GCD: Advancing Vision-Language Models for Incremental Object Detection via Global Alignment and Correspondence Distillation

Xu Wang, Zilei Wang*, Zihan Lin

University of Science and Technology of China, Hefei, China
xu-wang@mail.ustc.edu.cn, zlwang@ustc.edu.cn, myustc@mail.ustc.edu.cn

Abstract

Incremental object detection (IOD) is a challenging task that requires detection models to continuously learn from newly arriving data. This work focuses on incremental learning for vision-language detectors (VLDs), an under explored domain. Existing research typically adopts a local alignment paradigm to avoid label conflicts, where different tasks are learned separately without interaction. However, we reveal that this practice fails to effectively preserve the semantic structure. Specifically, aligned relationships between objects and texts would collapse when handling novel categories, ultimately leading to catastrophic forgetting. Though knowledge distillation (KD) is a common approach for tackling this, traditional KD performs poorly when directly applied to VLDs, as for different phases, a natural knowledge gap exists in both encoding and decoding processes. To address above issues, we propose a novel method called Global alignment and Correspondence Distillation (GCD). Differently, we first integrate knowledge across phases within the same embedding space to construct global semantic structure. We then enable effective knowledge distillation in VLDs through a semantic correspondence mechanism, ensuring consistent proposal generation and decoding. On the top of that, we distill teacher model's informative predictions and topological relationships to maintain stable local semantic structure. Extensive experiments on COCO 2017 demonstrate that our method significantly outperforms existing approaches, achieving new state-of-the-art in various IOD scenarios.

Code — https://github.com/Never-wx/GCD

Introduction

Typically, object detection models follow fully supervised learning paradigm, requiring the network to learn from annotated data within a predefined label space. This paradigm generally assumes that the data distribution is fixed and stationary, while in real-world applications, data often arrives continuously in a non-stationary manner (Feng, Wang, and Yuan 2022). Directly fine-tuning the model on new coming data will lead to a significant performance decline on old tasks, which is known as catastrophic forgetting (Goodfellow et al. 2013; McCloskey and Cohen 1989). To enable models to continuously learn new knowledge, the incremental learning paradigm has been proposed. In this paper, we mainly focus on class-incremental learning on object detection, namely incremental object detection(IOD).

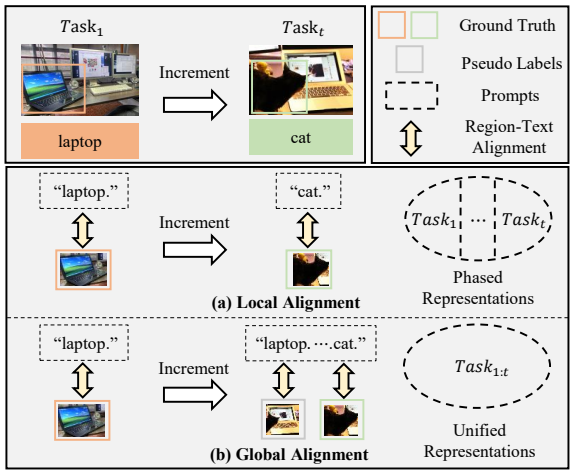


Figure 1: Existing research on VLDs employs local alignment paradigm, which focuses on incremental learning within a localized label space. This approach typically results in phased vision-language representations. In contrast, our method emphasizes global alignment, enabling the maintenance of cohesive and unified representations.

In incremental object detection (IOD), prior research primarily addresses catastrophic forgetting using knowledge distillation (KD). For instance, (Feng, Wang, and Yuan 2022) distills classification and regression responses separately, while (Liu et al. 2023c) combines pseudo-labels and exemplar replay for DETR-based detectors. However, these studies mainly focus on vision detectors. Concurrently, recent research trends in class incremental learning have increasingly centered on vision-language models (VLMs) (Zheng et al. 2023; Zhou et al. 2023). Following this, a pioneering work (Zhang et al. 2024) explored the application of GLIP (Li et al. 2022) in IOD, a vision-language detector (VLD) that reformulates object detection as phrase grounding. They propose to address background-foreground conflicts through local alignment, where only objects and

*Corresponding author.

texts of current task are aligned, as shown in Fig. 1 (a). However, despite avoiding label conflicts, catastrophic forgetting still occurs and they attempt to mitigate it through parameter isolation. To understand why VLDs forget even without label conflicts, we probe into the embedding space and find the problem lies in the semantic structure($e.g.$ the topological relationships of text features and object features). More details in Fig. 4. This approach fails to effectively preserve the local semantic structure of old knowledge. Additionally, learning knowledge from different phases in disjoint embedding spaces, with no interaction across phases, hinders the formation of global semantic structure and ultimately weakens the potential of vision-language representations.

We assert that maintaining a robust semantic structure is essential for overcoming catastrophic forgetting in vision-language detectors. A fixed semantic structure lacks plasticity, while a loose one lacks stability. Therefore, we develop a flexible global semantic structure to integrate new knowledge, while maintain a stable local semantic structure to preserve old knowledge. For the global aspect, we propose integrating old and new knowledge within the same embedding space through global alignment, as shown in Fig. 1 (b). This process involves contrastive learning over both old and new samples to shape a global semantic structure. For the local aspect, KD is typically used in IOD to preserve the local semantic structure of old knowledge. However, directly applying traditional KD methods (Li and Hoiem 2017; Feng, Wang, and Yuan 2022) performs poorly in VLDs. We attribute this to two reasons. First, the disruption of negative samples remains a core challenge. Additionally, text involvement in the detection process introduces new complexities. The teacher model detects old objects using old prompts, while the student learns new knowledge with new or combined prompts, leading to inconsistencies in encoding and decoding processes. This discrepancy results in different proposals and final predictions, making direct knowledge distillation impractical.

