# Progressive Domain-Invariant Injection for Cross-Domain Few-Shot Object Detection

Huiwen Huang[1], Shenghao Fu[2], Kun-Yu Lin[2],
Wei-Jin Huang[2], Shuxuan Li[2], Nan Lei[2], Xiaonan Luo[1,*], Wei-Shi Zheng[2,*]
[1]School of Computer Science and Information Security, Guilin University of Electronic Technology, China
[2]School of Computer Science and Engineering, Sun Yat-sen University, China
[*]Corresponding authors: luoxn@guet.edu.cn and wszheng@ieee.org

***Abstract*—Cross-domain few-shot object detection (CD-FSOD) aims to learn a detector that generalizes from a label-rich source domain to a novel target domain with disjoint categories and scarce annotations. A critical challenge is mitigating the domain gap between the source and target domains. To address this, we propose DP-CDFSOD, which progressively learns domain-invariant detection representations by leveraging cross-modal semantics as domain-invariant anchors. Our framework comprises three lightweight modules that inject cross-modal priors across the Encoder–Decoder–Head architecture. The Query-Side Semantic Calibration (QSC) and Support-Guided Query Refinement (SQR) modules inject text-aligned visual priors into the encoder and decoder, respectively. The Cross-Modal Alignment and Discrimination (CAD) module anchors the classification decision and realigns the visual priors with textual prototypes in the Head. This progressive cross-modal integration enables our framework to achieve state-of-the-art performance on the CD-FSOD benchmark, improving 3.7%mAP in the challenging 1-shot setting. Moreover, the framework supports efficient target-domain adaptation by fine-tuning a few lightweight projection layers with only 2.37M trainable parameters. It allows each domain to share the majority of the source-domain parameters while maintaining a small number of domain-specific ones. Code is available at https://github.com/HHuiwen/DP-CDFSOD.**

***Index Terms*—Object detection, Few-shot learning, Cross-domain adaptation, Domain-invariant priors, Multi-modal fusion**

## I. Introduction

This paper focuses on CD-FSOD [1], [2], where a general detector is pre-trained on a richly labeled source domain and then independently fine-tuned for multiple target domains. Each target domain offers only N-way K-shot labeled examples, with categories disjoint from the source domain and a highly divergent data distribution. Current meta-learning based FSOD [3], [4] methods commonly adopt a Support-Query mechanism for training and inference: the model extracts visual features from a small number of labeled examples in the support set and uses them to guide the detection of target objects in query images. However, under extremely scarce sample conditions, the visual class prototypes estimated from limited support samples tend to exhibit high variance and instability. To enhance prototype stability and generalization, some studies [5], [6] introduce text embeddings from vision-language models [7], [8] as "semantic anchors". Due to their inherent domain-invariant properties, text priors provide more consistent semantic representations of categories, thereby improving inter-class discriminability.

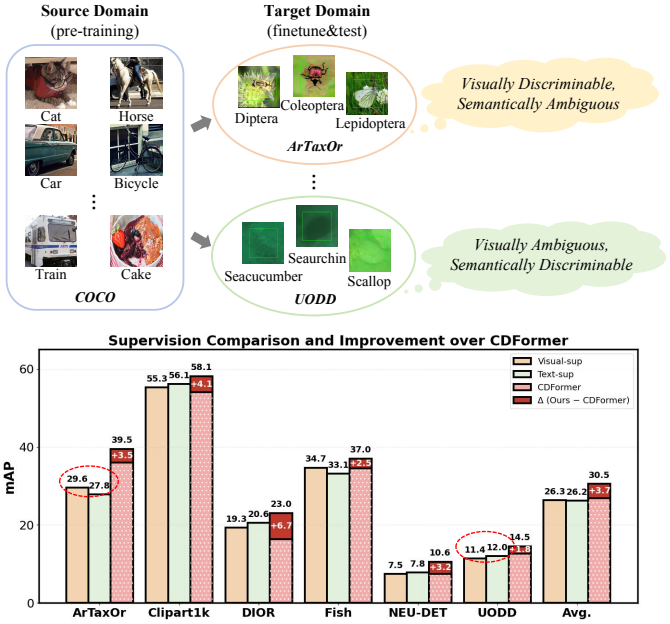


Fig. 1. In CD-FSOD, a detector is trained on a source domain and adapted to multiple target domains with limited annotations. Due to severe data scarcity and domain shifts, visual-only or semantic-only supervision fails to capture discriminative class prototypes. By integrating complementary visual cues with domain-invariant semantic priors, our method enables more robust cross-domain adaptation and consistently outperforms the state-of-the-art CDFormer across diverse target domains.

However, experiments show that in cross-domain scenarios, text supervision does not consistently outperform visual-only methods (see Fig. 1). The core contradiction lies in two aspects: although text features are largely domain-invariant, their generalization from natural image–text data often lacks sufficient discriminative power for fine-grained or abstract categories (e.g., insect subspecies in ArTaxOr [9]); in contrast, visual features, while rich in detail, are highly sensitive to domain shifts and image quality, leading to significant performance degradation under few-shot settings or challenging domains such as underwater imagery in UODD [10]. These findings reveal the inherent limitations of single-modal priors

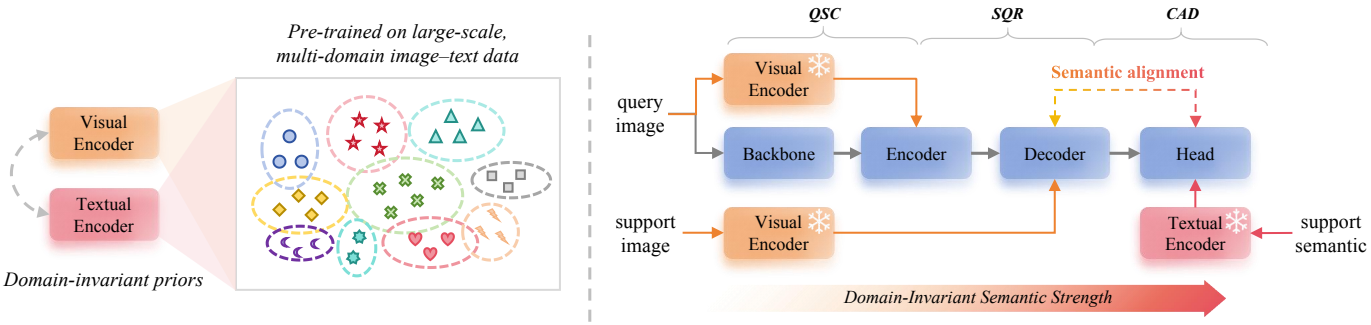


Fig. 2. Schematic architecture of the proposed framework. Vision–language models (e.g., CLIP), which are pretrained on massive image–text pairs from diverse domains, provide domain-invariant priors. Therefore, we exploit these priors to guide representation learning by progressively injecting and strengthening domain-invariant features through the Encoder-Decoder-Head architecture, thereby alleviating cross-domain shift and improving inter-class discriminability.

in cross-domain few-shot detection tasks and inspire us to explore a new multi-modal semantic injection mechanism: how to effectively leverage domain-invariant textual semantics as a stable anchor and domain-sensitive visual information as fine-grained supplements, so as to progressively mitigate domain bias during adaptation and enhance the discriminability of the classification representations.

