Learning Task-Aware Language-Image Representation for Class-Incremental Object Detection

Hongquan Zhang[1, 3*], **Bin-Bin Gao**[2*], **Yi Zeng**[2], **Xudong Tian**[1, 3], **Xin Tan**[1, 3†], **Zhizhong Zhang**[1, 3], **Yanyun Qu**[4], **Jun Liu**[2], **Yuan Xie**[1,3]

[1]East China Normal University
[2]Tencent YouTu Lab
[3]Chongqing Institute of East China Normal University
[4] Xiamen University
{51215901136,51194501066}@stu.ecnu.edu.cn, {xtan,zzzhang,yxie}@cs.ecnu.edu.cn,
{csgaobb,yizengstudy,junsenselee}@gmail.com, yyqu@xmu.edu.cn

Abstract

Class-incremental object detection (CIOD) is a real-world desired capability, requiring an object detector to continuously adapt to new tasks without forgetting learned ones, with the main challenge being catastrophic forgetting. Many methods based on distillation and replay have been proposed to alleviate this problem. However, they typically learn on a pure visual backbone, neglecting the powerful representation capabilities of textual cues, which to some extent limits their performance. In this paper, we propose task-aware language-image representation to mitigate catastrophic forgetting, introducing a new paradigm for language-image-based CIOD. First of all, we demonstrate the significant advantage of language-image detectors in mitigating catastrophic forgetting. Secondly, we propose a learning task-aware language-image representation method that overcomes the existing drawback of directly utilizing the language-image detector for CIOD. More specifically, we learn the language-image representation of different tasks through an insulating approach in the training stage, while using the alignment scores produced by task-specific language-image representation in the inference stage. Through our proposed method, language-image detectors can be more practical for CIOD. We conduct extensive experiments on COCO 2017 and Pascal VOC 2007 and demonstrate that the proposed method achieves state-of-the-art results under the various CIOD settings.

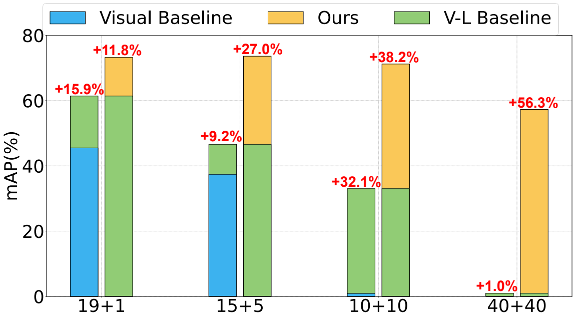

Figure 1: The performance comparisons (mAP[0.5]) of language-image and pure visual-based detectors on Pascal VOC 2007 and COCO 2017 with various incremental protocols. Here, we take Dyhead (Dai et al. 2021a) as a pure visual-based detector, and GLIP as a language-image detector, where GLIP replaces the classifier used in Dyhead with language-image alignment and others the same as Dyhead. We can see that the language-image detector (GLIP) brings clear improvements in most incremental settings. However, there is still catastrophic forgetting in some challenging settings (e.g., 40+40 on COCO). To further address this issue, we first learn task-aware language-image representation and then use a selective inference strategy for CIOD, which outperforms naive GLIP by a large margin.

Introduction

Object detection has shown remarkable advancements in facilitating various applications, including traffic monitoring, robotics (Xu et al. 2022) and autonomous driving (Li et al. 2023a). Most object detection works mainly focus on the offline training paradigm. However, online training plays a more important role in real-world applications in dynamic environments which urgently requires a model to continuously recognize new classes and maintain the ability on learned classes. Therefore, exploiting continual (incremental) object detection based on online data streaming has become an attractive yet challenging topic and aims to sequentially solve tasks with ideally no performance drop when inferred on the previously seen tasks.

In order to mitigate performance drop on previous classes, some class-incremental object detection (CIOD) methods either use knowledge distillation (Shmelkov, Schmid, and Alahari 2017; Hu et al. 2021; Peng et al. 2021; Feng, Wang, and Yuan 2022) on image features or replay a small number of previous exemplars (Shieh et al. 2020; Joseph et al. 2021a,b; Liu et al. 2023b). Commonly, these methods typi-

[*]These authors contributed equally, work done while at Tencent YouTu Lab.

[†]Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

cally learn on a pure visual backbone, and their performance is limited to training data (images and the corresponding annotations) to some extent. It's worth noting that it always suffers from serious background-foreground conflict, which means a proposal that belongs to the foreground in previous tasks is likely to be the background in future tasks in incremental detection scenarios. Therefore, the catastrophic forgetting issue will further be exacerbated by adopting pure visual-based CIOD methods due to background-foreground conflict.

Recently, language-image models have exhibited impressive results on zero-shot (Radford et al. 2021; Zhou et al. 2022; Zhai et al. 2022) and continual image classification (Li et al. 2023b). Meanwhile, this cross-modality learning paradigm has shown strong zero-shot and few-shot transferability to object detection, such as GLIP (Li et al. 2022), Grounding Dino (Liu et al. 2023a), and MQ-Det (Xu et al. 2023). Considering catastrophic forgetting issues in CIOD, we believe that this strong transferability should benefit incremental object detection tasks because of the separability between visual features and language representation. Here, we take GLIP as an example and simply extend it to the CIOD setting. The experimental comparisons of language-image GLIP and pure visual baseline are shown in Figure 1 on COCO and VOC with different incremental protocols. We can see that the language-image detector (GLIP) brings clear improvements (e.g., 32.1% mAP with 10+10 setting on VOC) compared pure visual method in most incremental settings. However, there is still catastrophic forgetting in some challenging settings (e.g., 40+40 on COCO).

