Rethinking Prototype-based Similarity Learning for Few-Shot Object Detection

KunHo Heo*, Seungjae Kim*, Wongyu Lee*, SuYeon Kim, and MyeongAh Cho†

Kyung Hee University, Republic of Korea
{hkh7710, tmdwo8814, fovert, spoiuy3, maycho}@khu.ac.kr

Abstract. Few-shot object detection aims to detect novel object categories from only a few labeled examples, avoiding costly large-scale annotation. Recent prototype-based similarity learning approaches enable training-free adaptation by matching query features with class prototypes. However, they suffer from two fundamental limitations: (i) **class confusion** arising from inter-class similarity margin collapse, and (ii) **insufficient visual cues** for precise localization, as similarity scores capture only class-level semantic affinity while providing limited spatial information. To address these issues, we introduce two complementary components. **Text-Anchored Semantic Mask (TSMa)** leverages class-level text features as semantic anchors to identify semantically aligned channels through channel-wise interaction between visual and text features. By suppressing style-induced spurious responses and emphasizing class-intrinsic signals, TSMa enlarges inter-class similarity margins and mitigates class confusion. We further propose **Stage-Aligned Hierarchical Autoregressive Regression (SHARe)**, which reformulates localization as a hierarchical autoregressive process that progressively refines bounding boxes across multiple stages. SHARe leverages the layer-wise characteristics of ViT representations by aligning feature abstraction levels with regression stages: deeper layers guide early coarse localization, while shallower layers rich in edge and texture cues refine spatial details in later stages. Experiments on COCO demonstrate a new state of the art, outperforming the previous best by **+10.1 nAP**, with extensive analysis validating each component. The code is available at `https://github.com/VisualScienceLab-KHU/ReSet`.

Keywords: Object Detection and Recognition · Few-shot Learning · Prototype-based Similarity Learning

1 Introduction

Object detection is a fundamental problem in computer vision [5,6,17,31], aiming to jointly localize objects and classify their semantic categories within an image. Despite remarkable progress in recent years, state-of-the-art detectors heavily

* Equal contribution
† Corresponding author

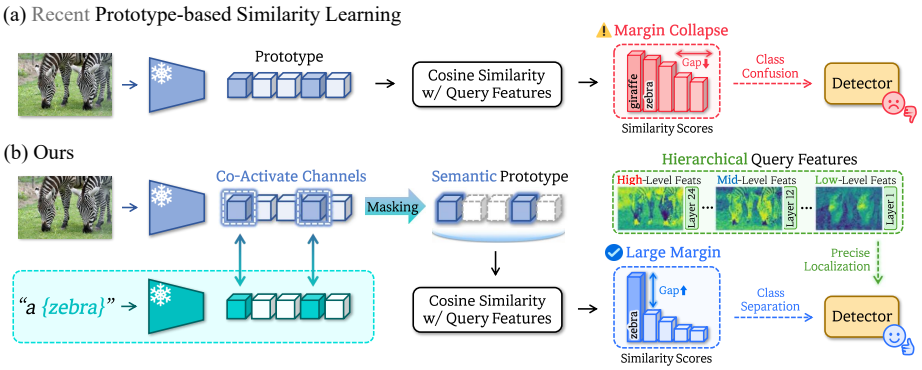

Fig. 1: (a) Recent prototype-based similarity learning methods construct class prototypes solely from visual features, and similarity scores computed with query features can cause inter-class margin collapse, leading to class confusion. (b) Our method constructs *Semantic Prototypes* by leveraging text features as anchors to retain only co-activated channels, enlarging inter-class similarity margins and thereby improving class separability. Additionally, multi-level hierarchical ViT features are injected to compensate for insufficient visual cues, enabling precise localization.

rely on large-scale datasets with exhaustive bounding-box annotations. In practical applications, however, novel object categories continuously arise, making it prohibitively expensive and often infeasible to collect sufficient labeled samples for every emerging class. As a practical solution to this limitation, Few-shot Object Detection (FSOD) [1,35] has emerged as a promising paradigm that enables the detection of novel classes from only a few annotated examples.

Early FSOD approaches [11, 29, 32, 33] first train a detector on *base classes* with abundant annotations and then fine-tune it on *novel classes* using a small *support set*. However, such fine-tuning-based approaches often suffer from training instability and overfitting due to the scarcity of support samples, limiting generalization to novel categories. To mitigate this issue, recent works [12,39,42] adopt a training-free, prototype-based similarity learning that handles novel classes at inference time without additional fine-tuning. Specifically, class prototypes are constructed by extracting and aggregating features from support images. Given a query image, dense features are obtained using a ViT backbone [28], and class-wise similarity scores are computed (e.g., dot products) between query features and class prototypes to guide both classification and localization. DE-ViT [39] and PiDiViT [42] are representative examples of this similarity-based learning approach. While they achieve strong performance without fine-tuning, they share two fundamental limitations.

The first limitation is **class confusion caused by inter-class similarity margin collapse.** Methods under the prototype-based similarity learning [39,42] classify objects by leveraging similarity scores between query features and class prototypes, where the relative separation among these scores plays a crucial role. When multiple classes produce comparable scores, leading to inter-class margin collapse, as shown in Fig. 1 (a), the model struggles to establish

clear decision boundaries, resulting in systematic misclassification that propagates through subsequent detection stages. The second limitation is **insufficient visual cues for precise localization.** In prototype-based similarity learning, localization is also primarily conditioned on class-wise similarity scores. Since these scores are computed via cosine similarity and thus mainly encode class-level semantic affinity (mostly based on vector direction), they provide limited spatial information such as object boundaries, shapes, and scale, making precise localization inherently challenging.

To mitigate **(i) class confusion**, we introduce **Text-Anchored Semantic Mask (TSMa)**. Visual features extracted from support images inevitably capture not only class-discriminative cues but also style-related components shared across different classes. Such style-driven components introduce spurious similarity responses across semantically unrelated classes, leading to inter-class similarity margin collapse. In contrast, text features provide a compact representation of class semantics, are largely free from such style components and thus carry a higher proportion of class-intrinsic information. Motivated by this property, as shown in Fig. 1 (b), we leverage text-derived class features as semantic anchors and identify semantically aligned channels via channel-wise interaction between visual and text features. The channels that are not jointly activated are then masked out, ensuring that similarity computation is dominated by semantically consistent channels. By amplifying contributions from class-intrinsic channels while attenuating style-induced responses, TSMa enlarges inter-class similarity margin and mitigates class confusion, as empirically validated in Sec. 4.4.