We hypothesize that **although visual features are sensitive to domain shifts, text-aligned visual priors (e.g., CLIP [7], [11] visual features) retain partial domain invariance, owing to their training on the large-scale image–text dataset**. Therefore, integrating such priors can enhance cross-domain adaptability. Guided by this insight, we propose DP-CDFSOD, which is designed to refine query image features from domain-sensitive to domain-invariant representations progressively (see Fig. 2). The framework systematically introduces and reinforces domain-invariant cross-modal priors throughout the Encoder → Decoder → Head architecture. This process guides the query image representations to become domain-invariant and aligned with the stable textual semantic space without compromising fine-grained appearance discrimination. Specifically, we design three lightweight modules: (1) Query-Side Semantic Calibration (QSC) injects text-aligned visual priors from the query image into the Encoder, thereby suppressing domain-specific perturbations. (2) Support-Guided Query Refinement (SQR) injects text-aligned object-aware visual priors from the support images into the Decoder, thereby iteratively refining the object queries to enhance discriminability. (3) Cross-Modal Alignment and Discrimination (CAD) anchors the classification decision space using domain-invariant textual class prototypes in the Head and enforces a cross-modal consistency loss to enhance inter-class separation.

By systematically integrating domain-invariant priors, we effectively refine query image features, transforming them from domain-sensitive to domain-invariant representations. On the challenging CD-FSOD benchmark [1], our DP-CDFSOD consistently outperforms the previous state-of-the-art model CDFormer [12] across six diverse target domains, achieving a notable average improvement of 3.7% mAP in the 1-shot setting. Furthermore, thanks to the distinctive domain-invariant representations learned by our model, we can adapt to new domains by fine-tuning only 2.37M domain projection parameters, enabling highly efficient transfer with minimal parameter updates.

## II. Related Work

### A. DETR-based Object Detector

Object detection [13]–[15] aims to predict both the locations and categories of objects. Recently, Transformer-based object detectors DETR [16] and its variants [17]–[19] have attracted increasing attention. DETR-based model formulates object detection as a set prediction problem and adopts a standard Encoder–Decoder–Head architecture [20]. The Encoder models global contextual information from multi-scale visual features, while the Decoder employs a set of learnable object queries to decode object-level representations. The Head then predicts object categories and bounding boxes based on the decoded features. This end-to-end formulation eliminates the need for handcrafted anchors and post-processing steps such as non-maximum suppression (NMS). In this work, we adopt Deformable-DETR [21] as the base detection framework. By introducing multi-scale feature representations and a sparse deformable attention mechanism, Deformable-DETR significantly accelerates training convergence and improves computational efficiency. In our approach, domain-invariant priors are integrated into the Encoder, Decoder, and Head, enhancing the model's adaptability from the source domain representation to diverse target domains.

### B. Few-Shot Object Detection

Traditional meta-learning-based few-shot object detection methods [3], [4], [22] construct visual class prototypes from the support set to assist detection. With the development of vision-language models (VLMs) [7], researchers [23] have introduced text prompts generated from class names to provide strong semantic priors. Some works [24] further integrate visual and text prompts to enhance class prototypes. However, these methods generally overlook the cross-domain setting: how to collaboratively leverage domain-specific visual prompts and domain-invariant text prompts to effectively mitigate the domain gap remains an understudied problem.

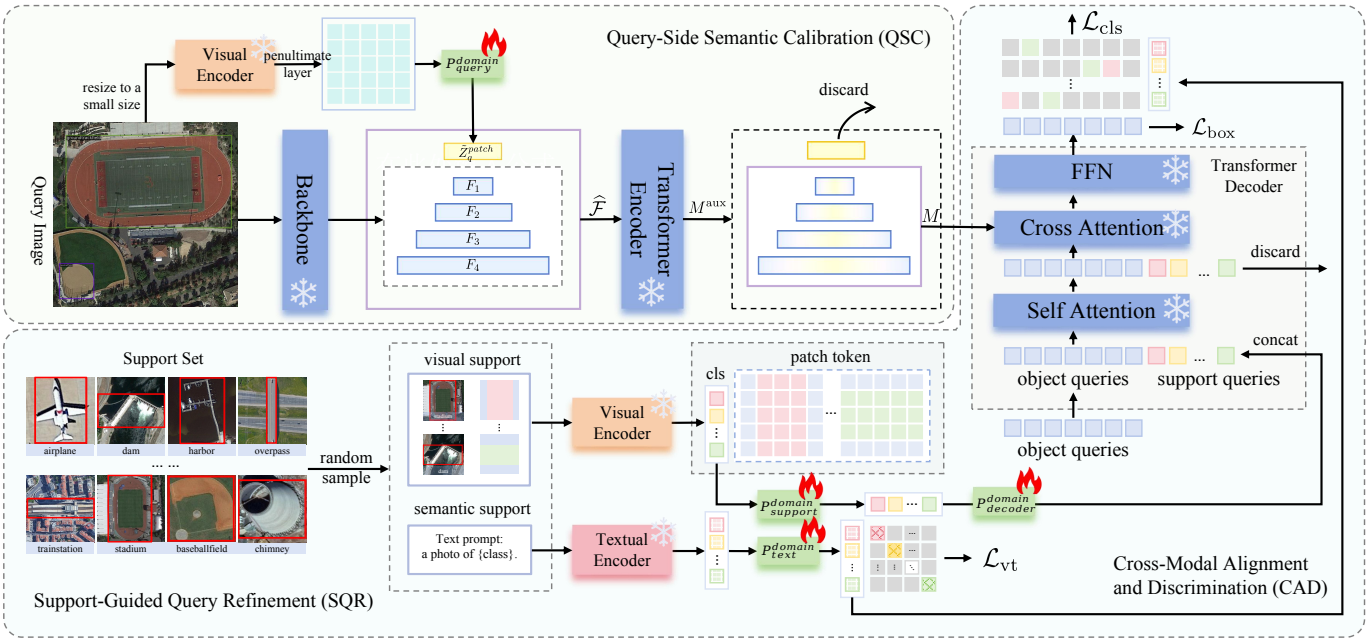


Fig. 3. **Overview of DP-CDFSOD.** The model progressively injects *domain-invariant cross-modal priors* across the Encoder–Decoder–Head architecture with three modules: **QSC (Sec. III-C)** fuses text-aligned features from the query image with backbone features in the encoder. For **SQR (Sec. III-D)**, text-aligned object visual priors from support set interact with object queries within the decoder. **CAD (Sec. III-E)** uses text prototypes as the classification head which are also aligned with support-derived visual prototypes via a symmetric InfoNCE loss. This staged integration yields domain-invariant query features and enables lightweight target adaptation by fine-tuning only a few projection layers (highlighted by a flame icon).