The above drawback is mainly attributed to serious background-foreground conflict due to the instance of different categories contained in an image becoming more dispersed across different tasks as the categories increase. More specifically, the separability between visual features and language representation has been insufficient to resolve this serious background-foreground conflict, resulting in poor performance in these challenging settings. Inspired by Der (Dai et al. 2021a) which uses an independent model to learn visual representation for each task, we consider that this learning paradigm can reinforce the separability to better mitigate catastrophic forgetting, as learning independent representation will no longer be subject to the background-foreground conflict dilemma.

To this end, we propose a learning task-aware language-image representation method that further separates the visual features and language representation to mitigate catastrophic forgetting. Specifically, in the training stage, a Task-Aware Module (TAM) is proposed to account for a part of the non-overlapping channels of the image feature map and the hidden states of text embedding for producing task-aware representation in each task. While in the inference stage, for the alignment scores predicted by language-image alignment, a Selective Inference Strategy (SIS) is proposed to use the task-aware portion of the alignment scores to unify a final clarifying prediction alignment score. Our contributions are summarized below:

- We are the first to apply the language-image detector to class-incremental object detection and identify its superiority in mitigating catastrophic forgetting over the pure visual-based detector.
- We propose a learning task-aware language-image representation method, which mitigates the background-foreground conflict by reinforcing the separability of the language-image detector.
- The leap in performance compared with all competitors on various benchmarks demonstrates its efficacy, while substantial qualitative evidence verifies each of our designs.

Related Work

Incremental Learning

Incremental learning algorithms intend to mitigate catastrophic interference while facilitating the transfer of skills whenever possible and achieve excellent performance in downstream tasks like classification, detection, *etc*. To this end, currently, popular studies can be roughly categorized as (1) the rehearsal-based approach aims to help the model not forget the old knowledge when learning a new task by saving a part of the old samples (Zhao et al. 2021; Petit et al. 2023); (2) the regularization-based method adds a penalty term to the loss function when learning new tasks so that the model is optimized to adapt all tasks (Yang et al. 2021; Zhao et al. 2023; Tian et al. 2023); (3) the method based on parameter isolation (Yan, Xie, and He 2021; Wang et al. 2022; Cai et al. 2023) separates the model parameters used by different tasks, so as to mitigate catastrophic forgetting. Most related to our work is Der (Yan, Xie, and He 2021), but it is only applicable for a few incremental steps due to the linear growth of model parameters. On the contrary, we use an approach based on learning task-aware representation, which is more compatible with incremental learning, and with arbitrary incremental steps, our model parameters keep constant.

Incremental Object Detection

Class-incremental object detection is a common scenario in practical applications (Shmelkov, Schmid, and Alahari 2017), where images could contain lots of instances that belong to different tasks, and the annotation of instances is provided only current task. To solve this problem, existing studies on this issue fall into two main categories: (1) Knowledge Distillation-based, which adds regularization terms to the learning objective as an attempt to preserve previous knowledge when training the model on new data (Cermelli et al. 2022; Feng, Wang, and Yuan 2022; Yang et al. 2022); (2) Rehearsal-based utilize a buffer to memorize some of the past training data, replaying them in the following phases to “call back” the old object categories (Shmelkov, Schmid, and Alahari 2017; Liu et al. 2023b). Several methods make different efforts to class-incremental object detection, e.g., meta learning-based (Joseph et al. 2021b), regularization-based (Liu et al. 2020), and pseudo labels (Guan et al. 2018).

However, these methods are based on visual-only detectors such as Faster-Rcnn (Girshick 2015), GFL (Li et al. 2020), and DETR-based detector (Zhu et al. 2021), but neglect the rich textual represents, leading to seriously catastrophic forgetting. In our work, we explore the application

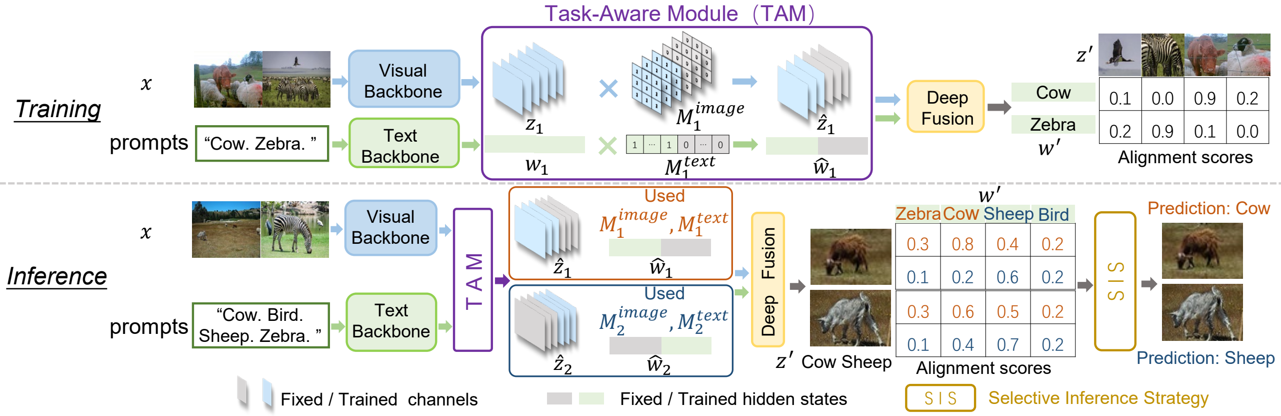

Figure 2: The whole pipeline of our method. For clear demonstration, we assume there are two tasks in the whole incremental learning process, and categories [Cow, Zebra] belong to Task 1 while [Sheep, Bird] belongs to Task 2. The TAM in the training stage and the corresponding inference strategy SIS are proposed to learn task-aware language-image representation.

of language-image detectors in class-incremental object detection. Although it has a stronger ability to mitigate catastrophic forgetting than visual-only detectors, it still faces the problem of task alignment confusion. To this end, we propose an effective method to solve it.

