Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ordered Support-Conditioned Query Exchange for Cross-Domain Few-Shot Remote Sensing Detection

Feng Jiang

jiangf@njust.edu.cn

Taizhou Institute of Science and Technology, NJUST

Zhipeng Ye

Taizhou Institute of Science and Technology, NJUST

Lili Zhou

Taizhou Institute of Science and Technology, NJUST

Research Article

Keywords: Remote sensing, cross-domain few-shot detection, query allocation, support evidence, decoder perturbation

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11608186/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ordered Support-Conditioned Query Exchange for Cross-Domain Few-Shot Remote Sensing Detection

Feng Jiang[1*], Zhipeng Ye[1] and Lili Zhou[2]

[1]School of Computer Science and Engineering, Taizhou Institute of Science and Technology, NJUST, Taizhou, 225300, China.

[2]School of Electronic and Electrical Engineering, Taizhou Institute of Science and Technology, NJUST, Taizhou, 225300, China.

*Corresponding author(s). E-mail(s): jiangf@njust.edu.cn;

Abstract

Generic proposal selection may overlook the target classes specified by a few-shot support set. We propose support-conditioned query exchange (SCQE) for cross-domain few-shot remote sensing detection using a frozen prompt-free universal region proposal network (PF-RPN). SCQE selects incoming encoder references and their receiving decoder slots, with proposal weights, learned query contents and query count fixed. Each replacement is evaluated by its complete support-conditioned coverage gain, including outgoing coverage loss, under a cumulative perturbation budget on protected native outputs. Vector–Jacobian product sketches estimate interactions among exchanges. The unchanged decoder verifies the complete plan, accepting changed proposals or restoring all cached native PF-RPN outputs before shared recognition. A support-adapted CD-ViTO backend provides class scores while retaining PF-RPN box geometry. SCQE achieves average precision (AP@[.50:.95]) of 24.86, 36.36 and 42.36 on DIOR and 21.47, 27.81 and 34.02 on DOTA with horizontal bounding boxes (DOTA-HBB) for 1-, 5- and 10-shot detection. Gains over native selection range from 7.11 to 8.68 and from 4.93 to 6.27 AP points, respectively. DIOR 5-shot ablations support receiving-slot assignment, complete exchange gain and cumulative perturbation control. Disabling the actual-risk gate reduces AP at an intersection-over-union threshold of 0.75 (AP75) by 2.88 points and yields deployed budget violations in 4.7% of query–support pairs. In the 5-shot setting, SCQE increases mean query-inference latency over the directly verified prefix-selection baseline by 12.4% on DIOR and 15.0% on DOTA-HBB, with the same reported peak allocated GPU memory.

Keywords: Remote sensing, cross-domain few-shot detection, query allocation, support evidence, decoder perturbation

1 Introduction

Remote-sensing detection faces changes in sensors, regions, scale, orientation and texture [1–3]. We study transfer from a source-pretrained system using K annotated target instances per class, with image-disjoint support and query sets. This annotation-budget setting permits target-category overlap with pretraining classes; checkpoint exposure is recorded.

Few-shot detection uses feature reweighting [4], RoI meta-learning [5], fine-tuning [6] and contrastive proposal encoding [7]. Attention-RPN

conditions proposals on support [8], and Meta-DETR aggregates support classes at image level [9]. CD-ViTO adapts support representations [10]; SI-ViTO adds sparse-instance learning [11], and FT-FSOD uses fine-tuning and parallel decoders [12]. We retain support adaptation but target reference allocation in a frozen proposal decoder.

DETR uses finite query collections [13]; Deformable DETR [14], DAB-DETR and DINO [15, 16] establish reference-based decoding, and Grounding DINO adds grounded pretraining [17]. PF-RPN self-prompts and selects centerness-guided references; its original evaluation already combines PF-RPN with CD-ViTO [18]. Unlike query pruning/fine-tuning in GPQ [19] or trajectory-based reliability correction in CQTR [20], SCQE fixes query count, learned contents and proposal weights. DINOv2 [21] and GLIP [22] further motivate controlling feature and recognition states across selectors.