To address these challenges, we propose a novel method called Global Alignment and Correspondence Distillation (GCD). GCD consists of a global pipeline and a local pipeline. In the global pipeline, we shape the global semantic structure by employing Global Alignment, where new knowledge is supervised by GT and old knowledge by pseudo-labels. In the local pipeline, we introduce a semantic correspondence mechanism (SCM). It includes a shared query to generate consistent proposals which are then combined with chunked text token to ensure a consistent decoding process. On this basis, we leverage teacher's responses to mitigate the overconfidence of noisy pseudo and preserve the activation of weak categories, termed Correspondence Response Distillation (CRD). Additionally, to maintain local semantic structure at feature level, we distill the teacher's relational topology of text and object prototypes to the student, termed Correspondence Topology Distillation (CTD).

Our contributions are as follows:

- We reveal that semantic structure collapse is the key factor leading to catastrophic forgetting in vision-language detectors (VLDs), which is crucial for exploiting VLDs' potential in IOD scenarios.
- We simultaneously develop a flexible global semantic structure via global alignment, while maintaining a stable local semantic structure through correspondence knowledge distillation, thus achieving a better balance between stability and plasticity.
- We conduct extensive experiments on COCO 2017 over various IOD settings. The results demonstrate that our method achieves new state-of-the-art performance.

Related Works

Incremental Learning

Incremental learning aims to develop artificially intelligent systems that can continuously learn to address new tasks from new data while preserving knowledge learned from previously learned tasks (Thrun 1995). Incremental image classification has been extensively studied. Current techniques can be mainly divided into knowledge distillation (KD) methods, exemplar replay (ER) methods. KD-based methods aim to preserve previous knowledge through logits distillation (Li and Hoiem 2017; Hou et al. 2019; Wu et al. 2019) and intermediate features distillation (Dhar et al. 2019; Douillard et al. 2020). ER-based methods replay previous knowledge in the incremental steps to better overcome forgetting (Rebuffi et al. 2017; Iscen et al. 2020; Liu et al. 2020b; Zhu et al. 2021). Furthermore, incremental learning research has also been extended to other tasks (Douillard et al. 2021; Lin, Wang, and Zhang 2022, 2023; Peng, Zhao, and Lovell 2020; Peng et al. 2021; Cermelli et al. 2022)

Incremental Object Detection

IOD is more complex than incremental classification given the occurrence of instances of classes that are unknown at the time, but can appear in subsequent tasks as a new class to be learned, resulting in missing annotations and conflicts with the background label (Menezes et al. 2023). RILOD (Shmelkov, Schmid, and Alahari 2017) performed proposal distillation based on Faster-RCNN (Ren et al. 2015). SID (Peng et al. 2021) distilled selected intermediate features in anchor-free detector FCOS (Tian et al. 2020). OWOD (Joseph et al. 2021) combined IOD with open world demands and propose a challenging new setting. Based on GFL (Li et al. 2020), ERD (Feng, Wang, and Yuan 2022) separately distilled classification responses and regression responses. Also, (Liu et al. 2020a, 2023b) revised sampling and replay strategies to conduct more efficient rehearsal. More recently, IOD has been extended to DETR-based (Carion et al. 2020) detectors. OW-DETR (Gupta et al. 2022) proposed a attention-driven pseudo labeling to discover potential objects. CL-DETR (Liu et al. 2023c) managed to make KD and ER compatible with DETR-based detectors. (Kang et al. 2023) took advantage of high-level semantic information to better preserve discriminativeness.

Incremental Learning for VLMs

Recent advances in Vision-Language Models (VLMs) (Radford et al. 2021; Jia et al. 2021; Yu et al. 2022) have shown promising capabilities in learning generalizable representations with the aid of textual information. Particularly,

CLIP (Radford et al. 2021) has demonstrated promising zero-shot performance on different downstream vision tasks. GLIP (Li et al. 2022) extended CLIP and reformulated object detection as phrase grounding, creating a new paradigm for object detection tasks. Grounding Dino (Liu et al. 2023a) further integrates detection transformers with grounding. In incremental classification field, (Zheng et al. 2023) observed that VLMs suffer from catastrophic forgetting when continually trained with new classes. (Zhu et al. 2023) introduced a new task Vision Language Continual Pretraining and proposed a compatible topology preservation method to flexibly update model. (Zhou et al. 2023) designed expandable projections to incrementally align visual and textual information. And in IOD, (Zhang et al. 2024) leveraged GLIP as baseline. To circumvent background-foreground conflict, they attempted to learn new objects in local alignment manner. However, they observed that catastrophic forgetting still occurs. They asserted forgetting occurs in the parameter space and that can be mitigated through parameter isolation.