Language-Image Pre-training

In recent years, language-image pre-training models have been widely developed and applied to various vision tasks like detection (Li et al. 2022) and classification (Radford et al. 2021). CLIP (Radford et al. 2021) is applied to incremental classification tasks, and main methods (Zhou et al. 2022; Wang et al. 2023) aims to design different prompts to better utilize the rich knowledge of pre-trained models to help mitigate catastrophic forgetting. Different from them, a phase grounding-based object detector GLIP is used as the baseline. We note that the pre-trained weights of GLIP are not used, and aim to explore the application of the language-image alignment model in class-incremental object detection.

Methodology

Preliminaries

Class-Incremental Object Detection. Let $\mathcal{C} = \{1, \ldots, c\}$ be the set of object categories, In CIOD, a task $\mathcal{T}_t$ is defined as a subset of $\mathcal{C}$, the detector is exposed to at time $t : \mathcal{T}_t \subset \mathcal{C}$, where $\mathcal{T}_i \cap \mathcal{T}_j = \emptyset$, for any $i, j \leq t$. Let $(\boldsymbol{x}, y) \in \mathcal{D}$ denote a dataset $\mathcal{D}$ which contains images $\boldsymbol{x}$ and their corresponding ground truth sets of objects y, *i.e.,* class labels and location information, such that $\mathcal{D}_t$ denote the images containing annotated class objects in $\mathcal{T}_t$. CIOD aims to maintain original performance on $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{t-1}\}$ while continually learning task $\mathcal{T}_t$ without access to all of $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{t-1}\}$.

Grounded Language-Image Learning. GLIP (Li et al. 2022) is a language-image detector that reformulates detection as a grounding task by aligning each region in the image to a phrase in language prompts. Given object categorizes [airplane, car, cow, ..., cat], the prompt is designed as:

"airplane. car. cow. ... cat".

GLIP is mainly composed of (1) a visual backbone $f_\theta(\cdot)$ and a language backbone $g_\psi(\cdot)$. Specifically, the image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ and the word token $\boldsymbol{e} \in \mathbb{R}^{D}$ are fed into $f_\theta(\cdot)$ and $g_\psi(\cdot)$ respectively to obtain the image feature map $\boldsymbol{z} \in \mathbb{R}^{H'\times W'\times C'}$ and the text embedding $\boldsymbol{w} \in \mathbb{R}^{D\times L}$, where H, W, and C are the heights, widths and channels of $\boldsymbol{x}$ respectively, while D and L indicate the amount of tokens and the length of each token; (2) a deep fusion module used to fuse the image feature maps and text embedding in the last few encoding layers and can be defined as:

$$\boldsymbol{z}', \boldsymbol{w}' = DeepFusion(\boldsymbol{z}, \boldsymbol{w}). \tag{1}$$

On this basis, the alignment scores are $\boldsymbol{z}'(\boldsymbol{w}')^{\intercal}$, and the alignment loss is formulated as follows.

$$\mathcal{L}_{ground} = loss(\boldsymbol{z}'(\boldsymbol{w}')^{\intercal}, T),$$

where *loss* is a focal loss (Lin et al. 2017b), T is the corresponding token labels which is 1 if $\boldsymbol{z}'$ and $\boldsymbol{w}'$ aligned, and 0 otherwise. The training objective of GLIP is defined as:

$$\mathcal{L}_{vl} = \mathcal{L}_{ground} + \mathcal{L}_{reg} + \mathcal{L}_{center}, \tag{2}$$

where $\mathcal{L}_{reg}$ and $\mathcal{L}_{center}$ denote the box regression loss and the centerness loss (Tian et al. 2019) respectively.

Incremental Language-Image Detector

Baseline Setting. We employ GLIP (Li et al. 2022) as our baseline detector and build a CIOD framework based on fine-tuning. In the first task, the visual backbone and language backbone are initiated by pre-trained weights on ImageNet (Deng et al. 2009) and BERT (Kenton and Toutanova 2019), respectively. Afterward, the model is updated by optimizing Eq. (2) with $\mathcal{D}_1$. While in the incremental task $\mathcal{T}_t$, the trained weight on $\mathcal{T}_{t-1}$ is used to initial the whole model and then updates it by optimizing Eq. (2) with $\mathcal{D}_t$. Here we note the prompts are disjoint for different tasks.

