PULSE: PERSISTENT UTILIZATION OF LIMITED SUPPORT EVIDENCE FOR CROSS-DOMAIN FEW-SHOT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain few-shot object detection (CD-FSOD) aims to localize and classify novel objects in unseen domains using only a few annotated target examples. However, fine-tuning-based adaptation can suffer from support underutilization: while limited target examples are used to update detector parameters during fine-tuning, the same target evidence need not remain directly accessible when query predictions are scored. To address this limitation, this work introduces Persistent Utilization of Limited Support Evidence (PULSE), a novel framework that continuously leverages annotated target appearances during prediction scoring. Specifically, after target-domain adaptation, PULSE caches multi-view support features in a frozen evidence bank. SSCA is then proposed to align these cached support representations with current query features and measure their support–query agreement. Based on this agreement, SSCA integrates a gated residual correction into the detector's base class logits, rewarding predictions that are consistent with the observed target support while suppressing confident yet unsupported detections. Importantly, this refinement operates only on classification scores and leaves predicted bounding box coordinates unchanged. Extensive experiments across six CD-FSOD benchmarks demonstrate the effectiveness of the proposed approach. PULSE achieves state-of-the-art mean AP under different shot settings. Furthermore, PULSE outperforms the corresponding FT-FSOD results in 16 of the 18 domain–shot settings and matches them in the remaining two, indicating that prediction-time support consultation provides a complementary signal to conventional parameter adaptation.

1 INTRODUCTION

Cross-domain few-shot object detection (CD-FSOD) adapts a source-pretrained detector to novel target categories using only K annotated bounding boxes per class (Xiong, 2023; Fu et al., 2024). Domain shifts can substantially change object appearance, scale, and boundary characteristics, while the limited support set captures only a small portion of the target-domain variation. As a result, the small annotation budget limits not only parameter adaptation but also the target evidence available for evaluating new predictions.

FT-FSOD (Yu et al., 2026) combines a Hybrid Ensemble Decoder (HED) with progressive fine-tuning and provides a strong baseline for CD-FSOD. Fine-tuning incorporates support information into the detector parameters, but the annotated target examples are not explicitly consulted when query predictions are subsequently scored. Consequently, a prediction may receive high detector confidence even when it is weakly supported by the observed target appearances. This work refers to this limitation as *support underutilization*: target evidence contributes to parameter adaptation, but is not explicitly retained as a prediction-time reference.

Support-conditioned detectors and prototype-based methods already exploit support examples during prediction. The focus here is specifically on fine-tuning-based CD-FSOD: how can the limited examples used for adaptation remain directly available when query predictions are scored? Persistent Utilization of Limited Support Evidence (PULSE) addresses this problem by retaining target-support representations after adaptation and reusing them during prediction. SSCA compares current query

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

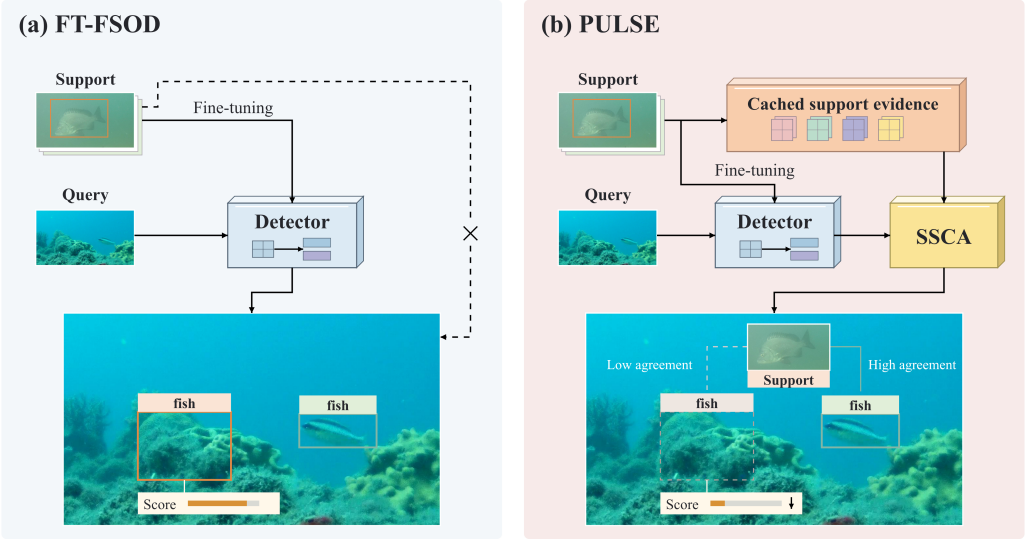

Figure 1: **Target-support use in FT-FSOD (Yu et al., 2026) and PULSE.** (a) FT-FSOD uses support annotations for target-domain fine-tuning and subsequently scores query predictions without explicitly consulting the support examples. (b) PULSE retains cached support evidence and uses SSCA to refine classification scores according to support–query agreement. Box tags denote predicted classes, and bar lengths illustrate schematic confidence. The dashed candidate illustrates score suppression at the same location. Test images from DeepFish (Fu et al., 2024) are used for illustration; the support examples, bounding boxes, and score changes are schematic.

features with the retained target evidence and uses their agreement to refine, rather than replace, the detector's base class logits. Figure 1 illustrates the difference between parameter-only scoring and the proposed evidence-consulting scoring process.

The proposed framework consists of two stages. In Stage I, HED is adapted to the target domain, and multi-view support representations are extracted and stored in a fixed evidence bank. In Stage II, the query detector is initialized from the Stage I checkpoint, while SSCA aligns current query features with the cached support evidence and converts their agreement into a gated residual correction of the base class logits. The detector therefore remains the primary source of prediction, while the retained target evidence provides an additional reference for classification-score refinement. The predicted bounding box coordinates remain unchanged.