Preliminaries

Grounding Dino

Grounding Dino (Liu et al. 2023a) is phrase grounding based vision-language detector which aligns objects with phrases in a text prompt. As shown in Fig. 2 (a), it includes a vision backbone f_v and language backbone f_l for extracting vanilla features. Given a pair of image-text data(I, T), the feature enhancer f_e for cross modality feature fusion will further align vanilla features, which can be defined as:

$$V, W = f_e(f_v(I)), f_l(T)), \tag{1}$$

After alignment, the fusion image embedding and text embedding are denoted as V, W respectively. The alignment scores can be defined as cosine similarity $S(V_i, W_j) = \frac{v_i \cdot w_j^\top}{\|v_i\| \|w_j\|}$. A language guided query selection module initialize a group of references according to $S(V, W)$ which is the positional part of learnable object queries. A cross-modality decoder will refine these object queries to produce high-semantic queries $Q \in \mathbb{R}^{N \times D}$, where N is the number of queries, D is the dimension of queries. The output $\hat{y} = (\hat{y}_i)_{i \in \mathcal{N}}$ is a sequence $\mathcal{N} = \{1, \ldots, N\}$ of object predictions $\hat{y}_i = (\hat{s}_i, \hat{b}_i)$ including logits and bounding boxes, where $\hat{s}_i = S(q_i, W)$ and $\hat{b}_i$ is decoded by the regression head. In the training stage, each ground truth will be assigned to a query through minimizing the matching cost between model predictions and ground truth. The optimal matching is solved as follows:

$$\hat{\sigma} = arg\,min_\sigma \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(\hat{y}_{\sigma_i}, y_i), \tag{2}$$

where σ_i is a permutation of N elements and $\hat{\sigma}$ is the optimal assignment. $y_i = (s_i, b_i)$ is i-th GT. $\mathcal{L}_{\text{match}}$ is a pair-wise matching cost defined as:

$$\mathcal{L}_{\text{match}}(\hat{y}_{\sigma_i}, y_i) = \mathcal{L}_{\text{align}}(\hat{s}_{\sigma_i}, s_i) + \mathcal{L}_{\text{reg}}(\hat{b}_{\sigma_i}, b_i), \tag{3}$$

The number of objects in an image is usually fewer than N. Thus, predictions not assigned to GT will be treated as negative samples. The corresponding alignment targets for negative samples are padded to zero noted as ϕ . Then, the detr loss can be defined as follows:

$$\mathcal{L}_{\text{detr}}(\hat{y}, y) = \sum_{i=1}^{N} \mathcal{L}_{\text{align}}(\hat{s}_{\sigma_i}, s_i) + \mathbf{1}_{s_i \neq \phi} \mathcal{L}_{\text{reg}}(\hat{b}_{\hat{\sigma}_i}, b_i), \tag{4}$$

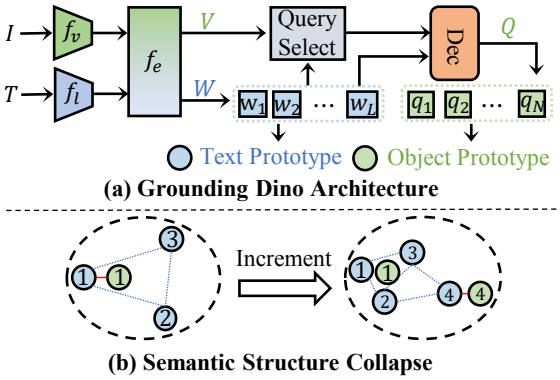


Figure 2: A Simple illustration of (a) Grounding Dino Architecture (b) Semantic Structure Collapse.

Forgetting in Grounding Dino

We take Fig. 2 (b) to intuitively understand semantic structure collapse. Assume a good embedding space is constructed in phase 1. The object prototype p_1 is close to corresponding text prototype $\hat{p}_1$ while keep a relatively far distance with other text prototype. In phase 2, only new objects are aligned with new texts while others are ignored. During training, $(p_4, \hat{p}_4)$ are pulled as close as possible, and the entire embedding space are shifting because of the pulling effect. As a result, other ignored prototypes are squeezed together and their original semantic relation can't sustain. More detailed visualization will be provided in experiments.

Method

IOD Definition

In IOD, the training procedure is composed of a series of Z learning phase. At each phase t where $t \in \{1, \ldots, Z\}$, a set of new categories C_t are introduced with a training set $D_t = \{(x_n, y_n)\}$ where x_n are images and y_n are the corresponding labels, n represents total number. The annotated categories of different phases are disjoint. In phase t, only class C_t of current task will be labeled. The detector is trained sequentially on each phase t, where it learn new objects of C_t according to D_t. After each training phase, the detector should be capable of detecting objects of all learned classes $C_{1:t} = C_{1:(t-1)} \cup C_t$. In the case of vision-language detector, the categories are transformed to text prompts and we denotes prompts of current task as $Prompts_t$ and $Prompts_{1:(t-1)}$ for previous tasks.