Forgetting Analysis. As analyzed in Sec. 1, the above baseline has a strong ability to mitigate catastrophic forgetting,

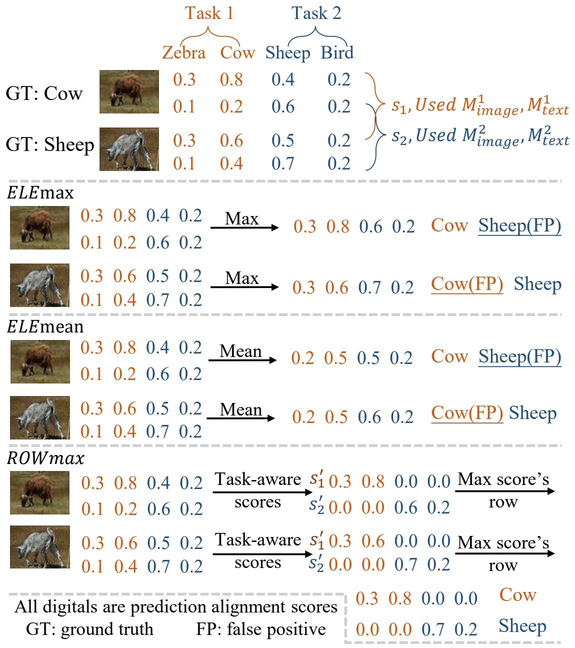

Figure 3: A visualized example of our selective inference strategy which links to the SIS of Figure 2. Our *ROWmax* strategy makes clearer predictions than naive solutions (*ELEmax* and *ELEmean*).

which can be attributed to the expressive power of the pre-trained language branch, hence we maintain a rather slow update during the training phase to have it appropriately adapt to each task. However, as the amount of categories increases, its superiority in mitigating catastrophic forgetting rapidly vanishes. We first conduct the following empirical study to reveal how and where forgetting occurs. As shown in Figure 4, the image feature map $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ have subtle differences, while significant disparities exist on $\boldsymbol{z}_1'$ and $\boldsymbol{z}_2'$. This phenomenon allows us to comprehend two aspects (1) seriously catastrophic forgetting occurs in the deep fusion module due to substantial distinctions of the deep fused image feature maps; (2) At Task 1, the total channels of the image feature map and hidden states of text embedding are used to deep fuse. But when it comes to Task 2, all of the channels and hidden states need to be reused, leading to only focusing on the labeled regions in the feature map of Task 2. Based on the above analysis, it becomes critical to address the catastrophic forgetting caused by using all channels and hidden states for deep fusion.

Task-Aware Representation Learning

To address the above problem, we propose to learn task-aware language-image representation by a Task-Aware Module (TAM) that selects partial channels and hidden states respectively for deep fusion. The whole incremental learning pipeline is shown in Figure 2. Specifically, for the training phase (the upper part of Figure 2), images $\boldsymbol{x}$ and prompts are fed into the visual and language backbone respectively, where image $\boldsymbol{x}$ and prompts consist of category labels belong to $\mathcal{D}_1$. Afterward, the feature map $\boldsymbol{z}$ and text embedding $\boldsymbol{w}$ are partially utilized by the TAM, and then fed into the deep fusion module to learn task-aware language-image representation. Finally, we update the whole model by optimizing Eq. (2). While in the incremental task, images $\boldsymbol{x}$ and prompts consist of category labels belonging to $\mathcal{D}_t$. What's more, we only utilize the unexploited part of channels and hidden states in the previous task $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{t-1}\}$ to learn task-aware language-image representation. For the inference phase (the bottom part of Figure 2), the test image $\boldsymbol{x}$ belongs to all learned categories, and prompts consist of all learned category labels. The TAM will produce two groups of selected channels and hidden states. After being fed into a deep fusion module, two groups of alignment scores are obtained. Finally, we propose a Selection Inference Strategy to unify these alignment scores for final prediction.

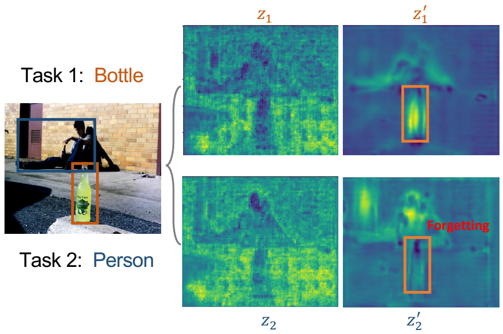

Figure 4: A visual illustration of where forgetting occurs, the yellow part of features with high activation. $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are the image features extracted by vision backbone which belong to Task 1 and Task 2 respectively, while $\boldsymbol{z}_1'$ and $\boldsymbol{z}_2'$ are the image features obtained by deep fusion module which belong to Task 1 and Task 2 respectively.

Task-Aware Module. The proposed task-aware module serves two purposes. On the visual side, we select different channels to learn task-aware visual representation to avoid reuse between different tasks. On the linguistic side, different hidden states are used to learn task-specific textual representation, hence the powerful representation ability of the pre-trained model can be applied to different tasks adaptively without interference from other tasks.

We denote two modal (image and text) masks as $\mathcal{M}_t^{image} \in \{0,1\}^{1\times 1\times c}$ and $\mathcal{M}_t^{text} \in \{0,1\}^{1\times l}$, where c and l represent the total number of channels for image feature $\boldsymbol{z}$ and hidden states $\boldsymbol{w}$ for text embedding, respectively. Then, we select partial channels of $\boldsymbol{z}$ and partial hidden states of $\boldsymbol{w}$ with the corresponding masks for learning the task-aware language-image representation. To do element-wise multiplication between representations and masks, we have to expand the dimension of mask $\mathcal{M}_t^{image}$ and $\mathcal{M}_t^{text}$ to the same spatial resolution of $\boldsymbol{z}$ and $\boldsymbol{w}$ as shown in the TAM of Figure 2. Formally, we have:

$$\hat{\boldsymbol{z}} = \boldsymbol{z}\mathcal{M}_t^{image}, \hat{\boldsymbol{w}} = \boldsymbol{w}\mathcal{M}_t^{text}, \quad (3)$$

where $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{w}}$ will be used to learn task-aware representa-