Two issues remain at this interface. Selecting a reference set does not specify its assignment to decoder slots: fixed learned contents can yield different outputs under different assignments. A replacement must also account for outgoing coverage loss and the combined perturbation of all exchanges. Weighted coverage [23], query-conditioned coverage [24] and output sensitivity [25] provide established tools; our contribution is their use in ordered candidate–slot decisions with complete gain and cumulative perturbation:

1. **Ordered reference-box exchange.** We separate selected-set coverage from receiving-slot effects in fixed-cardinality allocation, with each decision specifying an incoming reference and its eligible slot.
2. **Joint gain–risk decisions.** We evaluate complete removal–addition gain under a cumulative VJP budget, including estimated cross-slot interactions, and verify the whole plan before deployment.

Matched experiments on DIOR and DOTA-HBB isolate reference selection using shared adapted recognition. The 5-shot gains over native selection are 8.68 and 4.93 AP points; ablations examine assignment, complete gain, cumulative risk and verification.

2 Method

Figure 1 summarizes SCQE. Native decoding caches PF-RPN outputs, fixes protected slots and supplies the VJP reference point. Support evidence defines coverage; bounded search selects candidate–slot exchanges under cumulative perturbation control. The unchanged decoder verifies the plan, accepting changed outputs or restoring all cached native outputs. Both paths use shared support-adapted CD-ViTO class scoring and retain PF-RPN box geometry. B0 denotes native allocation with this backend.

2.1 Ordered exchange and joint objective

An episode supplies K annotated instances per target class in support set $\mathcal{S}$; query images are disjoint. PF-RPN weights, encoder memory, internal pseudo-text and learned slot contents q_r remain fixed during exchange. Let A_0 be its native allocation of Q distinct encoder references and $U_0 = \mathrm{set}(A_0)$. Candidate j has normalized (c_x, c_y, w, h) box $\xi_j = \mathrm{sigmoid}(a_j) \in [0,1]^4$, where a_j is its encoder box logit. An ordered allocation A assigns candidate $A[r]$ to slot r, giving

$$R(A)_{r,:} = \xi_{A[r]}^{\top}, \qquad R_0 = R(A_0). \tag{1}$$

With valid candidate pool $\mathcal{V}_s$, protected native slots H and exchange count $m(A) = |\{r : A[r] \neq A_0[r]\}|$, the search objective is

$$\begin{aligned} &\max_{A} F_{\mathcal{S}}(A) \\ &\text{s.t. } |A| = Q, \quad A[r] \text{ pairwise distinct}, \\ &\quad A[r] \in \{A_0[r]\} \cup (\mathcal{V}_s \setminus U_0), \\ &\quad A[r] = A_0[r] \ (r \in H), \\ &\quad m(A) \le M_{\max}, \quad \widehat{\mathcal{R}}(A) \le \epsilon_{\text{proxy}}. \end{aligned} \tag{2}$$

Here $F_{\mathcal{S}}$ measures support-conditioned set coverage and $\widehat{\mathcal{R}}$ estimates cumulative changes in protected native outputs; both are defined below. A slot keeps its own native reference or receives a previously unselected candidate: native indices cannot migrate between slots. For replacement $A^{r\leftarrow j}$, the complete exchange gain is

$$\Delta F_{rj}(A) = F_{\mathcal{S}}(A^{r\leftarrow j}) - F_{\mathcal{S}}(A). \tag{3}$$