Overview

We take Grounding-Dino-T as our baseline model, and we don't use O365 pretrained weights for fair compari-

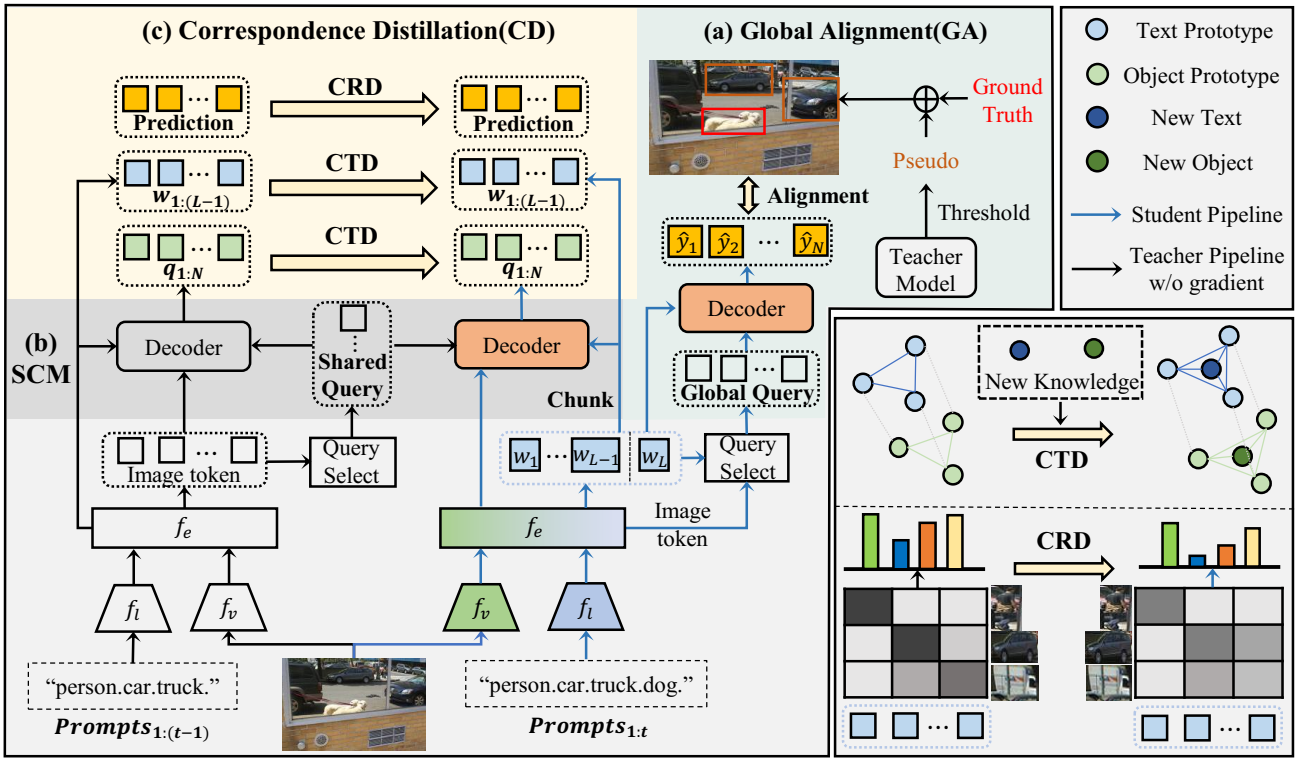


Figure 3: Overview of GCD framework. It consists of a global pipeline and a local pipeline. For global pipeline, the student leverages full text token $w_{1:L}$ to make predictions on global label space and perform Global Alignment(GA). For local pipeline, based on Semantic Correspondence Mechanism(SCM), chunked text tokens $w_{1:(L-1)}$ are used to carry out CRD and CTD.

son. In phase 1, we initialize vision backbone and language backbone with weights pretrained on ImageNet-1K and BERT$_{base}$. Then, we fine-tune the model with data D_1. As for incremental phase, new model will be initialized with weights trained in last phase.

The total framework of our method is shown in Fig. 3 which can be divided into: (a) Global Alignment for shaping global semantic structure. (b) Semantic Correspondence Mechanism (SCM) for enabling effective KD. (c) Correspondence Response Distillation (CRD) for preserving responses of weak categories and Correspondence Topology Distillation (CTD) for maintaining local semantic structure.

Global Alignment

Existing research (Zhang et al. 2024) focuses on learning new knowledge through local alignment, using parameter isolation to tackle catastrophic forgetting. In this approach, only objects and texts associated with current classes C_t are aligned, which fails to fully realize the potential of VLDs. In contrast, our method applies global supervision to both old and new knowledge through global alignment, where objects and texts across the entire label space $C_{1:t}$ are aligned within same embedding space. For new knowledge, supervision is provided by ground truth, ensuring that new objects are associated with their respective new texts and are distinct from others. However, label conflicts can cause detected old objects to be misaligned with all text representations, potentially leading to a collapse. As a solution, we employ pseudo labels as supervision for old knowledge, following previous studies (Gupta et al. 2022; Liu et al. 2023c).

In our framework, the teacher receives $Prompts_{1:(t-1)}$ to detect objects of previous tasks while the student uses $Prompts_{1:t}$ to predict across the entire label space $C_{1:t}$. The teacher's output predictions, including logits and regression results, are denoted as $\hat{y}^{old} = \{\hat{s}^{old}, \hat{b}^{old}\}, \hat{s}^{old} \in \mathbb{R}^{N\times(L-1)}$, where N is the number of queries and $L-1$ represents the length of the old text tokens. We apply a global threshold on the logits to select the most confident predictions, which are then transformed into one-hot pseudo labels and merged with the ground truth, resulting in a total label set $y = (y_i)_{i\in n}$, where n represents the total number of labels. A global bipartite matching is performed to align these labels with the student's predictions, and the detection loss, as defined in Eq. (4) is applied to provide global supervision. Pseudo labels play a crucial role in resolving label conflicts, enabling global alignment integrating knowledge of different phases in same embedding space to shape a global semantic structure. However, pseudo label generation inevitably suffers from information redundancy and noise, with high-confidence objects selected as pseudo labels while low-confidence ones are ignored. This leads to well-trained categories dominating the update process, while weak categories are overlooked. Additionally, overconfident noisy pseudo labels can disrupt optimization, causing insta-

bility. To address these issues, we utilize the teacher's informative responses as guidance.