### C. Cross-Domain Few-Shot Object Detection

Most prior research has focused on constructing visual prompts. CD-ViTO [1] optimizes class prototypes through contrastive learning between learnable domain prompts and visual prompts. CDFormer [12] introduces learnable background prompts to enhance foreground-background separation. VFMDETR [25] leverages a mixture of DETR and a vision foundation model (e.g., DINOv2 [26]) to handle domain shift through feature-level expert fusion rather than prototype prompting. These methods predominantly focus on visual representations, while neglecting complementary domain-invariant semantic knowledge for mitigating domain shift. A few studies [27] fuse visual features with lengthy category descriptions to augment prototypes, but their static concatenation induces semantic redundancy and granularity mismatch. Moreover, such enhancement is typically confined to the encoder, without encouraging domain-invariant representation learning throughout the full detection pipeline.

## III. Methodology

### A. Problem Setup

Let the label-sufficient source domain $\mathcal{S} = \{(x_i^s, Y_i^s)\}_{i=1}^{n_s}$ contain $n_s$ labeled images, where $Y_i^s$ includes $m_i^s$ object annotations of the form $(c_{ij}^s, b_{ij}^s)$ for category $c_{ij}^s \in \mathcal{C}_s$ and bounding box $b_{ij}^s \in \mathbb{R}^4$. We consider a target domain $\mathcal{T}$ with a class set $\mathcal{C}_t$ disjoint from $\mathcal{C}_s$ ($\mathcal{C}_t \cap \mathcal{C}_s = \varnothing$) and a data distribution $p_t(x)$ that differs from the source distribution $p_s(x)$. For the target domain $\mathcal{T}$, only $C_t$-way $K$-shot labels are available, formalizing as the training set $\mathcal{D}_{\text{train}} = \{(x_n, Y_n)\}_{n=1}^{C_t \times K}$. We first pre-train a general detector on $\mathcal{S}$ and then perform few-shot adaptation using $\mathcal{D}_{\text{train}}$, followed by evaluation on the corresponding test set $\mathcal{D}_{\text{test}}$.

### B. Overall Architecture

Fig. 3 illustrates our framework. Our DP-CDFSOD progressively injects domain-invariant cross-modal priors throughout the Encoder-Decoder-Head pipeline to align textual anchors: **First**, we design *Query-Side Semantic Calibration (QSC)* module to inject text-aligned CLIP visual priors from the query image into the Encoder, suppressing domain-specific perturbations. **Furthermore**, to enhance the discriminability of object queries, we propose *Support-Guided Query Refinement (SQR)* by injecting text-aligned object-aware CLIP visual priors from the support images into the Decoder. **Finally**, the *Cross-Modal Alignment and Discrimination (CAD)* module is adopted in the Head to anchor the classification decision and realign the object-aware visual priors with domain-invariant textual prototypes.).

In contrast to previous approaches [27] that directly fuse CLIP's multi-modal features and only enhance prototypes within the encoder, this study progressively injects domain-invariant cross-modal knowledge at different stages of the detection network, realigning the visual and textual semantic spaces. Consequently, query features become more domain-invariant and decision boundaries sharper.

### C. Query-Side Semantic Calibration (QSC)

Visual features extracted from the encoder are highly domain-sensitive. To enhance their domain invariance, we

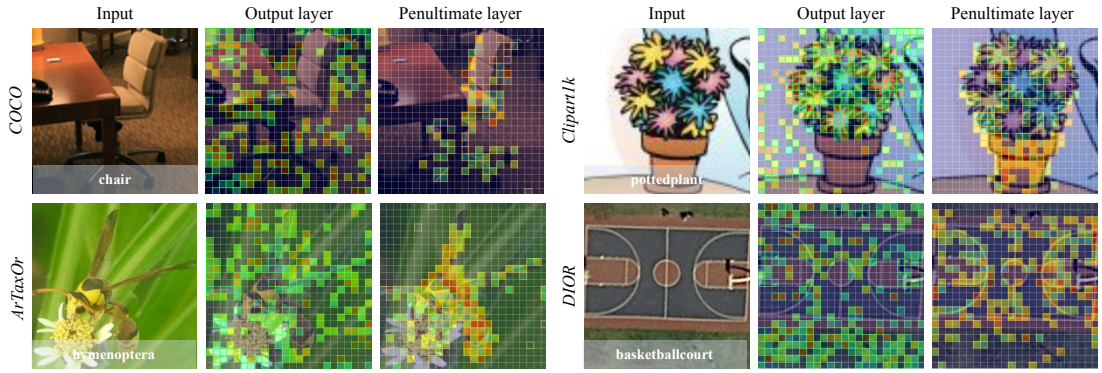


Fig. 4. Visualization of the similarity scores between patch tokens from the CLIP visual encoder and the text embedding from the textual encoder. The results show that patch tokens from the penultimate layer align more closely with textual semantics than those from the output layer. Fusing patch tokens from the penultimate layer with backbone features helps steer the source visual features towards learning domain-invariant representations.

inject text-aligned visual features extracted by CLIP into the encoder, effectively mitigating domain-specific perturbations.

Specifically, we use patch tokens from the penultimate layer of CLIP as visual priors, as they exhibit better alignment with text than those from the output layer, as shown in Fig. 4. Let $g_v$ denote the CLIP visual encoder, and $x$ be a query image. We extract the penultimate layer patch tokens and project them to obtain text-aligned visual features $Z_q^{\text{patch}} = P_{v\to t}^{\text{clip}}\big(g_v(x)_{\text{patch}}\big) \in \mathbb{R}^{N_p\times d_t^{\text{clip}}}$, where $g_v(x)_{\text{patch}} \in \mathbb{R}^{N_p\times d_v^{\text{clip}}}$ are the penultimate layer patch tokens, $N_p = H_pW_p$ denotes the number of patches, $d_v^{\text{clip}}$ and $d_t^{\text{clip}}$ are the CLIP visual and text embedding dimensions, respectively, and $P_{v\to t}^{\text{clip}} : \mathbb{R}^{d_v^{\text{clip}}} \to \mathbb{R}^{d_t^{\text{clip}}}$ is to map visual tokens into the text-aligned space.