Experiments are conducted on six CD-FSOD benchmarks, including ArTaxOr, Clipart1k, DIOR, DeepFish, NEU-DET, and UODD, under the 1-, 5-, and 10-shot settings. PULSE achieves state-of-the-art mean AP of 37.1, 46.0, and 48.5, respectively. Compared with the corresponding published FT-FSOD results of 34.9, 45.0, and 47.9 (Yu et al., 2026), PULSE improves the mean AP by 2.2, 1.0, and 0.6 points. These results show that retaining target evidence for prediction-time score refinement complements conventional target-domain adaptation.

The main contributions are summarized as follows:

- Persistent support utilization is introduced as a prediction-time design objective for fine-tuning-based CD-FSOD. PULSE preserves annotated target appearances after adaptation, allowing the same limited evidence to remain directly accessible during query scoring.
- SSCA is introduced to align cached support representations with current query features and quantify their support–query agreement. The resulting evidence is integrated with detector confidence through a gated residual correction that refines class logits while preserving the predicted bounding boxes.
- Extensive experiments on six target domains under the 1-, 5-, and 10-shot settings demonstrate the effectiveness of PULSE. The proposed framework achieves state-of-the-art mean

AP across all three annotation budgets, while the ablation study further analyzes the contribution of the main score-refinement components.

2 Related Work

Few-shot object detection. Few-shot object detection mainly exploits support examples through prediction conditioning or parameter adaptation. Meta-learning methods reweight features or model relations between support and query regions (Kang et al., 2019; Yan et al., 2019; Fan et al., 2020), while transfer-based methods emphasize selective fine-tuning, multi-scale representations, contrastive proposal learning, and decoupled optimization (Wang et al., 2020; Wu et al., 2020; Sun et al., 2021; Qiao et al., 2021). Meta-DETR extends support conditioning to image-level transformer features (Zhang et al., 2022), and BIOT studies the trade-off between adaptation and out-of-distribution generalization in query-based detectors (Yin et al., 2025). In contrast to methods that use support examples only for adaptation, PULSE retains target evidence as an explicit reference during prediction.

Foundation-model-based CD-FSOD. Vision foundation models provide transferable visual and semantic representations for open-set and example-based detection. Grounding DINO (Liu et al., 2024) introduces vision–language grounding for open-set detection, while DINOv2 (Oquab et al., 2024) provides general-purpose visual features that support example-based matching in DE-ViT (Zhang et al., 2025). SAM (Kirillov et al., 2023) and region–text models such as ViLD (Gu et al., 2022), GLIP (Li et al., 2022), and RegionCLIP (Zhong et al., 2022) provide complementary object-region and semantic information.

Recent CD-FSOD methods increasingly exploit these representations to improve target-domain adaptation. CD-ViTO (Fu et al., 2024) enhances open-set visual representations, DE-ViT (Zhang et al., 2025) performs example-based matching with DINOv2 features, and ETS (Pan et al., 2025) improves adaptation through augmentation search. Other approaches explore retrieval-guided generation (Li et al., 2025), cross-domain feature modeling (Meng et al., 2025), vision foundation-model integration (Liu et al., 2025a), and multimodal semantic information (Shangguan et al., 2025; Feng et al., 2026). Among fine-tuning-based approaches, FT-FSOD (Yu et al., 2026) combines a Hybrid Ensemble Decoder (HED) with progressive target-domain fine-tuning and is adopted as the base detector in this work. PULSE addresses a complementary problem after adaptation by preserving target-support evidence for prediction-time score refinement.

Support representations and prediction-time evidence. Prototype-based methods summarize class evidence or preserve per-sample and distributional information (Snell et al., 2017; Lee et al., 2022; Du et al., 2023). Recent CD-FSOD methods further adapt visual or multimodal prototypes (Yang & Xie, 2026; Wu et al., 2026; Wang et al., 2026), while IFC caches instance features for later retrieval (Huang et al., 2025). These methods demonstrate that retained support representations can provide useful prediction-time information. PULSE extends the role of the support set beyond target-domain fine-tuning. A fixed evidence bank keeps annotated target appearances explicitly accessible during prediction, while SSCA combines support–query agreement with detector confidence to residually refine the base HED logits rather than replacing them with a support-only score.

3 PULSE: Persistent Support Evidence and Score Calibration

3.1 Problem Setup and Overview

Let $\mathcal{S} = \{(I_{c,k}, b_{c,k}, c) : 1 \le c \le C,\ 1 \le k \le K\}$ denote the target-domain support set containing K annotated object instances for each of C target classes, where $I_{c,k}$ is the support image associated with the k-th instance of class c, and $b_{c,k}$ is its corresponding bounding box. Given $\mathcal{S}$, the goal is to localize and classify instances of the C target classes in unseen query images from the same target domain.

FT-FSOD with the Hybrid Ensemble Decoder (HED) (Yu et al., 2026) is adopted as the base detector. HED uses Grounding DINO (Liu et al., 2024) as its underlying detector, with a shared decoder stage followed by parallel prediction branches. The shared decoder provides query representations, while

the subsequent branches produce classification and localization outputs. For a query image, HED produces decoder features $Q \in \mathbb{R}^{N\times D}$, base class logits $S \in \mathbb{R}^{N\times C}$, and predicted bounding boxes $B \in \mathbb{R}^{N\times 4}$, where N denotes the number of object queries and D denotes the feature dimension. In PULSE, Q is used for support matching, S provides the base classification logits to be refined, and B provides the localization predictions that remain unchanged during score refinement. The score update is formulated at the class level. For token-level detector outputs, each class-level correction is mapped to its associated text tokens.