To address **(ii) lack of visual cues for localization**, we propose **Stage-Aligned Hierarchical Autoregressive Regression (SHARe)**. In existing prototype-based similarity learning, localization is performed in a single step conditioned on similarity scores, which provide limited structural and boundary information. Instead, SHARe reformulates localization as a hierarchical autoregressive process that refines bounding boxes across multiple stages. As shown in Fig. 1 (b), we further leverage the layer-wise property of the ViT backbone representations [8, 21] by injecting features of varying levels into each regression stage: deeper layers, which encode high-level semantics and global context, guide early coarse localization, while shallower layers, rich in low-level cues such as edges and textures, are progressively introduced to refine spatial details in later stages. This stage-aligned feature injection compensates for the representational limitations of similarity scores and enables precise object localization.

We evaluate our approach on the COCO benchmark [23] under one-shot and 10/30-shot settings reporting novel-class performance (nAP/nAP50/nAP75). Our method establishes a new state of the art, outperforming the previous best by +10.1 nAP, +7.1 nAP50, and +9.8 nAP75 in the 30-shot setting. Furthermore, extensive analysis and ablation studies provide systematic evidence for the effectiveness of each component and design rationale, offering a strong and reproducible baseline for prototype-based similarity learning.

– We identify class confusion induced by *inter-class similarity margin collapse* as a key bottleneck in prototype-based similarity learning, and pro-

pose **Text-Anchored Semantic Mask (TSMa)** that uses text features as semantic anchors to suppress style-induced spurious similarity.
- We propose **Stage-Aligned Hierarchical Autoregressive Regression (SHARe)**, which performs progressive box refinement via stage-aligned injection of hierarchical ViT features, enabling precise localization.
- We achieve a new state of the art on COCO benchmark with substantial gain over the previous SOTA (e.g. +10.1 nAP), and validate our contributions through rigorous analysis, establishing a strong baseline for FSOD.

2 Related Work

Few-Shot Object Detection (FSOD). FSOD aims to detect novel categories from only a handful of labeled examples, leveraging knowledge transferred from base categories with abundant annotations [1, 22, 35]. Most benchmarks follow a two-stage protocol that trains a detector on base classes and then adapts it to disjoint novel classes under a few-shot setting [33], evaluated in a generalized setting that includes both base and novel classes. Existing methods can be broadly grouped into finetuning-based and meta-learning-based approaches. Finetuning-based methods [11,13,29,32–34] adapt a pretrained detector to novel data in a simple and practical manner, but they are sensitive to training conditions and prone to overfitting toward base classes, leading to degraded novel class performance. Meta-learning methods train episodically across tasks to enable rapid adaptation. Early works rely on prototype-based matching [10,19,36], while later designs introduce richer support–query interactions such as dense feature interaction or transformer conditioning [3, 14–16, 38] to improve adaptability. These advances come at the cost of higher computation and potential amplification of support-set bias in low-shot regimes. Recently, the field has increasingly leveraged the rich representations of large pretrained vision transformers (ViTs) [28], often keeping the backbone frozen and shifting adaptation to inference-time mechanisms based on prototype-based similarity scores rather than heavy fine-tuning. Representative methods such as DE-ViT [39] and PiDiViT [42] demonstrate that freezing the backbone and relying on prototype-based similarity scores can simultaneously alleviate the base-biased overfitting of finetuning-based methods and the computational overhead of meta-learning approaches. This prototype-based similarity learning has emerged as a novel direction in modern FSOD.

Prototype-based Similarity Learning. Early few-shot detectors construct class prototypes from support examples and use them to modulate or classify query features [19, 36]. Subsequent meta-learning methods improve support–query alignment through dense feature interactions [3, 14–16, 38], yet at the cost of heavy computation and often benefiting from additional finetuning to achieve competitive novel-class accuracy. Departing from both lines, a recent paradigm keeps a pretrained ViT backbone [6, 28] and instead leverages prototype-based similarity scores as the core detection representation. DE-ViT [39] formalizes this direction by projecting query ViT features onto a prototype-spanned subspace via dot-product similarity and localizing objects via iterative

region propagation rather than direct coordinate regression. PiDiViT [42] inherits this pipeline and observes that frozen ViT features exhibit blurred boundary responses and excessively smooth center-to-boundary transitions, degrading object–background distinction. It mitigates these issues through pixel difference convolutions for boundary sharpening and multiscale feature fusion for scale-aware detection. Beyond in-domain benchmarks, this paradigm has also been extended to cross-domain few-shot detection with domain-aware adaptations such as instance features and domain prompting [12], demonstrating its broad applicability across diverse target domains. Nevertheless, prototype-based similarity learning methods still exhibit two structural bottlenecks. First, visual prototypes inevitably encode style-related and instance-specific components shared across classes. These spurious components introduce similarity responses across semantically unrelated categories, causing inter-class similarity margin collapse and systematic class confusion. Second, similarity scores primarily encode class-level semantic affinity and provide limited geometric cues (e.g., boundaries, shape, and scale), making precise localization challenging. In this paper, we aim to address both bottlenecks of the prototype-based similarity learning by improving the reliability of similarity computation itself to mitigate inter-class confusion and by progressively leveraging hierarchical representations of the frozen ViT backbone [8, 21] to supplement the missing geometric cues.