Then, we map the text-aligned visual features from the CLIP space to the detector dimension $d$:

$$\tilde{Z}_q^{\text{patch}} = P_{\text{query}}^{\text{domain}}\big(Z_q^{\text{patch}}\big) \in \mathbb{R}^{N_p\times d}, \tag{1}$$

where $P_{\text{query}}^{\text{domain}} : \mathbb{R}^{d_t^{\text{clip}}} \to \mathbb{R}^d$ is a learnable projection shared across tokens. Let multi-scale features of the backbone be $\mathcal{F} = \{F_\ell\}_{\ell=1}^4$ with $F_\ell \in \mathbb{R}^{N_\ell\times d}$. We treat $\tilde{Z}_q^{\text{patch}}$ as an *additional scale* and concatenate it along the token axis:

$$\begin{aligned}\widehat{\mathcal{F}} &= \big\{F_1, F_2, F_3, F_4, \tilde{Z}_q^{\text{patch}}\big\},\\ M^{\text{aux}} &= \mathcal{E}(\widehat{\mathcal{F}}) \in \mathbb{R}^{\left(\sum_{\ell=1}^4 N_\ell + N_p\right)\times d}.\end{aligned} \tag{2}$$

where $\mathcal{E}$ is the encoder. After encoding, we *discard the additional feature map*, obtaining the feature matrix $M \in \mathbb{R}^{(\sum_{\ell=1}^4 N_\ell)\times d}$ that serves as input to the decoder.

### D. Support-Guided Query Refinement (SQR)

The decoder refines object queries through stacked layers, each combining self-attention among object queries and cross-attention to the query-image features. However, these object queries inherit domain bias. To improve their domain invariance and class separability, we inject text-aligned object-aware visual features from the support set into the decoder, steering object queries toward semantically consistent and more discriminative representations. This module consists of two main steps: prototype extraction and injection.

Specifically, the first step is extract class prototypes from the CLIP visual encoder. However, using CLIP's global `[CLS]` token directly as a visual prototype is problematic [28]: it is biased by center cropping and sensitive to background clutter, leading to distorted representations. To address this, for a support image $x^{\text{sup}}$ with instance box $B$, we extract a square crop centered on $B$ and compute object-aware CLIP tokens $z_{\text{obj}} = [z_{\text{cls}}, Z_{\text{patch}}]$, along with a binary mask $\mathbb{M}_B$. This mask selects patch tokens whose receptive field centers lie within $B$ on the CLIP patch token grid, as shown in Fig. 5. We find that using an object-aware mask across all 24 layers of the visual encoder effectively improves the semantic confidence of the object and reduces background interference.

$$\begin{aligned}z_{\text{obj}} &= \text{MSA}\left(z_{\text{cls}};\, Z_{\text{patch}}[\mathbb{M}_B]\right),\\ v_c &= P_{\text{support}}^{\text{domain}}(z_{\text{obj}}) \in \mathbb{R}^d,\end{aligned} \tag{3}$$

where $MSA(\cdot)$ denotes *masked self-attention* and $P_{\text{support}}^{\text{domain}} : \mathbb{R}^{d_v^{clip}} \to \mathbb{R}^d$ projects the CLIP space to the detector dimension $d$. We then compute class-wise visual priors by $K$-shot averaging, $V_c = \frac{1}{K}\sum_{k=1}^K v_c^{(k)}$, and aggregate the prototypes of all classes in the episode into a matrix: $V = \text{stack}\big(V_c\big)_{c\in\mathcal{C}} \in \mathbb{R}^{|\mathcal{C}|\times d}$, where $\mathcal{C}$ is the episode's class set.

Building on the extracted prototypes, the second step is to inject them into each decoder layer to refine the object queries. Let $O^{(\ell-1)} \in \mathbb{R}^{N_q\times d}$ be the object queries before decoder layer $\ell$. We append the class-wise priors $V$ during self-attention and keep only the object queries for the subsequent cross-attention:

$$\hat{O}^{(\ell-1)} = \Pi_{\text{qry}}\Big[\text{SA}\Big(\big[O^{(\ell-1)};\, P_{\text{decoder}}^{\text{domain}}(V)^{(\ell-1)}\big]\Big)\Big], \tag{4}$$

$$O^{(\ell)} = \text{CA}\big(\hat{O}^{(\ell-1)}, M\big). \tag{5}$$

*where* $\Pi_{\text{qry}} : \mathbb{R}^{(N_q+|\mathcal{C}|)\times d} \to \mathbb{R}^{N_q\times d}$ selects the first $N_q$ object queries corresponding to the original positions and discards the appended $|\mathcal{C}|$ support object queries; $\text{SA}(\cdot)$ and $\text{CA}(\cdot,\cdot)$ denote multi-head *self-attention* and *cross-attention*, respectively. $P_{\text{decoder}}^{\text{domain}\,(l-1)} : \mathbb{R}^d \to \mathbb{R}^d$ *is* specific to decoder layer $\ell$.

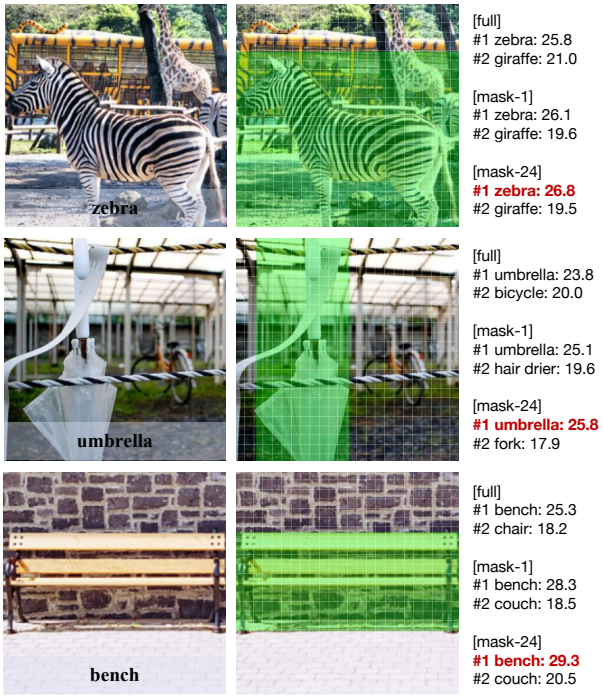


Fig. 5. Visualization of patch-token masking in CLIP's visual encoder, which confines the class token's attention to the object's bounding box. [full] means no modification, aggregating all patch tokens. [mask-k] means mask the last k-th layer(s) of the visual encoder. #1 and #2 represent the top-1 and top-2 confidence categories, respectively. Constraining the class token's attention to the object's bounding box with in all layers ([mask-24]) reduces background interference and enhances class discrimination.