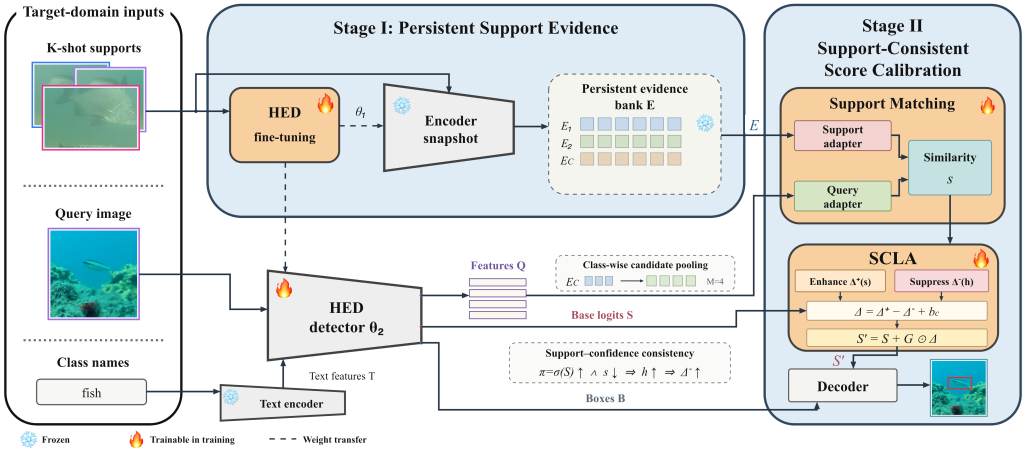

Figure 2: **Persistent support evidence and SSCA-centered calibration in PULSE.** In Stage I, HED is fine-tuned on the target support set, and the resulting encoder snapshot is frozen to construct the evidence bank E. In Stage II, the query detector is initialized with the Stage I weights and produces query features Q, base class logits S, and predicted boxes B. SSCA matches Q with the cached evidence E to obtain support agreement s and refines the base logits as $S' = S + G \odot \Delta$, while the box coordinates bypass score calibration and remain unchanged. The evidence bank remains fixed throughout Stage II training and inference. Flames indicate trainable parameters during training, whereas snowflakes indicate frozen modules or fixed evidence. All model parameters are fixed during inference. Test images from the DeepFish dataset (Fu et al., 2024) are adopted for illustration and do not correspond to the sampled adaptation episode; the displayed output box is schematic rather than an actual model prediction.

3.2 A Shared Bank of Target Appearance Evidence

Each annotated support box is represented by feature tokens extracted from three crop scales and their horizontally flipped counterparts. These view-specific features are extracted using the frozen Stage I encoder and stored in a class-structured evidence bank.

For target class c, the corresponding class-specific evidence set is denoted as

$$E_c = \{e_{c,j}\}_{j=1}^{L_c},$$

where $e_{c,j} \in \mathbb{R}^D$ denotes the j-th cached support-view feature of class c, and L_c denotes the total number of cached view features for that class. The complete evidence bank is therefore represented as

$$E = \{E_c\}_{c=1}^{C}.$$

Since several view-specific features are generated from each annotated support instance, the bank preserves complementary appearance variations across crop scales and horizontal flips without increasing the annotation budget K.

The evidence bank remains fixed throughout Stage II training and inference, providing a stable reference to the annotated target appearances. During Stage II, the query detector and SSCA continue to be optimized while this support reference remains fixed. Because the cached support features are extracted from the frozen Stage I encoder whereas the current query features continue to evolve during Stage II training, the two feature streams are not guaranteed to remain directly comparable.

SSCA therefore aligns the cached support representations and current query features into a shared comparison space before computing support–query agreement.

3.3 SUPPORT-CONSISTENT SCORE CALIBRATION

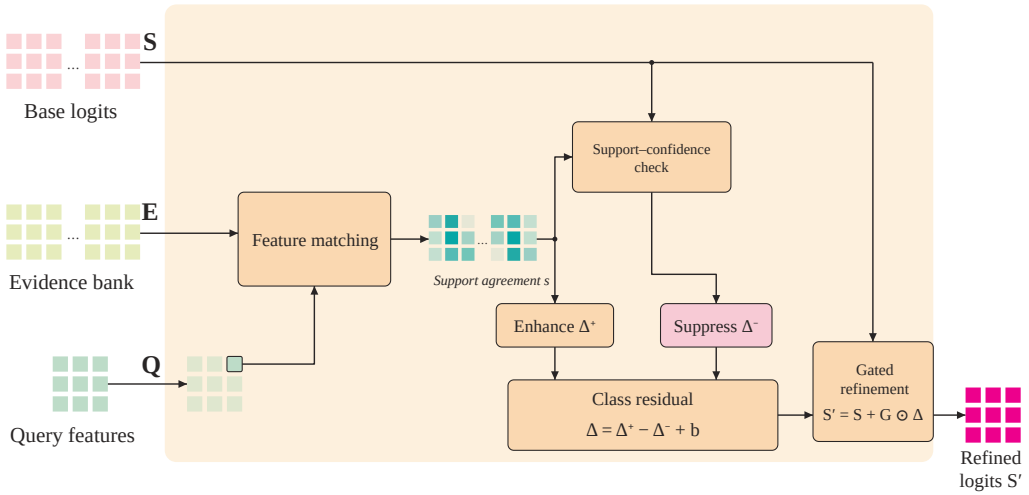

Figure 3: **The unified SSCA score-refinement path.** Feature alignment and matching compare the query features Q with the fixed evidence bank E to obtain class-wise support agreement s. Support agreement directly provides the support-based score adjustment, while its consistency with detector confidence determines an additional suppression penalty. The two terms are combined into the class residual Δ, which is further modulated by the language-score gate G before being added to the base logits S. The resulting logits S' provide the refined classification scores, while the predicted box coordinates B remain unchanged.