Semantic Correspondence Mechanism

Traditional response distillation (RD) methods (Feng, Wang, and Yuan 2022) depend on anchors to ensure knowledge distillation occurs at corresponding locations and classification heads can be directly divided for distillation. However, in our case, text is deeply integrated into the entire detection process, and serious errors can occur if these differences are overlooked. The teacher model detects old objects using old $Prompts_{1:(t-1)}$, while the student integrates new knowledge with full $Prompts_{1:t}$, resulting in different initialized object queries and an inconsistent decoding process. Consequently, output semantic queries lack spatial and semantic relationships between the teacher and student, making direct bipartite matching and distilling the teacher's responses impractical.

Drawing inspiration from recent advancements in knowledge distillation (Chang et al. 2023; Wang et al. 2024), we propose a semantic correspondence mechanism to ensure that the teacher and student produce corresponding predictions. We first introduce a shared query mechanism between the teacher and student, consisting of a content part and a positional part. The content part is initialized with the teacher's well-trained parameters, while the positional part is derived from the teacher's query selection results. This shared query enables both the teacher and student to generate corresponding initialized queries. In addition, we utilize the corresponding text tokens $w_{1:(L-1)}$ together with the corresponding initialized queries, as in the teacher model, to ensure a consistent decoding process, including coordinate refinement and semantic injection. Given that Grounding Dino uses sub-sentence-level text representations, we directly chunk the full text tokens $w_{1:L}$ to obtain local text tokens $w_{1:(L-1)}$ for compatibility with global alignment.

Correspondence Distillation

Correspondence Response Distillation. Based on semantic correspondence mechanism, we propose Correspondence Response Distillation (CRD) to transfer teacher's responses. With consistent decoding, the decoder produces a set of corresponding semantic queries, generating predictions denoted as $\{\hat{y}_i^{old}; \hat{y}_i\} = \{(\hat{s}_i^{old}, \hat{b}_i^{old}); (\hat{s}_i, \hat{b}_i)\}$, where $i \in \mathcal{N}$. For the alignment part, we apply KL-divergence with temperature scaling to transfer the teacher's soft probabilities to the student. The teacher's logits are transformed into probabilities as $\mathcal{P}_i^{old} = \mathrm{SoftMax}(\hat{s}_i^{old}/\tau)$, where τ is a temperature factor used to smooth the probability distribution, similarly applied to the student's logits. The alignment distillation loss is defined as follows:

$$\mathcal{L}_{CRD\_align} = \sum_{i=1}^{N} \alpha_i \mathcal{L}_{KL}(\mathcal{P}_i^{old}, \mathcal{P}_i) \tag{5}$$

where $\alpha_i = \max_{c \in C_{1:t-1}}(\hat{s}_i^{old}(c))$ represents the confidence of teacher's predictions. The information contained in the semantic queries is not uniform, where background responses may overshadow foreground responses if all responses are distilled indiscriminately. Queries with higher alignment scores are considered foreground predictions, which likely contain more valuable information for distillation and are given greater weight. For the regression part, we apply $\mathcal{L}_{reg}$, as defined in the detection loss, to ensure that the student outputs corresponding region predictions as the teacher.

$$\mathcal{L}_{CRD\_reg} = \sum_{i=1}^{N} \alpha_i \mathcal{L}_{reg}(\hat{b}_i^{old}, \hat{b}_i) \tag{6}$$

The total CRD loss can be defined as follows, where K represents the number of decoder layers, γ is the coefficient for the alignment component, and the coefficient for $\mathcal{L}_{CRD_{reg}}$ matches that of the regression component in detection loss.

$$\mathcal{L}_{CRD} = \sum_{k=1}^{K} \gamma \mathcal{L}_{CRD\_align}^{k} + \mathcal{L}_{CRD\_reg}^{k} \tag{7}$$

Correspondence Topology Distillation. Global alignment shapes a flexible global semantic structure and CRD ensures weak categories are not overlooked. However, these constraints are applied at the logits level, without directly enforcing consistency in feature-level relationships. The integration of new knowledge may alter the local semantic structure of the old due to the strong push-and-pull effect. To mitigate this, we propose Correspondence Topology Distillation (CTD). CTD imposes constraints on mini-batch samples during incremental process. Specifically, we ensure the topological relationships between the student's object prototypes and text prototypes remain consistent with those of the teacher. Using the corresponding semantic queries $\{q_1, q_2, \ldots, q_N\}$ and text tokens $\{w_1, w_2, \ldots, w_{(L-1)}\}$, we define the prototype of class c as follows:

$$p_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \alpha_i q_i \tag{8}$$

where p_c is object prototype, N_c represents the number of queries belong to class c in current batch. And α_i is a weight factor defined in Eq. (5) that helps in estimating a more accurate prototype. Similarly, the text prototype is defined as:

$$\hat{p}_c = \frac{1}{N_b} \sum_{i=1}^{N_b} w_i \tag{9}$$

where $\hat{p}_c$ is corresponding text prototype, N_b is the number of batch samples. The relationships are formulated as a distance matrix between each in-image prototype of old classes in current batch, with Euclidean distance used to capture small changes in structure. The relationships for objects and texts are defined as:

$$R_{ij} = \left\| p_i - p_j \right\|_2, \quad i, j \in C_{1:(t-1)} \tag{10}$$

$$\hat{R}_{ij} = \left\| \hat{p}_i - \hat{p}_j \right\|_2, \quad i, j \in C_{1:(t-1)} \tag{11}$$

where i, j refer to two different classes within $C_{1:(t-1)}$. Based on above definition, our CTD can be calculated as:

$$\mathcal{L}_{CTD} = \lambda_1 \left\| R - R^{old} \right\|_2 + \lambda_2 \left\| \hat{R} - \hat{R}^{old} \right\|_2 \tag{12}$$

where λ_1, λ_2 are the coefficients for object topology and text topology loss. And the overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{detr}} + \mathcal{L}_{CRD} + \mathcal{L}_{CTD} \tag{13}$$

Setting	Method	Baseline	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
40+40	CL-DETR (Liu et al. 2023c)	Deformable DETR	42.0	60.1	45.9	24.0	45.3	55.6
	DMD (Kang et al. 2023)	Deformable DETR	39.8	-	-	-	-	-
	SSDGR (Kim et al. 2024)	Deformable DETR	43.0	62.1	47.1	24.9	46.9	57.0
	TALIR (Zhang et al. 2024)	*GLIP	40.4	57.4	43.9	23.3	44.7	54.5
	LWF (Li and Hoiem 2017)	*Grounding Dino	21.8	29.4	23.7	12.4	23.0	30.4
	ERD (Feng, Wang, and Yuan 2022)	*Grounding Dino	29.7	41.8	32.1	19.0	33.1	38.8
	Ours	***Grounding Dino**	**45.7**	**62.9**	**49.7**	**28.4**	**49.3**	**60.0**
70+10	CL-DETR (Liu et al. 2023c)	Deformable DETR	40.4	58.0	43.9	23.8	43.6	53.5
	DMD (Kang et al. 2023)	Deformable DETR	37.6	-	-	-	-	-
	SSDGR (Kim et al. 2024)	Deformable DETR	40.9	59.5	44.8	23.9	44.7	54.0
	TALIR (Zhang et al. 2024)	*GLIP	42.9	59.2	45.2	24.3	45.1	54.1
	LWF (Li and Hoiem 2017)	*Grounding Dino	11.4	16.6	12.1	7.7	13.6	18.0
	ERD (Feng, Wang, and Yuan 2022)	*Grounding Dino	39.0	53.6	42.1	24.7	42.2	53.2
	Ours	***Grounding Dino**	**46.7**	**63.9**	**50.8**	**29.7**	**49.9**	**61.6**

Table 1: IOD results (%) on COCO 2017 in two-phase setting. Method with * indicates vision-language baseline, while others are vision baseline. The results are extracted from corresponding papers. Besides, we reproduce previous methods based on Grounding Dino for comparison. The "-" symbol indicates a missing value. The best performance is highlighted in bold.

Experiments

Implementation details. Our method is based on Grounding Dino-T, all settings are consistent with baseline model. All experiments were conducted on four NVIDIA Tesla A100 GPUs with a total batch size of 32. We used the ADAMW optimizer, setting the learning rate to 5e-5 for all components, except for the vision and language backbones, which were set to 5e-6. The model was trained for 12 epochs (1x mode). After the first phase, we fixed the language backbone and excluded the denoising loss for simplicity.

Datasets and Evaluation Metrics. Our IOD experiments are conducted on COCO 2017 (Lin et al. 2014) which is widely used in recent works. The standard COCO metrics are used for evaluation, i.e., AP, AP_{50}, AP_{75}, AP_S , AP_M, AP_L. And in our ablation study, we leverage forgetting percentage points (FPP) proposed by (Liu et al. 2023c) to evaluate the forgetting of old categories.

Experiment setup. In our experiments, we mainly focus on two-phase($40 + 40$, $70 + 10$) and multi-phase($40 + 10 \times 4$, $40+20\times2$) settings. In the first phase, we train the model via standard fine-tuning. For the subsequent incremental phases, we apply our proposed method. **Two-phase setting:** In the first phase, the model is trained with data D_1 which corresponds to the label space C_1. Then, in the second incremental phase, the model is further trained with data D_2 where only objects belonging to C_2 are labeled. After training, the model is evaluated on the entire label space $C_1 \cup C_2$. Following the data split in(Feng, Wang, and Yuan 2022), some images are shared between D_1 and D_2. **Multi-phase setting:** Similarly, in the first phase, the model is fine-tuned on data D_1 as beginning. In the subsequent incremental phases, at each phase t, new classes C_t are added. The model is evaluated on the cumulative label space $C_{1:t}$ after each phase.