Setting	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
70+10	SID (Peng et al. 2021)	32.8	49.0	35.0	17.1	36.9	44.5
	ERD (Feng, Wang, and Yuan 2022)	34.9	51.9	37.4	18.7	38.8	45.5
	$*$CL-DETR (Liu et al. 2023b)	37.6/40.1	56.5/57.8	39.4/43.7	20.5/23.2	39.1/43.2	49.9/52.1
	Ours	**42.9**	**59.2**	**45.2**	**24.3**	**45.1**	**54.1**
60+20	SID (Peng et al. 2021)	32.7	49.8	34.6	17.2	37.6	43.5
	ERD (Feng, Wang, and Yuan 2022)	35.8	52.9	38.4	20.6	39.4	46.5
	Ours	**38.9**	**55.3**	**42.2**	**22.2**	**42.6**	**53.3**
50+30	SID (Peng et al. 2021)	33.8	51.0	36.1	17.6	38.1	45.1
	ERD (Feng, Wang, and Yuan 2022)	36.6	54.0	38.9	19.4	40.4	48.0
	Ours	**41.2**	**58.5**	**44.8**	**23.0**	**45.4**	**57.2**
40+40	SID (Peng et al. 2021)	34.0	51.4	36.3	18.4	38.4	44.9
	ERD (Feng, Wang, and Yuan 2022)	36.9	54.5	39.6	21.3	40.4	47.5
	$*$CL-DETR (Liu et al. 2023b)	37.0/37.5	56.2/55.1	39.1/40.3	20.9/20.9	38.9/40.8	49.2/50.7
	Ours	**40.4**	**57.4**	**43.9**	**23.3**	**44.7**	**54.5**

Table 1: Incremental results (%) based on our detector on COCO benchmark under different scenarios, $*$ indicates CL-DETR's two detection baseline UP-DETR (Dai et al. 2021b)/Deformable DETR (Zhu et al. 2021) and the other compared results are borrowed from the ERD (Feng, Wang, and Yuan 2022).

tion for the current task. In this way, the future tasks are able to adopt completely independent image-text representations, i.e, $\boldsymbol{z}(1-\mathcal{M}_t^{image})$ and $\boldsymbol{w}(1-\mathcal{M}_t^{text})$, which is no overlap with the previous tasks. We expect to alleviate catastrophic forgetting via learning the task-aware representation.

Selective Inference Strategy. Given test images and prompts, we first extract their feature map $\boldsymbol{z}$ and text embedding $\boldsymbol{w}$ by visual and language backbone respectively. Then, $\mathcal{M}_t^{image}$ and $\mathcal{M}_t^{text}$ are used to select channels $\boldsymbol{z}\mathcal{M}_t^{image}$ and hidden states $\boldsymbol{w}\mathcal{M}_t^{text}$ that have been trained in different tasks. After that, deep fusion is used to produce task-aware language-image representation $\boldsymbol{z}'$ and $\boldsymbol{w}'$. Finally, a set of alignment scores $s_t^{A\times O}$ that focus on different tasks respectively for an image region is calculated as:

$$s_t^{A\times O} = \boldsymbol{z}'(\boldsymbol{w}')^{\intercal}, \tag{4}$$

where A is the amount of image regions, and O is the amount of all learned categories.

Please refer to Figure 3 for the graphic illustration of our Selective Inference Strategy, where we assume the total of image regions and tasks both as two, and each task includes two categories for simplicity. The s_1 is produced via using $\mathcal{M}_1^{image}$ and $\mathcal{M}_1^{text}$, and the same for s_2. The simple solution is to directly unify the maximum/average alignment scores to generate a final prediction score:

$$s_{max} = ELEmax(s_1, s_2), \tag{5}$$

$$s_{mean} = ELEmean(s_1, s_2), \tag{6}$$

where *ELEmax* is the element-wise maximum operation and *ELEmean* is the element-wise average operation. Since there is no overlap in the prompts used in the different tasks, Eq. (5) and Eq. (6) can consequently make each image region more prone to be assigned with a series of false categories, *i.e.,* False Positive predictions, and thus result in poor predictions. The Selective Inference Strategy is proposed to solve this dilemma, shown in the $ROWmax$ part of Figure 3, for any alignment scores s, *e.g.,* s_1, s_2, only the portion produced via task-aware representation is used, *i.e.,* s_1', s_2'. Finally, we unify these task-specific alignment scores by:

$$s_u = ROWmax(s_1', s_2'), \tag{7}$$

where *ROWmax* is the row-wise maximum operation. Specifically, we use the maximum confidence prediction between s_1' and s_2' as the final prediction.

Experiments

Datasets and Evaluation Metrics. Existing methods prove the validity of the method in two dataset settings, one using Pascal VOC 2007 and Microsoft COCO 2014, and the other using only Microsoft COCO2017. In order to better prove the validity of our method, the proposed method is evaluated on two benchmark datasets, *i.e.,* Pascal VOC 2007 and Microsoft COCO 2017. VOC 2007 has 20 object classes, and we use the trainval subset for training and the test subset for evaluation, the mean average precision (mAP) at 0.5 IoU threshold is used to measure the performance. We ensure consistency between data partitioning methods and CIOD (Dong et al. 2023) for VOC 2007. COCO 2017 has 80K images in the training set and 40K images in the validation set for 80 object classes, we use the train set for training and the minival set for testing, and the standard COCO protocols are used as the evaluation metrics, *i.e.,* AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. We ensure consistency between data partitioning methods and ERD (Feng, Wang, and Yuan 2022) for COCO 2017.