3 Method

3.1 Overview

As illustrated in Fig. 2, the proposed framework integrates **Text-Anchored Semantic Mask (TSMa)** and **Stage-Aligned Hierarchical Autoregressive Regression (SHARe)** into a unified pipeline. During training, a class prototype is constructed by averaging the visual features of object instances from each *base* class, and a class-specific semantic mask generated via **TSMa** from the same *base* class samples is applied to it to produce the *Semantic Prototype*. Subsequently, the semantic prototypes and query features are normalized and their similarity is computed through cosine similarity, then passed through a linear projection to produce a *similarity embedding* shared across two parallel branches: classification and regression. In the classification branch, the similarity embedding is fed into a classification head to produce a per-class score. In the regression branch, a mask logit corresponding to an initial RPN [31] proposal is fed into the first stage together with the similarity embedding, and progressively refined across subsequent stages in an autoregressive manner via **SHARe**. The resulting refined mask is converted to a bounding box and combined with the per-class score to produce the final detection output. Training losses comprise a focal loss for per-class scoring and ℓ_1 and GIoU losses for bounding box regression, supplemented by auxiliary BCE and Dice losses for mask refinement supervision. At inference, prototypes and semantic masks are constructed solely from the k-shot support examples of *novel* classes without any fine-tuning, and the remaining pipeline proceeds identically.

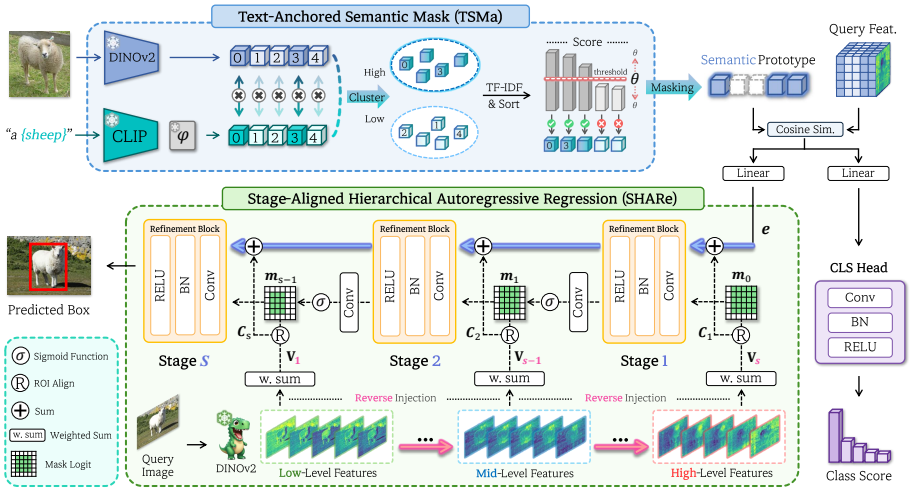

Fig. 2: Overview of the proposed framework. It consists of two main components: **Text-Anchored Semantic Mask (TSMa)**, which constructs semantic prototypes by leveraging text embeddings as anchors, and **Stage-Aligned Hierarchical Autoregressive Regression (SHARe)**, which injects multi-level ViT features in reverse order into each autoregressive regression stage for precise localization.

3.2 Text-Anchored Semantic Mask (TSMa)

The core premise of **TSMa** is that channels co-activated by both visual and text features more directly reflect class-specific semantics. Since the visual backbone DINOv2 [28] and the text backbone CLIP [30] operate in distinct feature spaces, we first learn a lightweight mapping function that projects CLIP text features into the DINOv2 feature space.

(1) Text-to-Vision Feature Mapping. We adopt the nonlinear warping proposed in Talk2DINO [2], which maps CLIP text features into the DINOv2 patch feature space. Both backbones are frozen, and only the mapping function $\psi(\cdot)$ is trained. Specifically, $\psi(\cdot)$ consists of two affine transformations with a tanh nonlinearity:

$$\psi(\mathbf{t}) = \mathbf{W}_b^\top \left(\tanh(\mathbf{W}_a^\top \mathbf{t} + \mathbf{b}_a)\right) + \mathbf{b}_b. \tag{1}$$

We use COCO [23] for training, extracting a text feature $\mathbf{t}$ for each class via a text prompt of the form `"{class}"`. The mapping function is optimized via contrastive learning [27], using the similarity between $\psi(\mathbf{t})$ and per-head region summary features obtained from DINOv2's self-attention as supervision.

(2) Channel Clustering and Scoring. For class c, let $\tilde{\mathbf{t}}_c = \psi(\mathbf{t}_c) \in \mathbb{R}^D$ denote the mapped text feature, where ψ is a pretrained mapping function and D is the feature dimension. Given the visual feature $\mathbf{v}_{c,n} \in \mathbb{R}^D$ of its n-th object instance, we measure the per-channel semantic alignment via element-wise product:

$$\mathbf{a}_{c,n} = \frac{\mathbf{v}_{c,n}}{|\mathbf{v}_{c,n}|_2} \odot \frac{\tilde{\mathbf{t}}_c}{|\tilde{\mathbf{t}}_c|_2} \in \mathbb{R}^D. \tag{2}$$

A large value of $\mathbf{a}_{c,n}[i]$ indicates that the i-th channel responds strongly to both the visual and text feature. We apply k-means clustering ($k = 2$) on the channels of $\mathbf{a}_{c,n}$, designating the higher-mean cluster as semantically-aligned active channels. This assignment is encoded as an indicator $g_{c,n}(i) \in \{0, 1\}$, where $g_{c,n}(i) = 1$ if channel i belongs to the higher-mean cluster and 0 otherwise. The activation frequency of channel i for class c is then aggregated over all N_c instances as:

$$\text{count}_c(i) = \sum_{n=1}^{N_c} g_{c,n}(i). \tag{3}$$

Subsequently, the importance of each channel is quantified as a score by adapting TF-IDF [25]:

$$\text{TF}_c(i) = \frac{\text{count}_c(i)}{N_c}, \tag{4}$$

$$\text{DF}(i) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{TF}_c(i),\ \text{IDF}(i) = \log\left(\frac{1}{\text{DF}(i) + \varepsilon}\right), \tag{5}$$

$$\text{Score}_c(i) = \text{TF}_c(i) \cdot \text{IDF}(i). \tag{6}$$

where $\mathcal{C}$ denotes the set of all classes. This formulation assigns high scores to channels that are frequently activated for class c (high TF) yet rarely activated across other classes (high IDF), ensuring that retained channels are not only class-relevant but also class-discriminative, thereby yielding a semantically discriminative channel ranking. This scoring is performed independently for base and novel classes, where novel classes rely solely on k-shot support examples.