Results

Two-phase setting. In two phase setting, we compare our methods with recent SOTA methods and results are shown in Tab. 1. We categorize these methods into two groups: those based on a vision baseline and those based on a vision-language(V-L) baseline. Compared to vision baseline methods, our approach outperforms current replay-based SOTA method, SSDGR, by 5.8% in the $70+10$ setting and 2.7% in the $40+40$ setting. This proves the great potential of V-L detectors in IOD. For V-L baselines, we compare with existing work TALIR and reproduce previous response-based methods for a comprehensive evaluation. The results indicate that our method achieves the highest AP, with 46.7% and 45.7% in the $70 + 10$ and $40 + 40$ settings. Notably, our baseline model's upper bound is comparable to that of TALIR. The significant lead in both settings demonstrates that our GCD effectively exploits the potential of V-L detectors.

Multi-phase setting. Tab. 3 evaluates our method on multi-phase setting. Though multi-phase settings are generally not a strong suit for KD-based methods, our approach still outperforms SSDGR(Kim et al. 2024) by 3.4% in the $40+10\times4$ setting and 2.9% in the $40 + 20 \times 2$ setting. In particular, the performance gap between our method and TALIR increases as the process progresses, with our method being 10.0% and 6.7% ahead in the $40 + 10 \times 4$ and $40 + 20 \times 2$ settings, respectively. This further demonstrates that GCD better balances stability and plasticity.

Ablation Study

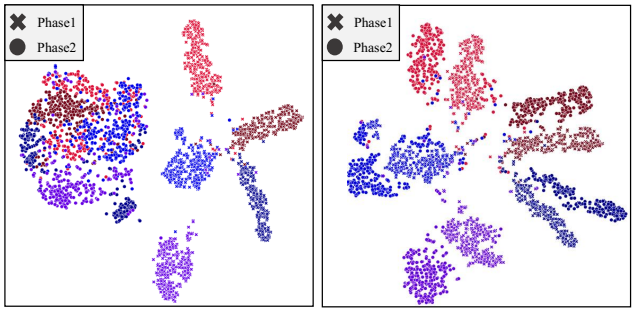
We validate each component of our method under the $70+10$ setting. In Tab. 2, "Local" and "Global" represent fine-tuning with only new text and fine-tuning with the full text, respectively. In Row 2, we introduce CRD based on local alignment, which still results in a weak baseline as new and old knowledge cannot interact within the same embedding space. In Row 4, using pseudo labels to address label conflicts makes global alignment a strong baseline (showing a 44.6% increase in performance for old categories compared to Row 3). In Row 5, naively applying response distillation to our framework decreases the strong baseline's per-

Row	Local	Global	Pseudo	Raw RD	CRD	CTD	All categories ↑		Old categories ↑		FPP ↓	
							AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
1	✓						6.4	8.7	1.8	2.8	48.2	61.8
2	✓				✓		41.0	56.0	41.7	57.1	8.3	11.0
3		✓					6.3	8.4	1.7	2.4	48.3	65.2
4		✓	✓				45.3	62.1	46.3	63.7	3.7	3.9
5		✓	✓	✓			42.9	58.6	44.0	60.4	6.0	7.2
6		✓	✓		✓		46.0	63.0	47.1	64.6	2.9	3.0
7		✓	✓		✓	✓	**46.7**	**63.9**	**48.1**	**65.9**	**1.9**	**1.7**

Table 2: Ablation study using COCO benchmark under $70+10$ setting. We evaluate the performance after completing all phases of training, focusing on three key aspects: the overall performance across all classes(higher is better), the performance of old class categories(higher is better), and the FPP(lower is better) reflects the performance degradation of the old classes compared to the previous phase. The best performance is highlighted in bold, with the final row indicating our method.

Method	Baseline	$40+10\times4$		$40+20\times2$	
		AP	AP_{50}	AP	AP_{50}
DMD	D-DETR	30.3	-	36.6	-
CL-DETR	D-DETR	28.1	-	35.3	-
SSDGR	D-DETR	36.8	54.7	41.1	59.5
TALIR	*GLIP	30.2	-	37.3	-
ERD	*G-DINO	7.2	9.6	18.3	25.2
Ours	*G-DINO	**40.2**	**55.1**	**44.0**	**60.4**

Table 3: IOD results (%) on COCO 2017 in multi-phase setting. The results are extracted from corresponding papers except ERD is reproduced based on Grounding Dino.



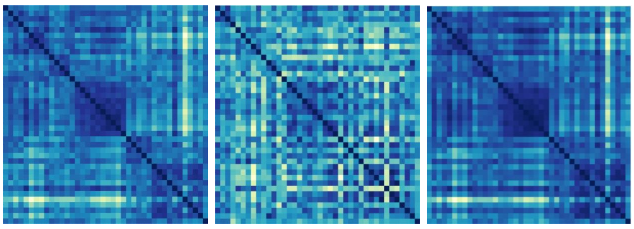
(a) Local Alignment (b) Ours

Figure 4: Visualization of semantic query features of old categories under $40+40$ setting.

formance. In Row 6, CRD effectively compensates for the baseline method's shortcomings and boosts performance by 0.7%. Finally, in Row 7, CTD further stabilizes the local relations of the old knowledge and reduces the FPP to 1.9%.

Further Analysis

In Fig. 4, we project the semantic query features from both phases into the same space to directly reflect the changes during the incremental process. Here, (a) represents fine-tuning with a local alignment approach, where old categories are ignored, leading to feature collapse. In contrast, (b) shows that our GCD framework preserves the semantic structure well, maintaining the relationships of the old categories. Since text features are naturally distinct, to more intuitively illustrate changes in the relationships between text features, we visualize them using a distance matrix, where each cell's value represents the distance between two text prototypes. A larger value indicates that the texts are closer in semantic space. As shown in Fig. 5, (a) depicts the text relationships in phase 1, (b) shows that these relationships change significantly after the increment when fine-tuned in a local alignment manner. As shown in (c), the text relationships remain similar to those in phase 1, demonstrating that our method effectively preserves the text relations.