Feature alignment and matching. Feature alignment and matching measures how well each query representation agrees with the retained target evidence of each class. The cached support views of class c are first aggregated into M candidate appearance representations $p_{c,m}$, allowing multiple appearance modes of the same class to be retained. Because the cached support features are extracted using the frozen Stage I encoder, whereas the current query features continue to evolve during Stage II training, the two feature streams may no longer be directly comparable. Therefore, two independently parameterized residual adapters (Houlsby et al., 2019) project the query and support representations into a shared normalized comparison space. With $\bar{q}_n = A_q(q_n)$ and $\bar{p}_{c,m} = A_p(p_{c,m})$, the class-wise support agreement is defined as

$$s_{n,c} = \max_{1 \le m \le M} \bar{q}_n^\top \bar{p}_{c,m}. \tag{1}$$

Here, $s_{n,c}$ measures how well query n matches the retained support appearances of class c. The maximum selects the candidate appearance that is most compatible with the query, yielding $s_{n,c} \in [-1, 1]$. Retaining multiple candidates allows the comparison to account for different appearance modes within the same class. Appendices A.2 and A.3 provide the details of candidate pooling and the residual adapters.

Support–confidence consistency. Support agreement alone indicates whether a query resembles the retained target appearances, but it does not reflect how confident the detector itself is in the corresponding class prediction. Detector confidence and support agreement therefore provide two complementary signals. For query n and class c, detector confidence indicates how strongly the adapted detector predicts the query as class c, whereas $s_{n,c}$ measures how well the same query

matches the cached target appearances of that class. The base detector confidence is defined as

$$\pi_{n,c} = \sigma(S_{n,c}),$$

where σ denotes the sigmoid function. Comparing $\pi_{n,c}$ with $s_{n,c}$ makes it possible to identify predictions that receive high detector confidence but are weakly supported by the retained target evidence. A mismatch score $h_{n,c}$ is therefore introduced to quantify this inconsistency, with larger values indicating stronger confidence–support mismatch. This mismatch signal is subsequently used to suppress predictions whose detector confidence is not sufficiently supported by the observed target appearances. Appendix A.4 provides the definition of $h_{n,c}$ and its associated parameters.

Enhancement and suppression. Enhancement and suppression convert the support-based signals into two complementary score adjustments. For query n and class c, the enhancement-branch adjustment is defined as

$$\Delta^{+}_{n,c} = \beta_{+} T s_{n,c},$$

where $s_{n,c}$ is the support agreement. Because $s_{n,c}$ can be either positive or negative, $\Delta^{+}_{n,c}$ is a signed score adjustment: positive agreement increases the class score, whereas negative agreement decreases it. The superscript $+$ denotes the enhancement branch rather than requiring $\Delta^{+}_{n,c}$ itself to be positive.

The suppression penalty is defined as

$$\Delta^{-}_{n,c} = \beta_{-} T h_{n,c},$$

where $h_{n,c}$ measures the confidence–support mismatch. A larger mismatch therefore produces a stronger penalty for predictions that receive high detector confidence but lack sufficient support from the retained target appearances. The learnable non-negative coefficients β_{+} and β_{-} control the strengths of the two branches, respectively, while T provides a shared scaling factor. Together, the two branches reinforce predictions that agree with the target evidence and impose additional suppression when detector confidence conflicts with that evidence.

Class residual. The two score adjustments are combined with a learnable class-specific bias b_c to form the class residual:

$$\Delta_{n,c} = \Delta^{+}_{n,c} - \Delta^{-}_{n,c} + b_c. \tag{2}$$

Here, $\Delta_{n,c}$ represents the overall support-conditioned correction for query n and class c before gating. The class bias b_c allows class-specific score offsets to be learned during Stage II.

Gated refinement. Although Δ provides a support-conditioned correction, applying the same correction strength to every query–class pair may over-adjust classes that are only weakly activated by the base detector. The language-score gate $G \in \mathbb{R}^{N \times C}$ therefore controls how strongly the residual correction is applied according to the base class confidence. In particular, the gate attenuates corrections for weakly activated classes while retaining stronger corrections for more relevant query–class pairs. The refined class logits are given by

$$S' = S + G \odot \Delta, \tag{3}$$

where $\odot$ denotes element-wise multiplication. The term $G \odot \Delta$ represents the gated residual correction added to the base class logits S to obtain the refined logits S'. Thus, SSCA preserves the detector's original prediction as the base score and modifies it only through a gated residual term rather than replacing it with a support-only score. Appendix A.4 provides the thresholds, scaling factors, and initialization settings. The refinement operates only on classification logits and leaves the predicted bounding boxes unchanged. Here, *calibration* specifically refers to this support-conditioned score-refinement process, whereas probabilistic calibration concerns the distinct relationship between prediction confidence and correctness (Guo et al., 2017).

3.4 Prediction and Evaluation

For each query image, HED produces query features Q, base class logits S, and predicted boxes B. SSCA then refines the base logits S using the fixed evidence bank E, while leaving the box predictions unchanged. During Stage II training, this score refinement is applied only to matching queries, whereas denoising queries retain the detector's original logits. At inference, both the evidence

bank and all model parameters remain fixed. A support-size box prior is subsequently applied to the SSCA-refined logits, followed by Soft-NMS (Bodla et al., 2017) on the decoded detections. Both post-processing operations further adjust classification scores without modifying the predicted box coordinates. Appendix A.5 provides their definitions and parameter settings.

3.5 TRAINING

Stage I: target-domain HED adaptation. HED is fine-tuned separately for each target domain and shot setting. The resulting Stage I checkpoint is used to construct the corresponding evidence bank and initialize the Stage II query detector.

Stage II: support-conditioned adaptation. Training continues for 50 additional epochs from the Stage I checkpoint while the evidence bank remains fixed. The detector and SSCA are optimized during this stage, whereas the evidence bank and language model remain frozen. Appendix A.1 provides the shared training schedule.