(3) Learnable Thresholding and Masking. Let $r_{c,i}$ denote the rank of channel i for class c sorted by $\text{Score}_c(i)$ in descending order. To avoid the discontinuity of hard thresholding, we employ a learnable rank-based mask:

$$M_{c,i}(\theta_c) = \sigma\left(\frac{\theta_c - r_{c,i}}{\tau}\right), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function and τ controls the softness of the mask boundary. Channels with $r_{c,i} < \theta_c$ receive $M \approx 1$, those with $r_{c,i} > \theta_c$ receive $M \approx 0$, and boundary channels are assigned continuous weights in $(0, 1)$, forming a soft selection. The threshold θ_c is a learnable scalar initialized to the lowest rank, ensuring that nearly all channels are retained initially to prevent model collapse, with unnecessary channels progressively suppressed during training. At inference time, the threshold θ_c for novel classes is set to the mean of the thresholds learned over all base classes, requiring no additional tuning.

By applying the resulting semantic mask to the class prototype $\mathbf{p}_c \in \mathbb{R}^D$, we obtain a *Semantic Prototype* $\tilde{\mathbf{p}}_c$, defined as:

$$\tilde{\mathbf{p}}_c = \frac{\mathbf{M}_c \odot \mathbf{p}_c}{|\mathbf{M}_c \odot \mathbf{p}_c|_2}, \tag{8}$$

which is then used to compute a similarity with the query feature via cosine similarity, yielding a *similarity embedding* $\mathbf{e}$ used for both classification and localization. By ensuring that the similarity computation is governed by semantically-aligned channels, TSMa directly mitigates the inter-class similarity margin collapse that acts as a bottleneck in prototype-based similarity learning.

3.3 Stage-Aligned Hierarchical Autoregressive Regression (SHARe)

Although TSMa removes style components to obtain *semantic prototypes*, the resulting *similarity embedding* $\mathbf{e}$ still lacks the visual cue required for precise localization. Since cosine similarity is primarily sensitive to the angle between feature vectors, the similarity embedding $\mathbf{e}$ mainly encodes their directional (angular) alignment rather than geometric cues. To address this limitation, we leverage the hierarchical depth property of frozen ViT representations. At each localization stage, we utilize visual representations from different ViT layers, which enables precise localization.

(1) Multi-layer Feature Aggregation of ViT. Let $\mathbf{X}^\ell \in \mathbb{R}^{D\times H\times W}$ denote the feature map from the ℓ-th layer of a frozen ViT backbone, where D is the feature dimension, $H \times W$ is the spatial resolution of the ViT patch grid, and L is the number of ViT layers. To provide stage-wise visual cues for all S localization stages, we split the ViT layers into S stage-aligned groups, where $\{\mathcal{G}_s\}_{s=1}^{S}$ denotes the set of layer indices belonging to each group (larger s corresponds to later ViT layers), and each group contains approximately L/S indices. For instance, $\mathcal{G}_1 = \{0, \ldots, L/S - 1\}$ and $\mathcal{G}_S = \{(S-1)\cdot L/S, \ldots, L-1\}$. We then apply learnable weighted aggregation within each group to obtain the visual cue $\mathbf{V}_s$:

$$\mathbf{V}_s = \sum_{\ell\in\mathcal{G}_s} \alpha_{s,\ell}\,\mathrm{LN}(\mathbf{X}^\ell), \qquad \alpha_{s,\ell} = \frac{\exp(w_{s,\ell})}{\sum_{j\in\mathcal{G}_s}\exp(w_{s,j})}, \tag{9}$$

where $\mathrm{LN}(\cdot)$ denotes channel-wise LayerNorm that reduces layer-wise scale discrepancy and $w_{s,\ell}$ are learnable weights within each group. The resulting set of stage-wise visual cue $\{\mathbf{V}_s\}_{s=1}^{S}$ will be used in the next step to condition the stage-wise refinement in a stage-aligned manner.

(2) Stage-Aligned Hierarchical Autoregressive Regression. Given an initial *similarity embedding* $\mathbf{e}$, **SHARe** performs localization by iteratively updating the $\mathbf{e}$ over S stages, progressively incorporating stage-wise visual cues at each step. For stage $t \in \{1, \ldots, S\}$, the *refinement block* updates the embedding $\mathbf{e}_{t-1}$ to $\mathbf{e}_t$ through a sequence of Conv→BatchNorm→ReLU operations. The updated $\mathbf{e}_t$ is then passed through a 1×1 convolution followed by a sigmoid activation to produce a mask logit $\mathbf{m}_t$ representing the object region. To effectively utilize the stage-wise visual cue $\{\mathbf{V}_s\}_{s=1}^{S}$, we reinterpret the hierarchical refinement process as proceeding in the *reverse order* of ViT representation levels across the layer depth. ViT representations transition from low-level patterns to high-level semantics as layer depth increases. Specifically, earlier layers tend to capture fine-grained geometric cues such as edges and boundaries, whereas later layers

increasingly encode high-level information such as semantic context [8,21]. Leveraging this property, we design stage-aligned autoregression to utilize high-level visual cue in early stages and progressively incorporate lower-level cue in later stages. This enables SHARe to first coarsely localize the object using semantic context and then refine it by injecting geometric details for precise localization. Accordingly, we align the stage progression with the hierarchy of visual cue in reverse order:

$$(\text{stage } 1 \rightarrow \text{stage } S) \iff (\mathbf{V}_S \rightarrow \mathbf{V}_1). \tag{10}$$

At stage t, the refinement is conditioned on the visual cue $\mathbf{V}_{\pi(t)}$ with $\pi(t) = S - t + 1$, which is used to form a stage condition $\mathbf{C}_t$. Specifically, we apply RoIAlign [17] on $\mathbf{V}_{\pi(t)}$ within the Region of Interest $\mathcal{R}$ to extract a region-specific visual cue, and then project it with a stage-specific linear layer to obtain $\mathbf{C}_t$:

$$\mathbf{C}_t = \text{Linear}_t\big(\text{RoIAlign}\big(\mathbf{V}_{\pi(t)}, \mathcal{R}\big)\big)\,. \tag{11}$$

A single RoI $\mathcal{R}$ is shared across all stages, obtained by the initial RPN proposal.

