isolates less-forgetting localization features from the coupled features, thereby leaving the fragile recognition forgetting to be conquered. To reduce recognition feature drift, we leverage semantic knowledge to establish a consistent optimization target throughout the incremental process. For future works, DCA offers a fresh perspective for tackling the challenges of transformer-based IOD, promising significant advancements in the field. Extensive experiments demonstrate that DCA achieves state-of-the-art, with the advantage of exemplar-free overhead.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant NO 62406318, 62376266, 62076195, 62376070), Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing, China, and by the Key Research Program of Frontier Sciences, CAS (Grant NO ZDBS-LY-7024).

References

Acharya, M.; Hayes, T. L.; and Kanan, C. 2020. Rodeo: Replay for online object detection. *British Machine Vision Conference*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, 213–229.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. AdaptFormer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. PODNet: Pooled outputs distillation for small-tasks incremental learning. In *Proceedings of the European Conference on Computer Vision*, 86–102.

Douillard, A.; Ramé, A.; Couairon, G.; and Cord, M. 2022. DyTox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9285–9295.

Everingham, M. 2007. The pascal visual object classes challenge,(voc2007) results. *http://pascallin. ecs. soton. ac. uk/challenges/VOC/voc2007/index. html.*

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Hao, Y.; Fu, Y.; Jiang, Y.-G.; and Tian, Q. 2019. An end-to-end architecture for class-incremental object detection with knowledge distillation. In *2019 IEEE International Conference on Multimedia and Expo*, 1–6.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 831–839.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Huang, L.; An, Z.; Zeng, Y.; Xu, Y.; et al. 2024a. KFC: Knowledge reconstruction and feedback consolidation enable efficient and effective continual generative learning. In *The Second Tiny Papers Track at ICLR 2024*.

Huang, L.; Zeng, Y.; Yang, C.; An, Z.; Diao, B.; and Xu, Y. 2024b. eTag: Class-incremental learning via embedding distillation and task-oriented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12591–12599.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, 709–727.

Joseph, K.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021a. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5830–5840.

Joseph, K.; Rajasegaran, J.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021b. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9209–9216.

Kang, M.; Zhang, J.; Zhang, J.; Wang, X.; Chen, Y.; Ma, Z.; and Huang, X. 2023. Alleviating catastrophic forgetting of incremental object detection via within-class and between-class knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18894–18904.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, L.; Kuang, Z.; Chen, Y.; Xue, J.-H.; Yang, W.; and Zhang, W. 2020a. IncDet: In defense of elastic weight consolidation for incremental object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6): 2306–2319.

Liu, X.; Yang, H.; Ravichandran, A.; Bhotika, R.; and Soatto, S. 2020b. Multi-task incremental learning for object detection. *arXiv preprint arXiv:2002.05347*.

Peng, C.; Zhao, K.; Maksoud, S.; Li, M.; and Lovell, B. C. 2021. SID: Incremental learning for anchor-free object de-

tection via Selective and Inter-related Distillation. *Computer Vision and Image Understanding*, 210: 103229.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCARL: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2001–2010.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28: 91–99.

Shmelkov, K.; Schmid, C.; and Alahari, K. 2017. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3400–3409.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. CODA-Prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11909–11919.

Wang, R.; Duan, X.; Kang, G.; Liu, J.; Lin, S.; Xu, S.; Lü, J.; and Zhang, B. 2023. AttriCLIP: A non-incremental learner for incremental knowledge learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3654–3663.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Yan, S.; Xie, J.; and He, X. 2021. DER: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3014–3023.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024. CLIP-KD: An empirical study of CLIP model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.

Yang, D.; Zhou, Y.; Hong, X.; Zhang, A.; and Wang, W. 2023a. One-Shot Replay: Boosting incremental object detection via retrospecting one object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3127–3135.

Yang, D.; Zhou, Y.; Hong, X.; Zhang, A.; Wei, X.; Zeng, L.; Qiao, Z.; and Wang, W. 2023b. Pseudo object replay and mining for incremental object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 153–162.

Yang, D.; Zhou, Y.; Shi, W.; Wu, D.; and Wang, W. 2022a. RD-IOD: Two-level residual-distillation-based triple-network for incremental object detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(1): 1–23.

Yang, D.; Zhou, Y.; Zhang, A.; Sun, X.; Wu, D.; Wang, W.; and Ye, Q. 2022b. Multi-view correlation distillation for incremental object detection. *Pattern Recognition*, 131: 108863.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, W.; Chang, S.; Sosa, N.; Hamann, H.; and Cox, D. 2020. Lifelong object detection. *arXiv preprint arXiv:2009.01129*.

Zhu, Z.; Hong, X.; Ma, Z.; Zhuang, W.; Ma, Y.; Dai, Y.; and Wang, Y. 2025. Reshaping the Online Data Buffering and Organizing Mechanism for Continual Test-Time Adaptation. In *European Conference on Computer Vision*, 415–433. Springer.