4 EXPERIMENTAL RESULTS

4.1 SETUP AND COMPARISON PROTOCOL

During the testing phase, performance is assessed under the 1-, 5-, and 10-shot settings on six target-domain datasets: ArTaxOr (Fu et al., 2024), Clipart1k (Fu et al., 2024), DIOR (Fu et al., 2024), DeepFish (Fu et al., 2024), NEU-DET (Fu et al., 2024), and UODD (Fu et al., 2024). These datasets cover diverse domains, including arthropod imagery, clip-art objects, remote-sensing scenes, underwater environments, and industrial surface defects. Performance is measured using COCO-style Average Precision (AP), averaged over intersection-over-union (IoU) thresholds from 0.50 to 0.95 (Lin et al., 2014). The cross-domain mean assigns equal weight to the six datasets.

The published benchmark comparison follows the original training and evaluation protocols of the compared methods. For PULSE, Stage II continues detector optimization for 50 epochs from the Stage I checkpoint while the evidence bank remains fixed. The SSCA-refined logits are further processed using the fixed support-derived box-size prior and Soft-NMS described in Section 3.4.

A separate frozen-detector analysis is conducted to examine the contribution of support-conditioned score refinement without further updating the detector. The adapted HED detector remains fixed, and successive SSCA variants are evaluated under the same detector outputs. This analysis isolates the score-refinement mechanism, whereas Table 1 reports the results of the full Stage II benchmark comparison.

4.2 RESULTS ACROSS DOMAINS AND SHOT SETTINGS

Table 1 compares PULSE with existing CD-FSOD methods under the 1-, 5-, and 10-shot settings. Overall, PULSE achieves the highest mean AP among the compared methods across all three settings, with AP values of 37.1, 46.0, and 48.5, respectively. Compared with the corresponding FT-FSOD results of 34.9, 45.0, and 47.9 (Yu et al., 2026), PULSE improves the mean AP by 2.2, 1.0, and 0.6 points under the 1-, 5-, and 10-shot settings, respectively.

Beyond the overall mean results, the strong performance of PULSE is also evident across individual target domains. Under the 1-shot setting, PULSE achieves the highest AP among the compared methods on five of the six datasets, including Clipart1k, DIOR, DeepFish, NEU-DET, and UODD. Under the 5-shot setting, the highest AP is achieved on ArTaxOr, Clipart1k, DIOR, DeepFish, and NEU-DET, while ETS obtains a slightly higher AP on UODD. Under the 10-shot setting, PULSE achieves the highest AP on ArTaxOr, Clipart1k, DIOR, and DeepFish, and matches FT-FSOD on NEU-DET and UODD. These results indicate that the strong performance of PULSE is not confined to a specific dataset, but extends across diverse target domains and shot settings.

Table 1: **CD-FSOD benchmark comparison.** COCO-style AP on six target domains under the 1-, 5-, and 10-shot settings. Published baseline results follow the comparison reported in FT-FSOD (Yu et al., 2026) and retain the original training and evaluation protocols of the corresponding methods. Bold and underlined values indicate the best and second-best results, respectively, in each column. † denotes the DE-ViT fine-tuning baseline reproduced by CD-ViTO (Fu et al., 2024).

Method	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Avg.
1-shot							
DE-ViT-FT† (Zhang et al., 2025)	10.5	13.0	14.7	19.3	0.6	2.4	10.1
CD-ViTO (Fu et al., 2024)	21.0	17.7	17.8	20.3	3.6	3.1	13.9
VFMDETR (Liu et al., 2025a)	26.1	20.1	20.6	24.2	9.1	9.0	18.2
CDFormer (Meng et al., 2025)	36.0	54.0	16.3	34.5	7.4	12.7	26.8
ETS (Pan et al., 2025)	28.1	<u>57.5</u>	12.7	40.7	11.7	21.2	28.7
Domain-RAG (Li et al., 2025)	**57.2**	56.1	18.0	38.0	12.1	20.2	33.6
FT-FSOD (Yu et al., 2026)	49.1	55.6	<u>24.6</u>	<u>42.7</u>	<u>15.5</u>	<u>22.1</u>	<u>34.9</u>
PULSE (ours)	<u>53.6</u>	**59.7**	**25.8**	**44.5**	**16.2**	**22.8**	**37.1**
5-shot							
DE-ViT-FT† (Zhang et al., 2025)	38.0	38.1	23.4	21.2	7.8	5.0	22.3
CD-ViTO (Fu et al., 2024)	47.9	41.1	26.9	22.3	11.4	6.8	26.1
VFMDETR (Liu et al., 2025a)	63.3	45.1	32.1	29.5	19.0	19.6	34.7
CDFormer (Meng et al., 2025)	65.0	58.9	28.1	31.7	15.0	23.8	37.1
ETS (Pan et al., 2025)	64.5	<u>60.1</u>	29.3	44.9	23.5	**28.6**	41.8
Domain-RAG (Li et al., 2025)	70.0	59.8	31.5	43.8	24.2	26.8	42.7
FT-FSOD (Yu et al., 2026)	<u>76.8</u>	59.4	<u>35.3</u>	<u>45.5</u>	<u>25.2</u>	27.5	<u>45.0</u>
PULSE (ours)	**77.9**	**61.4**	**36.5**	**45.7**	**26.2**	<u>28.3</u>	**46.0**
10-shot							
DE-ViT-FT† (Zhang et al., 2025)	49.2	40.8	25.6	21.3	8.8	5.4	25.2
CD-ViTO (Fu et al., 2024)	60.5	44.3	30.8	22.3	12.8	7.0	29.6
VFMDETR (Liu et al., 2025a)	71.3	49.9	37.8	34.1	23.7	22.1	39.8
CDFormer (Meng et al., 2025)	68.7	59.0	32.5	35.5	18.1	26.4	40.0
ETS (Pan et al., 2025)	71.2	<u>61.5</u>	37.5	44.1	26.1	29.8	45.0
Domain-RAG (Li et al., 2025)	73.4	61.1	39.0	41.3	<u>26.3</u>	<u>31.2</u>	45.4
FT-FSOD (Yu et al., 2026)	<u>79.2</u>	59.6	<u>41.5</u>	<u>46.3</u>	**28.4**	**32.3**	<u>47.9</u>
PULSE (ours)	**79.7**	**61.9**	**42.1**	**46.7**	**28.4**	**32.3**	**48.5**