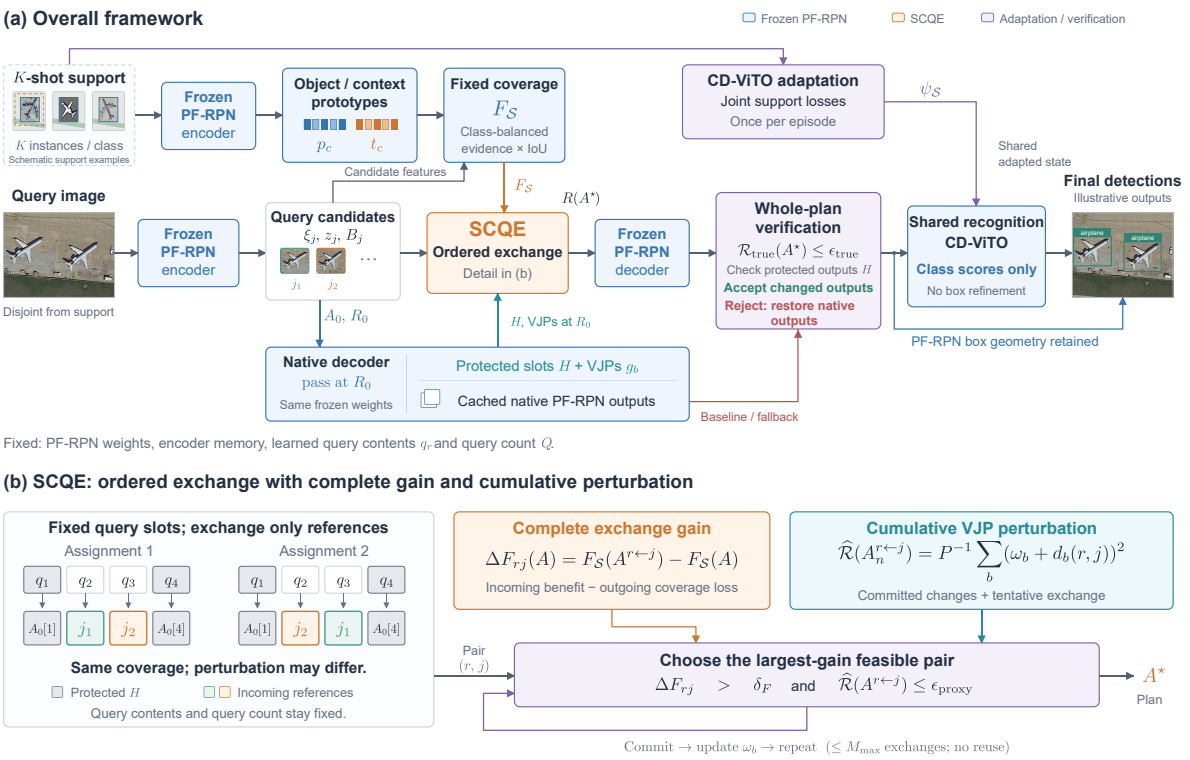


Fig. 1 SCQE framework. Native decoding caches all PF-RPN outputs, fixes protected slots H and supplies the VJP linearization point. Ordered exchanges use complete gain under cumulative perturbation control. Whole-plan verification accepts changed outputs or restores all cached native outputs before shared CD-ViTO class scoring. The cached path contains PF-RPN proposals before recognition; final boxes retain PF-RPN geometry.

This net gain includes outgoing coverage loss and incoming benefit evaluated after removal. Addition-only scoring omits that cost. The objective is a search surrogate, without an AP or global-optimality guarantee.

2.2 Support-conditioned coverage

2.2.1 Support evidence and coverage

Let $u_{c,k}, v_{c,k}, z_j \in \mathbb{R}^{256}$ be normalized support-object, support-context and candidate descriptors. All use the same frozen PF-RPN encoder and the common sampling rule in Section 3.2. Object descriptors use supplied support boxes; context boxes expand their width and height by 1.5 about the same center. Context sampling excludes points inside any exposed annotation in that support image, but can still include unannotated objects. For nonempty valid-instance sets $\mathcal{K}_c^p, \mathcal{K}_c^t \subseteq \{1, \dots, K\}$ and normalization $\nu(v) = v/\max(\|v\|_2, 10^{-6})$, define

$$
\begin{aligned}
p_c &= \nu\left(|\mathcal{K}_c^p|^{-1} \sum_{k \in \mathcal{K}_c^p} u_{c,k}\right), \\
t_c &= \nu\left(|\mathcal{K}_c^t|^{-1} \sum_{k \in \mathcal{K}_c^t} v_{c,k}\right), \\
\alpha_{jc} &= \mathrm{clip}\left(\frac{\langle z_j, p_c\rangle - \langle z_j, t_c\rangle}{2}, 0, 1\right).
\end{aligned}
\tag{4}
$$

Evidence α_{jc} is a relative score, not a probability. Usable classes $\mathcal{C}^+$ have valid prototypes and

positive $Z_c = \sum_{i \in \mathcal{V}_s} \alpha_{ic}$. For atom (i, c):

$$
\begin{aligned}
w_{ic} &= \frac{\alpha_{ic}}{|\mathcal{C}^+| Z_c}, \\
\kappa_{jic} &= \begin{cases} \alpha_{jc}\,\mathrm{IoU}(B_j, B_i), & j \in \mathcal{V}_s, \\ 0, & \text{otherwise}, \end{cases} \\
F_{\mathcal{S}}(U) &= \sum_{c \in \mathcal{C}^+} \sum_{i \in \mathcal{V}_s} w_{ic} \max_{j \in U} \kappa_{jic}.
\end{aligned}
\tag{5}
$$