Subsequently, the stage condition $\mathbf{C}_t$ is gated using the previous-stage mask logit $\mathbf{m}_{t-1}$ and injected into the latent embedding:

$$\tilde{\mathbf{e}}_{t-1} = \mathbf{e}_{t-1} + \gamma_t \cdot (\mathbf{C}_t \odot \sigma(\mathbf{m}_{t-1}))\,, \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\odot$ is element-wise product, and γ_t is a learnable scalar that controls the amount of injected visual cues. To carry forward mask predictions across stages, the previous mask logits are concatenated with the injected embedding to form the stage input:

$$\hat{\mathbf{e}}_{t-1} = \text{Concat}(\tilde{\mathbf{e}}_{t-1}, [\mathbf{m}_{t-1}, \ldots, \mathbf{m}_0]) \tag{13}$$

where $\text{Concat}(\cdot)$ concatenates tensors along the channel dimension. The refinement block then produces the updated embedding $\mathbf{e}_t$ from $\hat{\mathbf{e}}_{t-1}$, and predicts the corresponding mask logit $\mathbf{m}_t$. Finally, each stage's mask logit $\mathbf{m}_t$ is converted into a bounding box via the spatial integral layer [39]. More details are provided in the supplementary materials.

4 Experiment

4.1 Dataset and Evaluation Metrics

We evaluate on the COCO dataset [23] following standard base/novel class splits [33, 37]. Base classes are used for training, while novel classes are held out for evaluation. Novel class performance is measured using nAP, nAP50, and nAP75, and base class performance using bAP. For one-shot evaluation, we follow the conventional protocol [26] that divides the 80 COCO categories into four equal partitions (Split-1/2/3/4), using three partitions as base classes and the remaining one as novel classes. For Pascal VOC [9], we evaluate on Split-1/2/3 across 1/2/3/5/10-shot settings, reporting nAP50.

4.2 Implementation Details

We build on a two-stage detection framework [17], adopting a pre-trained RPN [41] trained solely on base classes for proposal generation. The vision backbone is a frozen DINOv2 [28] ViT-L, and text features for TSMa are extracted using a frozen CLIP [30] ViT-L text encoder. Background class prototypes are extracted from COCO-Stuff [4]. During evaluation, the model is tested on images containing objects of both base and novel classes, selecting the top-T (T=5) categories based on the similarity between query features and class prototypes.

Table 1: Results on COCO 2014 few-shot benchmark.

Method	Backbone	Finetune	10-shot				30-shot			
			bAP	nAP	nAP50	nAP75	bAP	nAP	nAP50	nAP75
FSCE (CVPR'21) [32]	RN101	✓	-	11.9	-	10.5	-	16.4	-	16.2
Retentive RCNN (CVPR'21) [11]	RN101	✓	39.2	10.5	19.5	9.3	39.3	13.8	22.9	13.8
HeteroGraph (ICCV'21) [14]	RN101	✓	-	11.6	23.9	9.8	-	16.5	31.9	15.5
FsDetView (TPAMI'22) [34]	RN50	✓	6.4	7.6	-	-	9.3	12	-	-
Meta Faster RCNN (AAAI'22) [15]	RN101	✓	-	12.7	25.7	10.8	-	16.6	31.8	15.8
LVC (CVPR'22) [20]	ViT-S/8	✓	28.7	19.0	34.1	19	34.8	26.8	45.8	27.5
CrossTransformer (CVPR'22) [16]	PVTv2-B2-Li	✓	-	17.1	30.2	17	-	21.4	35.5	22.1
NIFF (CVPR'23) [13]	RN101	✓	39.0	18.8	-	-	39.0	20.9	-	-
DiGeo (CVPR'23) [24]	RN101	✓	39.2	10.3	18.7	9.9	39.4	14.2	26.2	14.8
DE-ViT (CoRL'24) [39]	ViT-L/14	✗	29.4	34.0	52.9	37.0	29.5	34.0	53.0	37.2
CD-ViTO (ECCV'24) [12]	ViT-L/14	✗	-	35.3	54.9	37.2	-	35.9	54.5	38.0
PiDiViT (ICCV'25) [42]	ViT-L/14	✗	32.3	37.0	57.6	40.9	31.4	37.3	58.5	41.7
Ours	ViT-L/14	✗	**41.2**	**47.1**	**65.5**	**51.5**	**41.3**	**47.4**	**65.6**	**51.5**

4.3 Comparisons with SOTA Methods

Quantitative Evaluation. Our method achieves substantial improvements over the previous methods including DE-ViT [39] and PiDiViT [42], establishing a new state of the art across diverse metrics and shot settings. As reported in Tab. 1, we record gains of nAP +10.1, nAP50 +7.9, nAP75 +10.6, and bAP +8.9 in the 10-shot setting, and nAP +10.1, nAP50 +7.1, nAP75 +9.8, and bAP +9.9 in the 30-shot setting. Tab. 2 further shows consistent gains across all four COCO one-shot splits, with an average nAP50 improvement of +10.9 over the previous SOTA. On Pascal VOC [9], Tab. 3 reports an average nAP50 gain of +6.9% across all splits and shot settings. Taken together, these results firmly establish the superiority of our approach across two datasets and diverse shot conditions.

Table 2: Results on COCO 2017 one-shot benchmark.

Method	Backbone	nAP50				
		Split-1	Split-2	Split-3	Split-4	Avg
CoAE [18]	RN50	23.4	23.6	20.5	20.4	22.0
AIT [7]	RN50	26.0	26.4	22.3	22.6	24.3
SaFT [40]	RN101	27.8	27.6	21.0	23.0	24.9
BHRL [37]	RN50	26.1	29.0	22.7	24.5	25.6
DE-ViT [39]	ViT-L/14	27.4	33.2	27.1	26.1	28.4
PiDiViT [42]	ViT-L/14	32.0	38.6	30.4	30.2	32.8
Ours	ViT-L/14	**46.6**	**48.0**	**42.5**	**37.9**	**43.7**

Table 3: nAP50 results on Pascal VOC few-shot benchmark.