4.3 Ablation Study

To further examine the contribution of the main SSCA components, an ablation study is conducted with the adapted HED detector fixed. Under this setting, the detector features, base logits, and predicted boxes remain unchanged, while only the score-refinement components are varied. This design isolates the effect of the SSCA components from changes in the detector itself.

Progressive component analysis. Table 2 reports the 1-shot results across six target domains as the main SSCA components are introduced sequentially. Support–query matching provides the support-based score adjustment, IRCA aggregates support evidence into class-specific candidates, and the suppression branch penalizes high-confidence predictions with weak support agreement. The language-score gate then controls the strength of the final residual correction.

Table 2: **Progressive SSCA analysis with frozen HED (1-shot).** Each row adds the named score component to the preceding variant. Entries are AP; Δ is the step-wise change in mean AP. This analysis is distinct from the full Stage-II benchmark.

Variant	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Mean	Δ
Base HED	50.1	52.9	21.1	38.4	14.7	21.3	33.08	–
Basic SSCA	52.2	56.4	23.8	41.8	15.3	22.0	35.25	+2.17
+ IRCA pooling	52.8	58.1	24.7	43.1	15.7	22.5	36.15	+0.90
+ suppression	53.2	59.1	25.4	43.9	15.9	22.7	36.70	+0.55
+ **gate (Full SSCA)**	53.6	59.7	25.8	44.5	16.2	22.8	37.10	+0.40

The mean AP increases from 33.08 for HED to 35.25 after introducing support–query matching, corresponding to an improvement of 2.17 AP. Adding IRCA further increases the mean AP to 36.15. The suppression branch and language-score gate then improve the mean AP to 36.70 and 37.10, respectively. Among these progressive changes, support–query matching provides the largest stepwise improvement, indicating that direct use of the retained target evidence is beneficial for prediction scoring. IRCA, suppression, and the language-score gate provide further improvements by refining how the support evidence is represented and incorporated into the final score correction. Because each component is added on top of the preceding configuration, the reported AP changes represent stepwise effects rather than independent contributions.

Table 3: **Score reliability analysis with frozen HED (1-shot).** Values are averaged equally across the six target domains. The variant without suppression retains support–query matching, IRCA, and the language-score gate. HCFP is reported for detections with confidence scores above 0.7 and IoU below 0.5.

Variant	Mean AP ↑	ECE (%) ↓	HCFP (%) ↓
Base HED	33.08	17.82	14.35
Basic SSCA	35.25	13.60	11.20
Full w/o Δ^-	36.55	11.15	8.60
Full PULSE	37.10	8.12	5.24

Adding the confidence-mismatch suppression branch increases the mean AP from 36.55 to 37.10. At the same time, ECE decreases from 11.15% to 8.12%, while HCFP decreases from 8.60% to 5.24%. These results indicate that the suppression branch not only improves detection performance, but also improves the reliability of the predicted classification scores. In particular, the reduction in HCFP suggests that the model produces fewer high-confidence predictions when the corresponding query features are weakly supported by the retained target evidence. This behavior is consistent with the intended role of the suppression term Δ^-, which is designed to penalize predictions with high detector confidence but insufficient support–query agreement.

5 Conclusion

This work introduces PULSE to keep limited target-support evidence explicitly accessible beyond target-domain fine-tuning. A fixed multi-view evidence bank preserves annotated target appearances, while SSCA aligns the retained support representations with current query features and integrates their support–query agreement with detector confidence through a gated residual correction. This refinement adjusts the detector's base class logits without replacing them or modifying the predicted bounding box coordinates. Across six CD-FSOD benchmarks, PULSE achieves mean AP values of 37.1, 46.0, and 48.5 under the 1-, 5-, and 10-shot settings, respectively, exceeding the corresponding published FT-FSOD results by 2.2, 1.0, and 0.6 AP. These results demonstrate that retaining limited target evidence for prediction-time score refinement provides a useful complement to conventional parameter adaptation in CD-FSOD.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, English-language revision, figure preparation, LaTeX formatting, and consistency checks across the manuscript and result tables. All experimental results were obtained by the authors and were not generated by these tools. The authors remain responsible for the final claims, references, implementation, and experimental results.

ETHICS STATEMENT

This work studies object detection on established benchmark datasets, with no new collection of personal data or human-subject experiments. The evaluation focuses on transfer across these image domains. Deployment under different data distributions or application requirements may require separate evaluation.

REPRODUCIBILITY STATEMENT

The main paper describes the evidence bank, score-refinement mechanism, training stages, and evaluation protocol. Section 3.3 and Figure 3 present the SSCA score-refinement process, while Appendix A provides the implementation details, including candidate pooling, residual adapters, gating, and evaluation processing.

REFERENCES

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS – improving object detection with one line of code. In *ICCV*, 2017.

Jinhao Du, Shan Zhang, Qiang Chen, Haifeng Le, Yanpeng Sun, Yao Ni, Jian Wang, Bin He, and Jingdong Wang. σ-Adaptive Decoupled Prototype for Few-Shot Object Detection. In *ICCV*, pp. 18950–18960, 2023.

Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-RPN and multi-relation detector. In *CVPR*, 2020.

Chen-Bin Feng, Youyang Sha, Longfei Liu, Yongjun Yu, Chi Man Vong, Xuanlong Yu, and Xi Shen. FSOD-VFM: Few-shot object detection with vision foundation models and graph diffusion. In *ICLR*, 2026.

Yuqian Fu, Yu Wang, Yixuan Pan, Xingyu Qiu, Lian Huai, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *ECCV*, pp. 247–264, 2024.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

Yali Huang, Jie Mei, Yiming Yang, Mi Guo, Mingyuan Jiu, and Mingliang Xu. Instance feature caching for cross-domain few-shot object detection. In *CVPR Workshops*, pp. 1567–1575, 2025.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Hojun Lee, Myunggi Lee, and Nojun Kwak. Few-shot object detection by attending to per-sample-prototype. In *WACV*, pp. 2445–2454, 2022.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, pp. 10965–10975, 2022.

Yu Li, Xingyu Qiu, Yuqian Fu, Jie Chen, Tianwen Qian, Xu Zheng, Danda Pani Paudel, Yanwei Fu, Xuanjing Huang, Luc Van Gool, and Yu-Gang Jiang. Domain-RAG: Retrieval-guided compositional image generation for cross-domain few-shot object detection, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Changhan Liu, Xunzhi Xiang, Zixuan Duan, Wenbin Li, Qi Fan, and Yang Gao. Don't need retraining: A mixture of DETR and vision foundation models for cross-domain few-shot object detection. In *NeurIPS*, 2025a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024.

Xin Liu, Jie Liu, Jie Tang, and Gangshan Wu. CATANet: Efficient content-aware token aggregation for lightweight image super-resolution. In *CVPR*, pp. 17902–17912, 2025b.

Boyuan Meng, Xiaohan Zhang, Peilin Li, Zhe Wu, Yiming Li, Wenkai Zhao, Beinan Yu, and Hui-Liang Shen. CDFormer: Cross-domain few-shot object detection transformer against feature confusion, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Jiancheng Pan, Yanxing Liu, Xiao He, Long Peng, Jiahao Li, Yuze Sun, and Xiaomeng Huang. Enhance then search: An augmentation-search strategy with foundation models for cross-domain few-shot object detection. In *CVPR Workshops*, pp. 1548–1556, 2025.

Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-CNN for few-shot object detection. In *ICCV*, 2021.

Zeyu Shangguan, Daniel Seita, and Mohammad Rostami. Cross-domain multi-modal few-shot object detection via rich text. In *WACV*, pp. 6570–6580, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *CVPR*, 2021.

Wanqi Wang, Jingcai Guo, Yuxiang Cai, and Zhi Chen. Learning multi-modal prototypes for cross-domain few-shot object detection, 2026.

Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020.

Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *ECCV*, 2020.

Man Wu, Lili Wen, and Gaocai Wang. DICC-ViT: Improving cross-domain few-shot object detection via dynamic information coupling and contrastive learning. *Journal of King Saud University Computer and Information Sciences*, 38:357, 2026. doi: 10.1007/s44443-026-00716-4.

Wuti Xiong. CD-FSOD: A benchmark for cross-domain few-shot object detection. In *ICASSP*, pp. 1–5, 2023.

Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *ICCV*, 2019.

Xi Yang and Quantao Xie. StyleProto: Style-augmented prototype learning for cross-domain few-shot object detection. In *AAAI*, volume 40, pp. 11748–11756, 2026. doi: 10.1609/aaai.v40i14.38160.

Yuantao Yin, Ping Yin, Xue Xiao, Liang Yan, Siqing Sun, and Xiaobo An. Balanced ID–OOD tradeoff transfer makes query based detectors good few shot learners. *High-Confidence Computing*, 5:100237, 2025.

Xuanlong Yu, Youyang Sha, Longfei Liu, Xi Shen, and Di Yang. A closer look at cross-domain few-shot object detection: Fine-tuning matters and parallel decoder helps. In *CVPR*, pp. 26593–26603, 2026.

Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias. Detect everything with few examples. In *CoRL*, 2025.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *CVPR*, pp. 16793–16803, 2022.

A SSCA IMPLEMENTATION DETAILS

This appendix provides the implementation details of the five SSCA operations illustrated in Figure 3, together with the fixed evaluation processing used in the experiments. The main configuration includes candidate pooling and a scalar query gate. The implementation terminology is retained where necessary to maintain a clear correspondence between the equations and the accompanying code.

A.1 TRAINING SCHEDULE

Stage I fine-tunes HED for 16 epochs on DeepFish, 50 epochs on Clipart1k, and 100 epochs on the remaining target domains. Stage II continues from the corresponding Stage I checkpoint for 50 additional epochs, including an initial five-epoch staged-freezing period. The base learning rate is 5×10^{-5}, with learning-rate multipliers of 20 for the support adapter and its submodules, 0.1 for the backbone, and 0 for the text encoder. The learning rate is reduced by a factor of 0.1 at epochs 30 and 42. During Stage II, the detector and SSCA parameters are optimized, while the evidence bank and text encoder remain fixed. Except for the Stage I training duration, the remaining optimization settings and module hyperparameters are shared across domains and specified in the accompanying implementation configuration.