(a) Phase1 (b) Phase2(Local) (c) Phase2(Ours)

Figure 5: Distance matrix of semantic text features of old categories under $40+40$ setting.

Conclusion

In this paper, we propose a new perspective on catastrophic forgetting in vision-language detectors, termed semantic structure collapse, and demonstrate how GCD effectively addresses this issue. By integrating knowledge across different phases through global alignment, GCD maintains a unified embedding space for both old and new knowledge, facilitating flexible and coherent knowledge integration. Furthermore, we propose semantic correspondence mechanism(SCM) to enable effective KD in VLDs. On this basis, GCD mitigates the overconfidence of noisy pseudo labels and preserves the activation of weak categories using correspondence response distillation (CRD). Additionally, GCD ensures consistency in the local relationships of both modalities with the teacher model through correspondence topology distillation (CTD). Extensive experiments on COCO 2017 validate the efficacy of GCD.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62176246. This work is also supported by Anhui Province Key Research and Development Plan (202304a05020045) and Anhui Province Natural Science Foundation (2208085UD17). This work is also supported by National Natural Science Foundation of China under Grant 62406098. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC. This research was also supported by the advanced computing resources provided by the Supercomputing Center of the USTC.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

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.

Chang, J.; Wang, S.; Xu, H.-M.; Chen, Z.; Yang, C.; and Zhao, F. 2023. Detrdistill: A universal knowledge distillation framework for detr-families. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6898–6908.

Dhar, P.; Singh, R. V.; Peng, K.-C.; Wu, Z.; and Chellappa, R. 2019. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5138–5146.

Douillard, A.; Chen, Y.; Dapogny, A.; and Cord, M. 2021. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4040–4050.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*, 86–102. Springer.

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.

Goodfellow, I. J.; Mirza, M.; Xiao, D.; Courville, A.; and Bengio, Y. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Gupta, A.; Narayan, S.; Joseph, K.; Khan, S.; Khan, F. S.; and Shah, M. 2022. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9235–9244.

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.

Iscen, A.; Zhang, J.; Lazebnik, S.; and Schmid, C. 2020. Memory-efficient incremental learning through feature adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 699–715. Springer.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Joseph, K.; Khan, S.; Khan, F. S.; and Balasubramanian, V. N. 2021. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5830–5840.

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.

Kim, J.; Cho, H.; Kim, J.; Tiruneh, Y. Y.; and Baek, S. 2024. Sddgr: Stable diffusion-based deep generative replay for class incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28772–28781.

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, 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. *Advances in Neural Information Processing Systems*, 33: 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.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Lin, Z.; Wang, Z.; and Zhang, Y. 2022. Continual semantic segmentation via structure preserving and projected feature alignment. In *European Conference on Computer Vision*, 345–361. Springer.

Lin, Z.; Wang, Z.; and Zhang, Y. 2023. Preparing the Future for Continual Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11910–11920.

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

Liu, Y.; Cong, Y.; Goswami, D.; Liu, X.; and van de Weijer, J. 2023b. Augmented box replay: Overcoming foreground shift for incremental object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11367–11377.

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

Liu, Y.; Su, Y.; Liu, A.-A.; Schiele, B.; and Sun, Q. 2020b. Mnemonics training: Multi-class incremental learning without forgetting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 12245–12254.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

Menezes, A. G.; de Moura, G.; Alves, C.; and de Carvalho, A. C. 2023. Continual object detection: a review of definitions, strategies, and challenges. *Neural networks*, 161: 476–493.

Peng, C.; Zhao, K.; and Lovell, B. C. 2020. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern recognition letters*, 140: 109–115.

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.

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. PMLR.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE 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.

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.

Thrun, S. 1995. Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*, 8.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2020. FCOS: A simple and strong anchor-free object detector. *IEEE transactions on pattern analysis and machine intelligence*, 44(4): 1922–1933.

Wang, Y.; Li, X.; Weng, S.; Zhang, G.; Yue, H.; Feng, H.; Han, J.; and Ding, E. 2024. KD-DETR: Knowledge Distillation for Detection Transformer with Consistent Distillation Points Sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16016–16025.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 374–382.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhang, H.; Gao, B.-B.; Zeng, Y.; Tian, X.; Tan, X.; Zhang, Z.; Qu, Y.; Liu, J.; and Xie, Y. 2024. Learning Task-Aware Language-Image Representation for Class-Incremental Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7096–7104.

Zheng, Z.; Ma, M.; Wang, K.; Qin, Z.; Yue, X.; and You, Y. 2023. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19125–19136.

Zhou, D.-W.; Zhang, Y.; Ning, J.; Ye, H.-J.; Zhan, D.-C.; and Liu, Z. 2023. Learning without forgetting for vision-language models. *arXiv preprint arXiv:2305.19270*.

Zhu, F.; Zhang, X.-Y.; Wang, C.; Yin, F.; and Liu, C.-L. 2021. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5871–5880.

Zhu, H.; Wei, Y.; Liang, X.; Zhang, C.; and Zhao, Y. 2023. Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22257–22267.