Method	Backbone	Novel Split 1					Novel Split 2					Novel Split 3					Avg
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
TFA [33]	RN101	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	39.9
Multi-Relation Det [10]	RN50	37.8	43.6	51.6	56.5	58.6	22.5	30.6	40.7	43.1	47.6	31.0	37.9	43.7	51.3	49.8	43.1
Retentive RCNN [11]	RN101	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1	41.1
Meta Faster R-CNN [15]	RN101	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6	50.5
LVC [20]	ViT-S/8	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6	52.3
CrossTransformer [16]	PVTv2	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7	50.9
HeteroGraph [14]	RN101	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5	48.0
DiGeo [24]	RN101	37.9	39.4	48.5	58.6	61.5	26.6	28.9	41.9	42.1	49.1	30.4	40.1	46.9	52.7	54.7	44.0
NIFF [13]	RN101	**62.8**	**67.2**	68.0	70.3	68.8	38.4	42.9	54.0	56.4	54.0	56.4	62.1	61.2	64.1	63.9	59.4
DE-ViT [39]	ViT-L/14	55.4	56.1	68.1	70.9	71.9	43.0	39.3	58.1	61.6	63.1	58.2	64.0	61.3	64.2	67.3	60.2
PiDiViT [42]	ViT-L/14	57.3	56.9	68.1	73.7	73.1	43.5	44.7	61.2	61.2	62.4	58.4	64.2	61.4	64.1	67.6	61.2
Ours	ViT-L/14	59.2	64.3	**75.6**	**79.5**	**79.7**	**50.8**	**56.0**	**67.6**	**67.7**	**67.3**	**65.5**	**72.0**	**68.9**	**72.8**	**74.6**	**68.1**

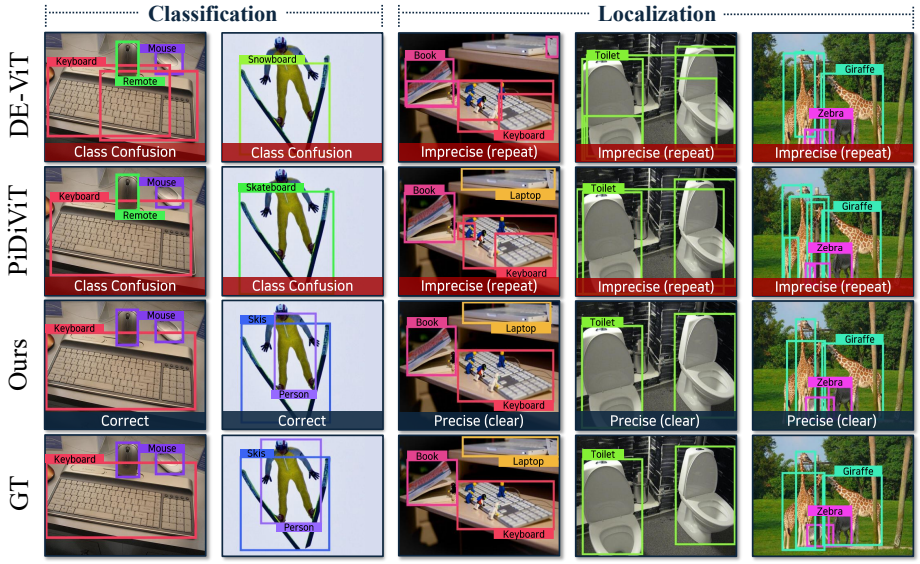

Fig. 3: Qualitative comparison with existing methods (30-shot).

Qualitative Evaluation. Fig. 3 compares detection results of DE-ViT [39], PiDiViT [42], and our method from classification and localization perspectives. **Classification**: in the first and second columns, DE-ViT and PiDiViT misclassify a `Mouse` as `Remote`, whereas our method correctly detects it. **Localization**: in the third through fifth columns, both baselines produce multiple redundant detections for a single object, while our method yields a single accurate prediction consistent with the ground truth. These results directly validate the contribution of each component: the classification improvements reflect **TSMa**'s ability to mitigate *class confusion* from inter-class similarity margin collapse, while the localization improvements demonstrate **SHARe**'s effectiveness in compensating for *insufficient visual cues* through hierarchical ViT feature injection. Additional qualitative samples are provided in the supplementary materials.

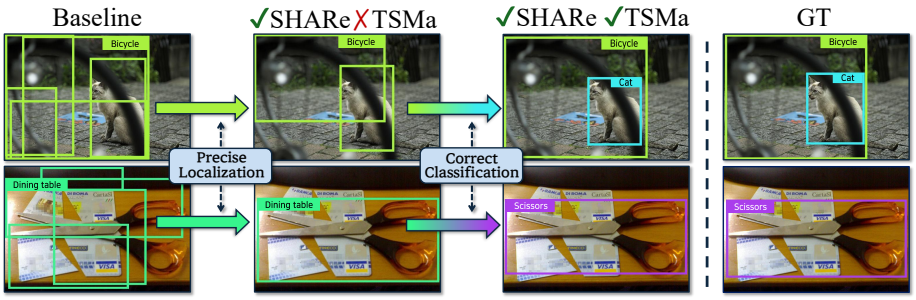

Fig. 4: Visualization of ablation results on the proposed components.

4.4 Discussions

Table 4: Ablation results on the proposed components.

SHARe	TSMa	bAP	nAP	nAP50	nAP75
–	–	30.0	34.3	53.6	36.8
✓	–	37.1	43.1	60.8	47.4
–	✓	38.6	43.7	63.5	46.1
✓	✓	**41.3**	**47.4**	**65.6**	**51.5**

Ablation Studies on the Proposed Components. We present ablation results both quantitatively (Tab. 4) and qualitatively (Fig. 4). SHARe alone yields consistent gains across all metrics, with the largest improvement on nAP75 (+10.6), which is most sensitive to localization precision at high IoU, suggesting that SHARe most prominently enhances *localization*. This is further corroborated in Fig. 4, where redundant box detections observed in the baseline are resolved upon adding SHARe, demonstrating that injecting hierarchical ViT features from high-level semantics to low-level spatial cues effectively induces precise localization. TSMa alone likewise improves across all metrics, with the largest gain on nAP50 (+9.9), a metric more sensitive to classification under its lenient IoU criterion, suggesting that TSMa most prominently enhances *inter-class discriminability*. Fig. 4 further shows that misclassifications present in the SHARe-only model are corrected upon adding TSMa, demonstrating that constructing Semantic Prototypes by suppressing class-agnostic channels and emphasizing class-specific ones leads to improved inter-class separation.