A.2 CANDIDATE POOLING

Candidate pooling aggregates the fixed evidence bank into $M = 4$ class-specific candidate representations. This module is referred to as IRCA in the implementation and adopts content-aware token grouping (Liu et al., 2025b). Let $\hat{e}_{c,j}$ denote the normalized evidence feature of class c, and let $\hat{a}_m$ denote the m-th normalized shared center. The binary validity mask $v_{c,j}$ excludes empty evidence slots, while the non-negative reliability weight $r_{c,j}$ controls the contribution of each valid feature. The soft assignment of evidence feature j to candidate m is defined as

$$\alpha_{c,j,m} = \frac{\exp(\hat{e}_{c,j}^{\top} \hat{a}_m / \tau_{\mathrm{c}})}{\sum_{m'=1}^{M} \exp(\hat{e}_{c,j}^{\top} \hat{a}_{m'} / \tau_{\mathrm{c}})}, \tag{4}$$

$$u_{c,m} = \frac{\sum_j v_{c,j} r_{c,j} \alpha_{c,j,m} \hat{e}_{c,j}}{\max\left(\sum_j v_{c,j} r_{c,j} \alpha_{c,j,m}, \epsilon\right)}. \tag{5}$$

The temperature and numerical-stability constants are set to $\tau_{\mathrm{c}} = 0.15$ and $\epsilon = 10^{-6}$, respectively. The centers are shared across classes, whereas the pooling operation is performed independently for each class. If the assignment mass of a candidate falls below ϵ, the normalized mean of the valid evidence features from the corresponding class is used as a fallback.

Each pooled candidate is further refined through a residual transformation:

$$p_{c,m} = \mathrm{norm}(u_{c,m} + W_r \, \mathrm{LN}(u_{c,m})) \,, \tag{6}$$

where norm denotes ℓ_2 normalization and W_r is zero-initialized.

During training, the shared centers are updated according to the implementation configuration, including an exponential moving average toward their assigned evidence directions with momentum 0.9. The centers remain fixed during evaluation. When candidate pooling is disabled, support–query matching is performed directly over the raw evidence features using maximum similarity.

For N queries, C classes, M candidates, and feature dimension D, the support-matching operation has complexity $O(NCMD)$, excluding feature adaptation and detector computation.

A.3 RESIDUAL ADAPTERS AND SCALAR QUERY GATE

The query and support streams use residual adapters with the same architecture but independent parameters:

$$A_x(h) = \mathrm{norm}(h + \rho U_x \, \mathrm{GELU}(V_x \, \mathrm{LN}(h))) \,, \qquad x \in \{q, p\}. \tag{7}$$

The feature dimension is 256, the bottleneck width is 128, and the residual scale is $\rho = 0.3$. The up-projection matrices U_x are zero-initialized, so the adapters initially preserve the input feature directions and subsequently learn stream-specific residual transformations for the shared comparison space.

The implementation configuration additionally contains a scalar query gate, denoted as LGAG in the code. Using the mean support context

$$g = \frac{1}{CM} \sum_{c,m} \bar{p}_{c,m},$$

the gate computes

$$a_n = 1 + \tanh(\psi(\mathrm{GELU}(W_g g + W_q \bar{q}_n))), \qquad \tilde{q}_n = \mathrm{norm}(a_n \bar{q}_n). \tag{8}$$

The hidden width is 64, and the scalar projection ψ is initialized with zero weights and bias.

Since a_n is a positive scalar, normalization removes its magnitude for nonzero query features:

$$\mathrm{norm}(a_n \bar{q}_n) = \bar{q}_n.$$

Consequently, this configured scalar gate does not alter the cosine direction used for support–query matching under the current normalized formulation. Feature alignment in Equation 1 is therefore provided by the residual adapters.

This scalar query gate is distinct from the language-score gate G used in SSCA. The latter modulates the class-level residual after support–query matching and directly affects the final logit correction.

A.4 SSCA Parameters

The confidence-mismatch score used by the suppression branch is defined as

$$h_{n,c} = \sigma\left(\frac{\pi_{n,c} - \tau_p}{\kappa_p}\right) \sigma\left(\frac{m_0 - s_{n,c}}{\kappa_m}\right), \tag{9}$$

where $\pi_{n,c}$ denotes detector confidence and $s_{n,c}$ denotes support–query agreement. The first factor increases as detector confidence exceeds τ_p, whereas the second increases as support–query agreement falls below m_0. Consequently, $h_{n,c}$ becomes large for predictions with high detector confidence but weak support evidence.

The enhancement and suppression strengths, β_+ and β_-, are constrained to be positive through a softplus parameterization and are initialized to 0.10. The scale factor is set to $T = 8$. The mismatch parameters are $\tau_p = 0.08$, $\kappa_p = 0.03$, $m_0 = 0.08$, and $\kappa_m = 0.08$. The language-score gate uses a threshold of 0.03, a temperature of 0.02, and a minimum gate value of 0.10. The class bias is initialized to zero.

Both the mismatch computation and the language-score gate use detached detector confidence, preventing these branches from propagating gradients through the confidence values used to construct the score correction. Background-margin scoring and similarity centering are disabled in the reported configuration.

For token-level detector outputs, the detector logits are first mapped to class-level confidence values. After SSCA computes the class-level correction, the correction for each class is mapped back to its associated text tokens before residual addition.

A.5 Fixed Evaluation Processing

A support-derived box-size prior is applied after SSCA score refinement. For a predicted box with image-normalized width w_n and height h_n, let $z_n = (\log w_n, \log h_n)$. The support annotations of class c determine the class mean μ_c and component-wise scale σ_c in this log-size space. The prior score is defined as

$$g_{n,c} = \exp\left[-\frac{1}{2} \left\|(z_n - \mu_c)/\sigma_c\right\|_2^2\right], \qquad S''_{n,c} = S'_{n,c} + \lambda_{\mathrm{box}}(g_{n,c} - 0.5), \tag{10}$$

where $\lambda_{\text{box}} = 0.7$. A minimum effective scale of 0.35 is used for each size dimension to avoid a degenerate prior when only a few support annotations are available, particularly in the 1-shot setting.

Soft-NMS (Bodla et al., 2017) is subsequently applied to the decoded detections with an IoU threshold of 0.4. Both operations adjust classification scores only and leave the predicted box coordinates unchanged.