This design steers object queries along class-consistent directions and suppresses background leakage. Since $|\mathcal{C}|$ is episode-way, the extra cost in self-attention is minor, and the decoder's overall footprint remains essentially the same.

### E. Cross-Modal Alignment and Discrimination (CAD)

The classification head provides decision boundaries under domain shift and enables few-shot adaptation. As analyzed in Sec. I, textual semantics are domain-invariant but weak on fine-grained or abstract categories; visual cues are discriminative but domain-sensitive. To harness both, we propose a dual-alignment strategy to enforce cross-modal consistency and sharpen inter-class margins.

*(i) Text–query alignment (classification).* To anchor the classification decision with domain-invariant information, we align object queries with textual priors. Let $g_t$ be the CLIP text encoder and $P_{\text{text}}^{\text{domain}} : \mathbb{R}^{d_t} \to \mathbb{R}^d$ the projection to detector width $d$. Given a class phrase, we compute $t_c = g_t(\text{phrase}_c)$ and its text prototype $T_c = P_{\text{text}}^{\text{domain}}(t_c) \in \mathbb{R}^d$. Let $h$ map a decoded query token $o$ to $h(o) \in \mathbb{R}^d$. Class logits are scaled cosine similarities:

$$z_c(o) = \alpha \, cos\big(h(o), T_c\big), \tag{6}$$

where $cos(\cdot,\cdot)$ is cosine similarity on $\ell_2$-normalized vectors and $\alpha > 0$ is a learnable scale.

*(ii) Text–prototype alignment (cross-modal).* To enhance the domain-invariant information of the visual class prototypes and leverage the complementary knowledge of visual and textual information, we realign the visual prototypes with the textual priors. Let $V_c \in \mathbb{R}^d$ be the object-aware visual prior from Sec. III-D. We couple $V_c$ and $T_c$ with a symmetric InfoNCE:

$$\mathcal{L}_{\text{vt}} = -\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \Bigg[ \log \frac{\exp\big(cos(V_c, T_c)/\tau\big)}{\sum_{c'} \exp\big(cos(V_c, T_{c'})/\tau\big)} + \log \frac{\exp\big(cos(T_c, V_c)/\tau\big)}{\sum_{c'} \exp\big(cos(T_c, V_{c'})/\tau\big)} \Bigg], \tag{7}$$

with temperature $\tau > 0$, which is set to 10 by default in all our experiments. This realignment ensures that the visual prototypes retain domain-agnostic properties while capturing rich class-specific semantics from both modalities, thereby improving the overall representation.

**Training objective.** We use Hungarian matching to assign positive predictions. The total loss is defined as $\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \lambda_{\text{vt}}\mathcal{L}_{\text{vt}}$, where $\mathcal{L}_{\text{cls}}$ denotes the focal classification loss and $\mathcal{L}_{\text{box}}$ represents the bounding box regression loss combining $\ell_1$ and GIoU losses, following DETR [21].

## IV. Experiment

### A. Experimental Setup

**Datasets and evaluation.** We adopt the standard CD-FSOD benchmark [1] in previous work [12]: the detector is pre-trained on the source dataset MSCOCO [33] and then independently adapted to six diverse target datasets with distinct domain shifts. MSCOCO is a large-scale dataset comprising 80 object categories captured in common and everyday environments. In contrast, the target domains differ significantly in terms of image style and object categories, creating a substantial domain gap. Specifically, ArTaxOr [9] is a photorealistic dataset with 6 classes of arthropods, while Clipart1k [34] consists of cartoon-style images representing 20 general object categories. DIOR [35] is an aerial imagery dataset containing 20 classes of annotated satellite objects, whereas DeepFish [36] offers underwater images of fish from various natural habitats. NEU-DET [37] focuses on industrial objects, specifically 6 classes of surface defects on hot-rolled steel strips, and UODD [10] is an underwater dataset containing three categories of aquatic objects. To assess cross-domain few-shot generalization, we conduct 1-, 5-, and 10-shot adaptations for each target domain and evaluate the performance using Mean Average Precision (mAP) on the corresponding target test sets. To ensure a fair comparison, all adaptations use the same source pre-trained weights and evaluation protocol.

**Implementation details.** Following prior works [12], [27], we use the self-supervised foundation model DINOv2-L [26] as the backbone. For classification, we replace the conventional C-way linear classifier with a 256-dimensional embedding head and compute cosine similarity against text prompts [38]. We use CLIP ViT-L-14-336px [7] for both the visual and

TABLE I

MAIN RESULTS (MAP) ON THE CD-FSOD BENCHMARK. BEST RESULTS ARE HIGHLIGHTED IN PURPLE. "MOD." INDICATES WHETHER THE MODEL USES SINGLE-MODAL (S) OR MULTI-MODAL (M) INPUTS. "AVG." DENOTES THE AVERAGE SCORES ACROSS ALL TARGET DOMAINS.