Table 5: Ablation results on TF-IDF scoring in TSMa.

TF-IDF	bAP	nAP	nAP50	nAP75
✗	40.1	45.8	64.1	50.0
✓	**41.3**	**47.4**	**65.6**	**51.5**

TF-IDF Scoring in TSMa. In the TSMa pipeline, after partitioning channels into high/low-mean clusters via k-means clustering, we further refine per-channel scores through TF-IDF scoring, whose effect is reported in Tab. 5. Among the per-channel magnitudes of the element-wise product between visual and text features, certain channels exhibit consistently large responses across all classes regardless of class identity, introducing confusion in inter-class discrimination. TF-IDF scoring mitigates this by penalizing such class-agnostic channels while emphasizing class-specific ones, and the performance drop when removing TF-IDF scoring corroborates this effect.

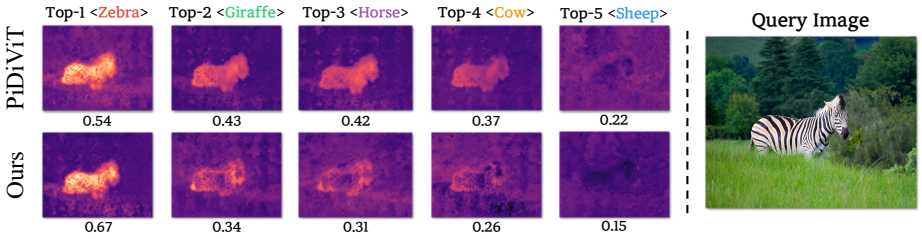

Fig. 5: Query–prototype similarity heatmaps of Top-5 class prototypes.

Table 6: Top-k accuracy comparison.

Method	Accuracy (%)			nAP50
	Top-1	Top-3	Top-5	
PiDiViT [42]	63.2	72.1	87.2	58.5
Ours	**84.3**	**89.6**	**96.8**	**65.6**
Δ	(+21.1%)	(+17.5%)	(+9.6%)	(+7.1%)

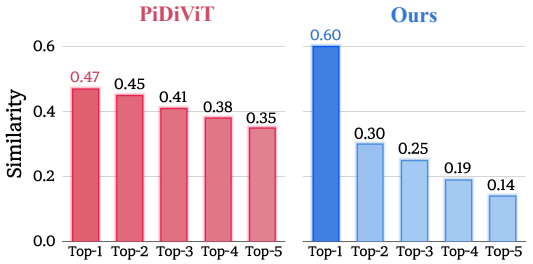

Fig. 6: Comparison of average similarity scores.

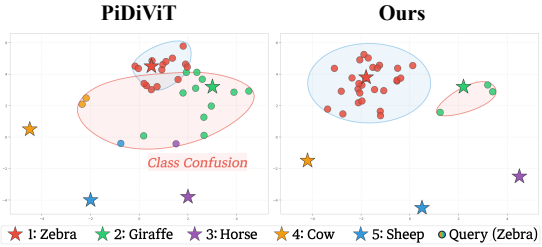

Fig. 7: Comparison of t-SNE visualization.

Inter-Class Similarity Margin Enlargement via TSMa. To validate that TSMa resolves inter-class similarity margin collapse and improves class separability, we conduct analyses on the COCO [23] test set. Tab. 6 reports Top-k Accuracy, defined as the probability that the ground-truth class ranks among the top-k most similar prototypes to the query feature, where our method outperforms PiDiViT by +21.1%, +17.5%, and +9.6% at Top-1, Top-3, and Top-5, respectively. Fig. 6 visualizes the average similarity scores of the top-5 class prototypes across all query features, showing substantially larger inter-class margins in our method. The enlarged Top-1/Top-2 gap indicates reduced confusion between the most confident class and its nearest competitor, while the wider margins across all top-5 classes corroborate improved inter-class discriminability. Fig. 5 further illustrates this through similarity heatmaps: in PiDiViT, the gap between Top-1 and Top-2 is narrow, and Top-2 through Top-4 classes maintain high scores, whereas our method produces a pronounced gap beyond Top-1. Fig. 7 presents t-SNE maps of 30 `Zebra` query features alongside the top-5 prototypes, where PiDiViT shows clear class confusion with many queries clustering near incorrect prototypes, while ours largely eliminates confusion. Collectively, these analyses confirm that TSMa successfully mitigates inter-class similarity margin collapse and improves class separability.

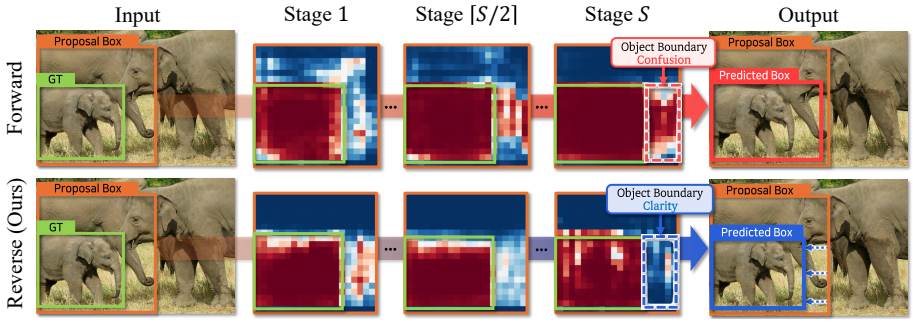

Fig. 8: Visual comparison of ViT feature injection strategies in SHARe.