Method	19+1				15+5				10+10			
	1-19	20	1-20	Avg	1-15	16-20	1-20	Avg	1-10	11-20	1-20	Avg
Upper	77.2	80.3	77.4	78.8	78.8	73.0	77.4	75.9	77.2	77.5	77.4	77.4
Fine-tuning	61.4	62.8	61.5	62.1	46.6	59.2	40.9	52.9	33.0	68.1	50.6	50.6
LOD (Zhou et al. 2020)	70.5	53.0	69.6	61.8	-	-	-	-	-	-	-	-
Meta (Joseph et al. 2021b)	70.9	57.6	70.2	64.3	71.7	55.9	67.8	63.8	68.3	64.3	66.3	66.3
MVCD (Yang et al. 2022)	70.2	60.6	69.7	65.4	69.4	57.9	66.5	63.7	66.2	66.0	66.1	66.1
CIOD (Dong et al. 2023)	70.3	65.3	70.1	67.8	71.4	57.5	67.9	64.5	69.8	64.4	67.1	67.1
Ours	**73.2**	**66.5**	**72.9**	**69.9**	**73.6**	**60.2**	**70.3**	**66.9**	**71.2**	**70.0**	**70.6**	**70.6**

Table 2: mAP@0.5% results on single incremental step on Pascal-VOC 2007, all compared results are borrowed from the corresponding papers.

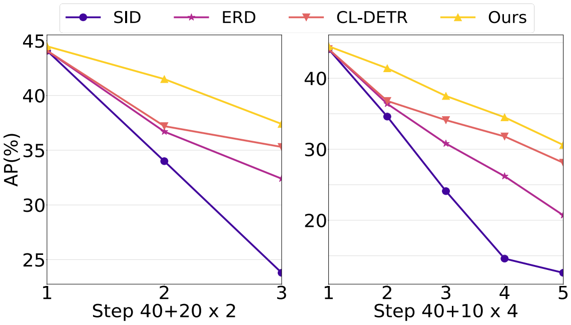

Figure 5: Incremental results (mAP%) on COCO 2017 dataset under different scenarios, all compared results are borrowed from the corresponding papers.

TAM Strategies		1-10	11-20	1-20
1-10	11-20			
first 75%	last 25%	68.0	71.9	70.0
first 25%	last 75%	67.6	72.0	69.9
random 50%	rest 50%	69.7±1.0	71.0±1.1	70.4±1.0
first 50%	last 50%	71.2	70.0	70.6

Table 3: mAP@0.5% results on different feature map channels and text embedding hidden states selection strategies on VOC 2007 dataset with 10+10 setting.

Experiments Setup. Specifically, we conduct experiments with different splits in the following class-incremental learning scenarios. **One-step:** we notate this setup as $B + I$, *i.e., Base + Incremental*. we observe a fraction $\frac{B}{B+I}$ of the training samples with B categories annotated in the first step. Then, in the second step, we observe the remaining $\frac{I}{B+I}$ of the training samples, where I new categories are annotated. Four settings for the COCO 2017 dataset, *i.e.,* $B + I$ = 40+40, 50+30, 60+20, 70+10 and three settings for VOC 2007 dataset, *i.e.,* $B + I$ = 19 + 1, 15 + 5, 10 + 10. **Multi-step:** we notate this setup as $B + I \times N$, where N is the incremental number of times. For the COCO 2017 dataset, two-step, and four-step settings with 20 and 10 new classes respectively added each time, *i.e.,* $B + I \times N = 40 + 20 \times 2$ and $40 + 10 \times 4$. We run each experiment three times in different random orders of categories and report the average mAP.

Implementation Details. We build our method on GLIP, which uses Swin-Tiny (Liu et al. 2021) with FPN (Lin et al. 2017a) and BERT (Kenton and Toutanova 2019) as visual and language backbone respectively. All the experiments are performed on 8 NVIDIA Tesla V100 GPUs, with a batch size of 16, we use the ADAMW as the optimizer with the learning rate of the language backbone is 5×10^{-6} and other parts are 5×10^{-5}. For the usage of feature map channels and text embedding hidden states, we divided them according to the proportion of categories, for example, the amount of channels and the amount of hidden states used in the two tasks are 50% and 50% respectively in the VOC 2007 dataset with the 10+10 setting.

Overall Performance

For the COCO 2017, we compared with CL-DETR (Liu et al. 2023b), ERD (Feng, Wang, and Yuan 2022), SID (Peng et al. 2021), and LwF (Li and Hoiem 2017), while for the VOC 2007, we compared with LOD (Zhou et al. 2020), Meta (Joseph et al. 2021b), MVCD (Yang et al. 2022), and CIOD (Dong et al. 2023). The total of the above methods is based on visual-only detectors.

One-step. For the COCO 2017, Figure 1 demonstrates the performance under four settings, all of our method outperforms the current state-of-the-art CL-DETR and other class-incremental object detection methods. For 70+10 and 40+40 settings, the AP of our method is 2.8 and 2.9 percentage points higher than CL-DETR, respectively; for 60+20 and 50+30 settings, on which CL-DETR has not experimented, the AP of our method is 3.1 and 4.6 percentage points higher than ERD, respectively. Table 2 shows the experimental results on the VOC 2007, the Avg metric equally weights new and old classes averaging their aggregated mAP. Under the three experimental settings of 19+1,15+5,10+10, our mAP is 2.8, 2.4, 3.5 percentage points higher than the state-of-the-art method, and outperforms the CIOD on both the new and old classes. All analysis above illustrates that the proposed method can effectively overcome background-foreground conflict even in these challenging settings.