| Method | Backbone | Mod. | ArT | Cli | DIO | UOD | Fish | NEU | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| **1-shot** | | | | | | | | | |
| Distill-cdfsod [29] | ResNet50 | S | 5.1 | 7.6 | 10.5 | 5.9 | – | – | / |
| ViTDeT-FT [30] | ViT-B/14 | S | 5.9 | 6.1 | 12.9 | 4.0 | 0.9 | 2.4 | 5.4 |
| DE-ViT [31] | ViT-L/14 | S | 0.4 | 0.5 | 2.7 | 1.5 | 0.4 | 0.4 | 1.0 |
| DE-ViT-FT [31] | ViT-L/14 | S | 10.5 | 13.0 | 14.7 | 2.4 | 19.3 | 0.6 | 10.1 |
| CD-ViTO [1] | ViT-L/14 | S | 21.0 | 17.7 | 17.8 | 3.1 | 20.3 | 3.6 | 13.9 |
| CDFormer [12] | ViT-L/14 | S | 29.4 | 49.6 | 8.0 | 12.2 | 27.7 | 3.0 | 21.7 |
| CDFormer-FT [12] | ViT-L/14 | S | 36.0 | 54.0 | 16.3 | 12.7 | 34.5 | 7.4 | 26.8 |
| VFMDETR [25] | ResNet50 | S | 26.1 | 20.1 | 20.6 | 9.0 | 24.2 | 9.1 | 18.2 |
| Detic [32] | ViT-L/14 | M | 0.6 | 11.4 | 0.1 | 0.0 | 0.9 | 0.0 | 2.2 |
| Detic-FT [32] | ViT-L/14 | M | 3.2 | 15.1 | 4.1 | 4.2 | 9.0 | 3.8 | 6.6 |
| CDMM-FSOD [27] | ResNet101 | M | 15.1 | – | 14.3 | 6.9 | – | – | / |
| **Ours** | ViT-L/14 | M | **39.5** | **58.1** | **23.0** | **14.5** | **37.0** | **10.6** | **30.5** |
| **5-shot** | | | | | | | | | |
| Distill-cdfsod [29] | ResNet50 | S | 12.5 | 23.3 | 19.1 | 12.2 | 15.5 | 16.0 | 16.4 |
| ViTDeT-FT [30] | ViT-B/14 | S | 20.9 | 23.3 | 23.3 | 11.1 | 9.0 | 13.5 | 16.9 |
| DE-ViTcite [31] | ViT-L/14 | S | 10.1 | 5.5 | 7.8 | 3.1 | 2.5 | 1.5 | 5.1 |
| DE-ViT-FT [31] | ViT-L/14 | S | 38.0 | 38.1 | 23.4 | 5.0 | 21.2 | 7.8 | 22.3 |
| CD-ViTO [1] | ViT-L/14 | S | 47.9 | 41.1 | 26.9 | 6.8 | 22.3 | 11.4 | 26.1 |
| CDFormer [12] | ViT-L/14 | S | 38.4 | 53.6 | 7.9 | 16.5 | 24.6 | 3.6 | 24.1 |
| CDFormer-FT [12] | ViT-L/14 | S | **65.0** | **58.9** | 28.1 | 23.8 | 31.7 | 15.0 | 37.1 |
| VFMDETR [25] | ResNet50 | S | 63.3 | 45.1 | **32.1** | 19.6 | 29.5 | 19.0 | 34.8 |
| Detic [32] | ViT-L/14 | M | 0.6 | 11.4 | 0.1 | 0.0 | 0.9 | 0.0 | 2.2 |
| Detic-FT [32] | ViT-L/14 | M | 8.7 | 20.2 | 12.1 | 10.4 | 14.3 | 14.1 | 13.3 |
| CDMM-FSOD [27] | ResNet101 | M | 48.7 | – | 26.9 | 12.5 | – | – | / |
| **Ours** | ViT-L/14 | M | 57.0 | 58.6 | 30.2 | **27.8** | **36.2** | **19.8** | **38.3** |
| **10-shot** | | | | | | | | | |
| Distill-cdfsod [29] | ResNet50 | S | 18.1 | 27.3 | 26.5 | 14.5 | 15.5 | 21.1 | 20.5 |
| ViTDeT-FT [30] | ViT-B/14 | S | 23.4 | 25.6 | 29.4 | 15.6 | 6.5 | 15.8 | 19.4 |
| DE-ViT [31] | ViT-L/14 | S | 9.2 | 11.0 | 8.4 | 3.1 | 2.1 | 1.8 | 5.9 |
| DE-ViT-FT [31] | ViT-L/14 | S | 49.2 | 40.8 | 25.6 | 5.4 | 21.3 | 8.8 | 25.2 |
| CD-ViTO [1] | ViT-L/14 | S | 60.5 | 44.3 | 30.8 | 7.0 | 22.3 | 12.8 | 29.6 |
| CDFormer [12] | ViT-L/14 | S | 37.3 | 53.5 | 7.9 | 16.7 | 25.7 | 4.0 | 24.2 |
| CDFormer-FT [12] | ViT-L/14 | S | 68.7 | 59.0 | 32.5 | 26.4 | 35.5 | 18.1 | 40.0 |
| VFMDETR [25] | ResNet50 | S | **71.3** | 49.9 | **37.8** | 22.1 | 34.1 | **23.7** | 39.8 |
| Detic [32] | ViT-L/14 | M | 0.6 | 11.4 | 0.1 | 0.0 | 0.9 | 0.0 | 2.2 |
| Detic-FT [32] | ViT-L/14 | M | 12.0 | 22.3 | 15.4 | 14.4 | 17.9 | 16.8 | 16.5 |
| CDMM-FSOD [27] | ResNet101 | M | 61.4 | – | 31.4 | 17.5 | – | – | / |
| **Ours** | ViT-L/14 | M | 63.3 | **59.7** | 33.3 | **30.2** | **38.3** | 22.4 | **41.2** |

text encoders. Unless otherwise noted, we adopt the prompt template "`a photo of {class name}`". For source domain pre-training, we use a total batch size of 32 and set the maximum query image size to 512. The model is trained for 20 epochs on eight NVIDIA RTX 3090 GPUs, using the AdamW optimizer with an initial learning rate of 1e-4, which is reduced to 1e-5 at the 16th epoch. For target domain fine-tuning, we use a total batch size of 16 and set the maximum query image size to 800. The AdamW optimizer is employed with a learning rate of 5e-5. Fine-tuning experiments are conducted on two NVIDIA RTX 3090 GPUs for 300 epochs, with early stopping applied.

### B. Main Results on CD-FSOD

Table I reports consistent gains over state-of-the-art baselines across six target datasets under the 1-, 5-, and 10-shot settings, achieving 30.5% mAP, 38.3% mAP, and 41.2% mAP. The high performance demonstrates our strong transferable capacity. In particular, our method outperforms CDFormer [12] by +3.7% mAP in the challenging 1-shot setting, showing that domain-invariant textual semantics complemented with fine-grained visual prototypes can serve as a robust classifier across multiple domains. In contrast, CDFormer relies primarily on in-domain adaptation to construct visual class prototypes. As a result, the representations learned on the source domain tend to be domain-specific and generalize poorly to unseen target domains, leading to inferior performance, especially in low-shot settings.