Ablations on SHARe Feature Injection. Stage-wise injection is central to SHARe, enabling the use of hierarchical ViT cues during autoregressive mask refinement. We report the effect of varying the injection order of ViT feature groups across stages in Tab. 7. Compared to "`None`", "`Forward`" substantially improves all metrics, confirming that explicit stage-wise visual conditioning is crucial as similarity embeddings alone provide insufficient geometric cues. However, "`Reverse`" achieves the best performance across all metrics, with a particularly large gain on nAP75 (+3.3 over "`Forward`"), demonstrating superior localization at high IoU. This validates our key insight that progressive localization is best aligned with the reverse of ViT's depth hierarchy: high-level semantic features establish coarse object extent in early stages, while lower-level spatial cues progressively sharpen boundaries in later stages. Fig. 8 further corroborates this qualitatively; while "`Forward`" produces reasonable early-stage localization, it increasingly confuses overlapping instances as refinement progresses, whereas "`Reverse`" consistently resolves such confusion across stages.

Table 7: Ablation results on injection strategies in SHARe.

Strategy	bAP	nAP	nAP50	nAP75
None	38.6	43.7	63.5	46.1
Forward	40.0	45.9	64.4	48.2
Reverse	**41.3**	**47.4**	**65.6**	**51.5**

5 Conclusion

We address two fundamental limitations of prototype-based similarity learning in FSOD: inter-class similarity margin collapse causing class confusion, and insufficient visual cues for precise localization. To mitigate the former, **TSMa** leverages text features as semantic anchors to identify semantically aligned channels, suppressing style-induced spurious similarity and enlarging inter-class similarity margins. To address the latter, **SHARe** reformulates localization as a hierarchical autoregressive process, injecting hierarchical ViT features in reverse depth order across refinement stages to progressively recover spatial cues lost in similarity embeddings. Experiments on COCO and Pascal VOC demonstrate substantial gains over existing methods, establishing a strong and reproducible baseline for prototype-based similarity learning in FSOD.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT)(RS-2024-00456589).

References

1. Antonelli, S., Avola, D., Cinque, L., Crisostomi, D., Foresti, G.L., Galasso, F., Marini, M.R., Mecca, A., Pannone, D.: Few-shot object detection: A survey. ACM computing surveys (CSUR) **54**(11s), 1–37 (2022)
2. Barsellotti, L., Bianchi, L., Messina, N., Carrara, F., Cornia, M., Baraldi, L., Falchi, F., Cucchiara, R.: Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22025–22035 (2025)
3. Bulat, A., Guerrero, R., Martinez, B., Tzimiropoulos, G.: Fs-detr: Few-shot detection transformer with prompting and without re-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11793–11802 (2023)
4. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chen, D.J., Hsieh, H.Y., Liu, T.L.: Adaptive image transformer for one-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12247–12256 (2021)
8. Dorszewski, T., Tětková, L., Jenssen, R., Hansen, L.K., Wickstrøm, K.K.: From colors to classes: Emergence of concepts in vision transformers. In: World Conference on Explainable Artificial Intelligence. pp. 28–47. Springer (2025)
9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
10. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4013–4022 (2020)
11. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4527–4536 (2021)
12. Fu, Y., Wang, Y., Pan, Y., Huai, L., Qiu, X., Shangguan, Z., Liu, T., Fu, Y., Van Gool, L., Jiang, X.: Cross-domain few-shot object detection via enhanced open-set object detector. In: European Conference on Computer Vision. pp. 247–264. Springer (2024)
13. Guirguis, K., Meier, J., Eskandar, G., Kayser, M., Yang, B., Beyerer, J.: Niff: Alleviating forgetting in generalized few-shot object detection via neural instance feature forging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24193–24202 (2023)

14. Han, G., He, Y., Huang, S., Ma, J., Chang, S.F.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3263–3272 (2021)
15. Han, G., Huang, S., Ma, J., He, Y., Chang, S.F.: Meta faster r-cnn: Towards accurate few-shot object detection with attentive feature alignment. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 780–789 (2022)
16. Han, G., Ma, J., Huang, S., Chen, L., Chang, S.F.: Few-shot object detection with fully cross-transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5321–5330 (2022)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
18. Hsieh, T.I., Lo, Y.C., Chen, H.T., Liu, T.L.: One-shot object detection with co-attention and co-excitation. Advances in neural information processing systems **32** (2019)
19. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8420–8429 (2019)
20. Kaul, P., Xie, W., Zisserman, A.: Label, verify, correct: A simple few shot object detection method. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14237–14247 (2022)
21. Kim, J., Kim, J., Shim, Y., Kim, J., Jung, S., Hwang, S.J.: Interpreting vision transformers via residual replacement model. arXiv preprint arXiv:2509.17401 (2025)
22. Köhler, M., Eisenbach, M., Gross, H.M.: Few-shot object detection: A comprehensive survey. IEEE transactions on neural networks and learning systems **35**(9), 11958–11978 (2023)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
24. Ma, J., Niu, Y., Xu, J., Huang, S., Han, G., Chang, S.F.: Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3208–3218 (2023)
25. Manning, C.D.: Introduction to information retrieval. Syngress Publishing, (2008)
26. Michaelis, C., Ustyuzhaninov, I., Bethge, M., Ecker, A.S.: One-shot instance segmentation. arXiv preprint arXiv:1811.11507 (2018)
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
29. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8681–8690 (2021)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

32. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7352–7362 (2021)
33. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020)
34. Xiao, Y., Lepetit, V., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3090–3106 (2022)
35. Xin, Z., Chen, S., Wu, T., Shao, Y., Ding, W., You, X.: Few-shot object detection: Research advances and challenges. Information Fusion **107**, 102307 (2024)
36. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9577–9586 (2019)
37. Yang, H., Cai, S., Sheng, H., Deng, B., Huang, J., Hua, X.S., Tang, Y., Zhang, Y.: Balanced and hierarchical relation learning for one-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7591–7600 (2022)
38. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 14832–14845 (2022)
39. Zhang, X., Liu, Y., Wang, Y., Boularias, A.: Detect everything with few examples. arXiv preprint arXiv:2309.12969 (2023)
40. Zhao, Y., Guo, X., Lu, Y.: Semantic-aligned fusion transformer for one-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7601–7611 (2022)
41. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16793–16803 (2022)
42. Zhou, H., Liu, Y., Mo, C., Li, W., Peng, B., Liu, L.: When pixel difference patterns meet vit: Pidivit for few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24309–24318 (2025)