Lines	VB	LB	Ave	Max	Sel	Params	1-10	11-20	1-20
1	Fine-tuning (Baseline)					232M	33.0	68.1	50.6
2			✓			464M	42.9	69.7	56.3
3				✓		464M	42.2	71.0	56.6
4					✓	464M	71.9	72.5	72.2
5	✓				✓	232M	70.2	70.0	70.1
6		✓			✓	232M	35.2	71.0	53.1
7	✓	✓	✓			232M	38.8	65.8	52.3
8	✓	✓		✓		232M	33.4	68.3	50.9
9	✓	✓			✓	232M	71.2	70.0	70.6

Table 4: mAP@0.5% results on VOC 2007 dataset with 10+10 setting. VB and LB indicate the vision branch and language branch respectively. Ave, Max, and Sel indicate three inference strategies, *i.e.,* Eq. (5), Eq. (6), and Eq. (7).

Multi-step. Figure 5 shows the AP in $40+20\times2$ and $40+10\times4$ settings on the COCO 2017. The AP for our first phase and related are 44.5 and 44.1, respectively. Compared with the state-of-the-art method CL-DETR, the AP of our final stage improved by 2.1 and 2.5 percentage points respectively at $40+20\times2$ and $40+10\times4$ settings. This fully demonstrates that our method is stable and can still maintain its ability to mitigate catastrophic forgetting under different scenarios.

Ablation Study

We validate the effectiveness of the various parts of our method on the VOC 2007 dataset, and all experiments are performed in the 10+10 setting.

Sensitivity of Hyper-parameters Effectiveness of TAM and SIS. Table 4 illustrates the effectiveness of using the channel selection strategy on different branches and different inference strategies. Shown in lines 2-4, for each task, we utilize an alone model to learn independent representation and ensemble all predictions by Eq. (6), Eq. (5), and Eq. (7), this is likely Der (Dai et al. 2021a). We find that our inference strategy Eq. (7) is effective. Compared to line 4 with line 9, although the previous one gets better performance, the model parameters are twice as much as our method, with the increase of incremental tasks, the model parameters grow linearly, resulting in a huge storage burden.

Lines 5-6 and 9 illustrate the effectiveness of the TAM on different branches. When TAM is used in the language branch (LB) only, there is an improvement (line 6) in the old classes' performance compared to directly fine-tuning (line 1). There is a huge improvement when applying TAM only in the visual branch (line 5). The results demonstrate that our method reinforces the separability between image features and language representation, and has effectively solved the background-foreground conflict problem of class-incremental object detection.

Line 7 and line 8 use Eq. (6) and Eq. (5), respectively, to directly average or maximize the alignment scores, which produces a lot of false negative predictions due to the confusion of alignment between tasks, making it very ineffective. Line 9 uses Eq. (7) and achieves the peak performance by selectively using the alignment score strategy.

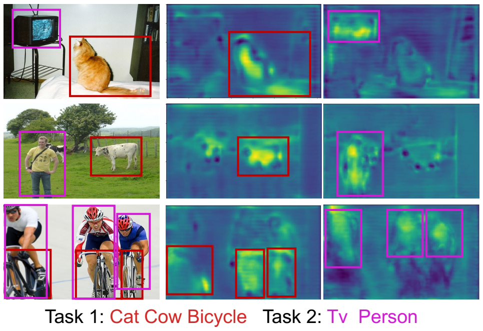

Figure 6: Visualisation of the VOC 2007 dataset under 10+10 setting. The first column is the original image, the second and third columns are feature maps using $\{\mathcal{M}_1^{image}, \mathcal{M}_1^{text}\}$ and $\{\mathcal{M}_2^{image}, \mathcal{M}_2^{text}\}$ respectively.

Analyze of Selection Strategies. We made four different selection strategies for image feature map channels and text embedding hidden states (shown in Table 3). The first and second lines use 75% and 25% of the channels and hidden states in the first task, respectively, and we find that the performance of the old and new classes is related to the amount of used channels and hidden states, so we divided the amount used in the different tasks according to the amount of classes to achieve the best results (line 4). Specifically, in line 3, when we randomly select the channels and hidden states (the channels and hidden states are not all consecutive), the results do not differ much from those using consecutive ones, which demonstrates the generalisability of our method.

Visualized Analysis

We conduct a visual analysis of the feature maps of the images after deep fusion, and Figure 6 illustrates the results. The first column displays the original image, while the second column shows the feature map obtained after deep fusion using $\mathcal{M}_1^{image}$ and $\mathcal{M}_1^{text}$. The third column represents the feature map obtained using $\mathcal{M}_2^{image}$ and $\mathcal{M}_2^{text}$. From the visualization, it is evident that our task-aware language-image learning method effectively segregates the categories of different tasks in the feature maps. It focuses solely on the regions specific to each task, ensuring task-specific information is captured accurately.