TABLE II

RESULTS OF ABLATION STUDIES. THE **TEXT** COLUMN INDICATES WHETHER TEXTUAL CLASS PROTOTYPES ARE USED FOR CLASSIFICATION. THE RED FONT DENOTES IMPROVEMENTS, WHEREAS THE BLUE FONT INDICATES DECREASES.

| Text | QSC | CAD | SQR | ArT | Cli | DIO | UOD | Fish | NEU | Avg. | gain |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 29.6 | 55.3 | 19.3 | 11.4 | 34.7 | 7.5 | 26.3 | - |
| ✓ | | | | 27.8 | 56.1 | 20.6 | 12.0 | 33.1 | 7.8 | 26.2 | -0.1 |
| ✓ | ✓ | | | 34.5 | 57.4 | 20.8 | 11.3 | 34.2 | 8.2 | 27.7 | +1.4 |
| ✓ | | ✓ | | 32.2 | 55.4 | 19.9 | 11.5 | 33.7 | 8.7 | 26.9 | +0.6 |
| ✓ | ✓ | ✓ | ✓ | 37.6 | 57.9 | 22.0 | 10.3 | 34.8 | 9.1 | 28.6 | +2.3 |

TABLE III

COMPARISON OF TRAINABLE PARAMETERS (TRAINP↓) AND PERFORMANCE (MAP↑) IN THE 1-SHOT SETTING.

| Trainable Component | ArT | Cli | DIO | UOD | Fish | NEU | Avg. | TrainP. |
|---|---|---|---|---|---|---|---|---|
| proj (Ours) | 37.6 | 57.9 | 22.0 | 10.3 | 34.8 | 9.1 | 28.6 | 2.37M |
| proj+head | 35.5 | 58.2 | 21.2 | 9.7 | 34.9 | 9.0 | 28.1 | 3.76M |
| proj+enc+dec+head | 35.6 | 56.6 | 22.3 | 9.3 | 34.2 | 8.4 | 27.7 | 11.45M |

### C. Ablations on the Proposed Modules

Table II presents ablation studies of the proposed modules under the 1-shot setting. Using only text prototypes as classifiers fails to transfer the detector to target domains, as a single class name cannot capture the nuances of fine-grained or abstract categories. In contrast, our stage-wise prior injection effectively transforms domain-specific features into domain-invariant ones: QSC enhances feature discriminability, SQR improves representations with support-guided cues, and CAD ensures cross-modal consistency and inter-class separation. The full model, incorporating all three components, achieves the highest performance (28.6% mAP), demonstrating the effectiveness of DP-CDFSOD's overall architecture.

Fig. 6 shows the qualitative results of the proposed method compared to single-modal prototype supervision. As discussed in Sec. I and Fig. 1, visual prototypes are sensitive to domain shifts and image quality, which makes detection challenging in low-quality underwater scenes such as UODD. Meanwhile, text prototypes may not fully cover domain-specific label semantics, leading to difficulties in fine-grained categories like insects in ArTaxOr. Our progressive cross-modal design introduces domain-invariant cues, leveraging the complementary strengths of both modalities to mitigate individual weaknesses and enhance domain generalization.

### D. Efficient Target Domain Adaptation

A key advantage of our framework is efficient adaptation, which only requires fine-tuning a few projection layers with

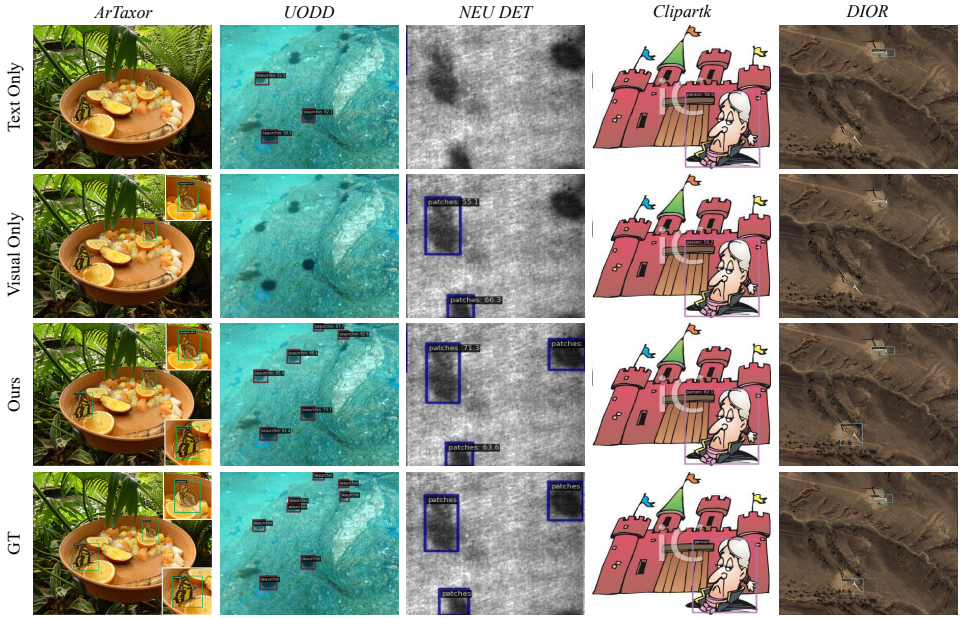


Fig. 6. Qualitative comparison of detection results using text-only prototypes, visual-only prototypes, and our full proposed approach. Our method leverages the complementary strengths of both modalities, enabling successful transfer to various domains.

TABLE IV
MODEL EFFICIENCY COMPARISON. "PROTO" MEANS PROTOTYPE OVERHEAD. THE RED FONT DENOTES IMPROVEMENTS, WHEREAS BLUE FONT INDICATES DECREASES.

| Method | mAP↑ | FPS↑ | GFLOPs↓ | Proto(s)↓ | TrainP.↓ |
|---|---|---|---|---|---|
| CDFormer [12] | 26.8 | 3.87 | 2129.29 | 161.92 | 19.88M |
| Ours | 30.5 | 3.33 | 1223.99 | 7.49 | 2.37M |
| gain | +3.7 | -0.54 | -905.30 | -154.43 | -17.51M |

minimal trainable parameters. As shown in Table III, fine-tuning solely the projection layer achieves the highest 28.9% mAP with only 2.37M trainable parameters. In contrast, fine-tuning the detection head or all components yields lower performance (28.1% and 27.7% mAP, respectively) while requiring significantly more parameters (3.76M and 11.45M). Thanks to the progressive injection of domain-invariant visual and semantic knowledge throughout the framework, features learned during source domain training become more generalizable across domains. As a result, adapting a small number of projector parameters is sufficient for domain transfer.

Table IV compares the efficiency of CDFormer [12] and the proposed method in terms of accuracy, runtime, and model complexity. All evaluation experiments are conducted on a single NVIDIA RTX 3090 GPU. Specifically, trainable parameters drop from 19.88M to 2.37M (88.1% reduction), and the inference speed (fps) is comparable, showing that the introduced injection modules (QSC, SQR, and CAD) are effective and lightweight. Furthermore, the prototype computation overhead (Proto(s)) is dramatically reduced from 161.92s to 7.49s, a 95.4% reduction, indicating the high efficiency of our prototype module. CDFormer requires multiple iterations to extract visual prototypes from support images, with each iteration involving full forward passes through the backbone and transformer encoder. In contrast, our method constructs the support set prototypes in a single iteration, resulting in significant reductions in prototype processing time. Additionally, by using a lower-resolution query image (800 compared to 1152), the GFLOPs decrease from 2129.29 to 1223.99 (42.5% reduction). Overall, the results show that our method achieves higher accuracy with significantly lower computational burden and trainable parameters, offering a superior trade-off between performance and efficiency compared to CDFormer.

## V. CONCLUSION

The inherent limitations of single-modality learning motivate us to develop a systematic approach for progressively refining query image features throughout the Encoder-Decoder-Head architecture. Therefore, we present DP-CDFSOD, a detector that progressively injects domain-invariant cross-modal priors to learn robust, semantics-driven representations. Through three lightweight modules (QSC, SQR, and CAD), the framework stabilizes query image features, refines object queries with support-aware priors, and anchors classification with textual prototypes. This staged integration enables efficient adaptation by fine-tuning only a few projection layers (2.37M parameters) and achieves the state-of-the-art results with +3.7% mAP (1-shot) on the CD-FSOD benchmark.

## ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (62495081), in part by the Guangxi Science and Technology Project (AA24263013).

## REFERENCES


[1] Y. Fu, Y. Wang, Y. Pan, L. Huai, X. Qiu, Z. Shangguan, T. Liu, Y. Fu, L. Van Gool, and X. Jiang, "Cross-domain few-shot object detection via enhanced open-set object detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 247–264.

[2] J. Pan, Y. Liu, X. He, L. Peng, J. Li, Y. Sun, and X. Huang, "Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025.

[3] G. Zhang, Z. Luo, K. Cui, S. Lu, and E. P. Xing, "Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, no. 11, pp. 12 832–12 843, 2022.

[4] G. Han, S. Huang, J. Ma, Y. He, and S.-F. Chang, "Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36, no. 1, 2022, pp. 780–789.

[5] Y. Xu, M. Zhang, C. Fu, P. Chen, X. Yang, K. Li, and C. Xu, "Multi-modal queried object detection in the wild," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[6] G. Han and S.-N. Lim, "Few-shot object detection with foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 28 608–28 618.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[8] K.-Y. Lin, H. Wang, W. Ren, and K. Han, "Panoptic captioning: An equivalence bridge for image and text," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[9] F. M. A. Mazen, "Arthropod taxonomy orders object detection in artaxor dataset using yolox," *Journal of Engineering and Applied Science (JEAS)*, vol. 70, p. 113, 2023.

[10] L. Jiang, Y. Wang, Q. Jia, S. Xu, Y. Liu, X. Fan, H. Li, R. Liu, X. Xue, and R. Wang, "Underwater species detection using channel sharpening attention," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 4259–4267.

[11] K.-Y. Lin, H. Ding, J. Zhou, Y. Peng, Z. Zhao, C. C. Loy, and W. Zheng, "Rethinking clip-based video learners in cross-domain open-vocabulary action recognition," *CoRR*, vol. abs/2403.01560, 2024.

[12] B. Meng, X. Zhang, P. Li, Z. Wu, Y. Li, W. Zhao, B. Yu, and H.-L. Shen, "Cdformer: Cross-domain few-shot object detection transformer against feature confusion," in *Proceedings of the IEEE International Conference on Multimedia*, 2025.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015, pp. 91–99.

[14] Q. Mo, Y. Gao, S. Fu, J. Yan, A. Wu, and W.-S. Zheng, "Bridge past and future: Overcoming information asymmetry in incremental object detection," in *European Conference on Computer Vision*, 2024, pp. 463–480.

[15] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, X. Xie, and W.-S. Zheng, "Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 14 987–14 997.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 30, 2020, pp. 5998–6008.

[17] S. Fu, J. Yan, Y. Gao, X. Xie, and W.-S. Zheng, "Asag: Building strong one-decoder-layer sparse detectors via adaptive sparse anchor generation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 515–11 524.

[18] S. Fu, J. Yan, Q. Yang, X. Wei, X. Xie, and W.-S. Zheng, "Frozen-detr: Enhancing detr with image understanding from frozen foundation models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024, pp. 105 949–105 971.

[19] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "Deim: Detr with improved matching for fast convergence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 15 162–15 171.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.

[22] Y. Gao, K.-Y. Lin, J. Yan, Y. Wang, and W.-S. Zheng, "Asyfod: An asymmetric adaptation paradigm for few-shot domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3261–3271.

[23] Y. Du *et al.*, "Text generation and multi-modal knowledge transfer for few-shot object detection," *Pattern Recognition (PR)*, vol. 161, p. 111283, 2025.

[24] A.-K. Nguyen Vu, Q.-T. Truong, V.-T. Nguyen, T. D. Ngo, T.-T. Do, and T. V. Nguyen, "Multi-perspective data augmentation for few-shot object detection," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025, pp. 350–366.

[25] C. Liu, X. Xiang, Z. Duan, W. Li, Q. Fan, and Y. Gao, "Don't need retraining: A mixture of detr and vision foundation models for cross-domain few-shot object detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[26] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research (TMLR)*, 2024.

[27] Z. Shangguan, D. Seita, and M. Rostami, "Cross-domain multi-modal few-shot object detection via rich text," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 6570–6580.

[28] G. Yu, J. Zhu, J. Yao, and B. Han, "Self-calibrated tuning of vision-language models for out-of-distribution detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, pp. 56 322–56 348.

[29] W. Xiong, "Cd-fsod: A benchmark for cross-domain few-shot object detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[30] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 280–296.

[31] X. Zhang, Y. Liu, Y. Wang, and A. Boularias, "Detect everything with few examples," in *Proceedings of the Conference on Robot Learning (CoRL)*, vol. 270, 2024, pp. 3986–4004.

[32] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 13669, 2022, pp. 350–368.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[34] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5001–5009.

[35] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS JPRS)*, pp. 4259–4267, 2020.

[36] A. Saleh, I. H. Laradji, D. A. Konovalov, M. Bradley, D. Vazquez, and M. Sheaves, "A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis," *Scientific Reports*, vol. 10, p. 14671, 2020.

[37] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects," *Applied Surface Science*, vol. 285, pp. 858–864, 2013.

[38] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 38–55.