Conclusion

In this paper, we implement the first application of a visual-language detector for class incremental object detection. The language-image detector is found to have a better ability to mitigate catastrophic forgetting when there are fewer categories, which fails when there are more categories due to increased task alignment confusion. To this end, we propose to learn task-aware language-image representation to segregate visual feature map channels and text embedding hidden states for different tasks. State-of-the-art results are achieved on both VOC 2007 and COCO 2017 benchmark datasets, demonstrating the effectiveness of our approach.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2021ZD0111000); National Natural Science Foundation of China (No.62222602, 62302167, U23A20343, 62176092, 62106075, 62176224), Shanghai Science and Technology Commission (No.21511100700), Development Project of Ministry of Industry and Information Technology (ZTZB-23-990-016), Shanghai Sailing Program under Grant (23YF1410500), Natural Science Foundation of Shanghai (23ZR1420400), Natural Science Foundation of Chongqing (CSTB2023NSCQ-JQX0007, CSTB2023NSCQ-MSX0137), and CCF-Tencent Rhino-Bird Young Faculty Open Research Fund (RAGR20230121).

References

Cai, T.; Zhang, Z.; Tan, X.; Qu, Y.; Jiang, G.; Wang, C.; and Xie, Y. 2023. Multi-Centroid Task Descriptor for Dynamic Class Incremental Inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7298–7307.

Cermelli, F.; Geraci, A.; Fontanel, D.; and Caputo, B. 2022. Modeling missing annotations for incremental learning in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3700–3710.

Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; and Zhang, L. 2021a. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7373–7382.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021b. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1601–1610.

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 conference on computer vision and pattern recognition*, 248–255.

Dong, N.; Zhang, Y.; Ding, M.; and Bai, Y. 2023. Class-incremental object detection. *Pattern Recognition*, 139: 109488.

Feng, T.; Wang, M.; and Yuan, H. 2022. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9427–9436.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Guan, L.; Wu, Y.; Zhao, J.; and Ye, C. 2018. Learn to detect objects incrementally. In *Proceedings of 2018 IEEE Intelligent Vehicles Symposium*, 403–408.

Hu, X.; Tang, K.; Miao, C.; Hua, X.-S.; and Zhang, H. 2021. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 3957–3966.

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.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Li, J.; Xu, R.; Ma, J.; Zou, Q.; Ma, J.; and Yu, H. 2023a. Domain adaptive object detection for autonomous driving under foggy weather. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 612–622.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, W. L.; Gao, B.-B.; Xia, B.; Wang, J.; Liu, J.; Liu, Y.; Wang, C.; and Zheng, F. 2023b. Cross-Modal Alternating Learning with Task-Aware Representations for Continual Learning. *IEEE Transactions on Multimedia*.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Proceedings of Advances in Neural Information Processing Systems*, 21002–21012.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017b. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, L.; Kuang, Z.; Chen, Y.; Xue, J.-H.; Yang, W.; and Zhang, W. 2020. Incdet: In defense of elastic weight consolidation for incremental object detection. *IEEE transactions on neural networks and learning systems*, 32(6): 2306–2319.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, Y.; Schiele, B.; Vedaldi, A.; and Rupprecht, C. 2023b. Continual detection transformer for incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23799–23808.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Peng, C.; Zhao, K.; Maksoud, S.; Li, M.; and Lovell, B. C. 2021. SID: Incremental learning for anchor-free object detection via Selective and Inter-related Distillation. *Computer vision and image understanding*, 210: 103229.

Petit, G.; Popescu, A.; Schindler, H.; Picard, D.; and Delezoide, B. 2023. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3911–3920.

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 *Proceedings of International conference on machine learning*, 8748–8763.

Shieh, J.-L.; Haq, Q. M. u.; Haq, M. A.; Karam, S.; Chondro, P.; Gao, D.-Q.; and Ruan, S.-J. 2020. Continual learning strategy in one-stage object detection framework based on experience replay for autonomous driving vehicle. *Sensors*, 20(23): 6777.

Shmelkov, K.; Schmid, C.; and Alahari, K. 2017. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, 3400–3409.

Tian, X.; Zhang, Z.; Tan, X.; Liu, J.; Wang, C.; Qu, Y.; Jiang, G.; and Xie, Y. 2023. Instance and Category Supervision are Alternate Learners for Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5596–5605.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9627–9636.

Wang, F.-Y.; Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2022. Foster: Feature boosting and compression for class-incremental learning. In *Proceedings of European conference on computer vision*, 398–414.

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.

Xu, G.; Khan, A. S.; Moshayedi, A. J.; Zhang, X.; and Shuxin, Y. 2022. The Object Detection, Perspective and Obstacles In Robotic: A Review. *EAI Endorsed Transactions on AI and Robotics*, 1(1).

Xu, Y.; Zhang, M.; Fu, C.; Chen, P.; Yang, X.; Li, K.; and Xu, C. 2023. Multi-modal Queried Object Detection in the Wild. In *Proceedings of Neural Information Processing Systems*.

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

Yang, Y.; Zhou, D.-W.; Zhan, D.-C.; Xiong, H.; Jiang, Y.; and Yang, J. 2021. Cost-effective incremental deep model: Matching model capacity with the least sampling. *IEEE Transactions on Knowledge and Data Engineering*, 35: 3575–3588.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18123–18133.

Zhao, H.; Wang, H.; Fu, Y.; Wu, F.; and Li, X. 2021. Memory-efficient class-incremental learning for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10): 5966–5977.

Zhao, Z.; Zhang, Z.; Tan, X.; Liu, J.; Qu, Y.; Xie, Y.; and Ma, L. 2023. Rethinking Gradient Projection Continual Learning: Stability/Plasticity Feature Space Decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3718–3727.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

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

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable detr: Deformable transformers for end-to-end object detection. In *Proceedings of International Conference on Learning Representations*.