Here B_j is the clipped sampling copy of ξ_j and $F_{\mathcal{S}}(A) = F_{\mathcal{S}}(\mathrm{set}(A))$. Each usable class has equal total demand weight; native entries outside $\mathcal{V}_s$ contribute zero utility. Empty evidence returns B0. Evidence, atoms and weights stay fixed, giving a monotone submodular utility [23]. Candidate density can increase demand mass, so coverage is not instance-balanced recall.

Candidates must have native-valid finite references and valid clipped sampling geometry/descriptors. Invalid native references trigger bypass; invalid sampling descriptors contribute zero coverage but leave unprotected slots exchangeable. Descriptor validity and bypass rules are specified in Online Resource 1.

2.3 Cumulative decoder perturbation

2.3.1 Protected outputs

Let $\hat{b}_r(R)$, $\pi_r(R)$ and $\gamma_r(R)$ denote the normalized decoded box, internal-token scores and native transformed confidence. The $T \geq 1$ valid internal pseudo-text positions are fixed by the native token mask, in native order, for every slot and trial allocation; masked and padded positions are excluded. These tokens are distinct from target classes. For final-decoder token logits ℓ_{rt}^{tok} and confidence logit ℓ_r^{conf}, the native PF-RPN transformations are

$$
\begin{aligned}
\pi_{rt} &= \mathrm{sigmoid}(\ell_{rt}^{\mathrm{tok}}), \\
\gamma_r &= \mathrm{sigmoid}(\mathrm{sigmoid}(\ell_r^{\mathrm{conf}})), \\
s_r^{\mathrm{PF}} &= \max_t \pi_{rt} + \gamma_r.
\end{aligned}
\tag{6}
$$

The two sigmoid operations in the confidence conversion reproduce the released PF-RPN scoring path. The protection fraction $\eta = 0.5$ fixes H as the top $h = \lceil \eta Q \rceil$ native-score slots before exchange, with ascending slot-index ties. Their references remain unchanged, but interactions with modified slots can alter their outputs. The protected outputs are concatenated in fixed slot order:

$$
\begin{aligned}
y_H(R) &= \mathrm{col}_{r \in H}(\hat{b}_r(R), \pi_r(R), \gamma_r(R)), \\
\mathcal{R}_{\mathrm{true}}(A) &= \left\| W^{1/2}[y_H(R(A)) - y_H(R_0)] \right\|_2^2.
\end{aligned}
\tag{7}
$$

Diagonal W assigns $1/(12h)$ per box coordinate, $1/(3Th)$ per token activation and $1/(3h)$ per confidence, giving each group total coefficient weight $1/3$. This pre-NMS aggregate risk does not bound individual-object IoU loss or detection accuracy.

2.3.2 Cumulative VJP estimate

For $\Delta R = R(A) - R_0$, cache P vector–Jacobian products using independent Rademacher probes v_b of dimension $h(5 + T)$:

$$
\begin{aligned}
g_b &= \nabla_R [v_b^\top W^{1/2} y_H(R)] \Big|_{R_0}, \\
\widehat{\mathcal{R}}(A) &= P^{-1} \sum_{b=1}^{P} \langle g_b, \Delta R \rangle_F^2.
\end{aligned}
\tag{8}
$$

The functional decoder removes iterative-reference stop-gradients while preserving native forward values. Each $g_b \in \mathbb{R}^{Q \times 4}$ projects the complete change, retaining estimated interactions among exchanges. For Jacobian block $J_r = \partial y_H / \partial R_{r,:}$ at R_0 and reference change $\delta_r = \xi_{A[r]} - \xi_{A_0[r]}$, the sketch estimates $\left\| W^{1/2} \sum_r J_r \delta_r \right\|_2^2$, including expected cross terms $2\delta_r^\top J_r^\top W J_s \delta_s$. Unbiasedness holds for fixed probe-independent linearized changes, not adaptively selected finite plans. Summing isolated exchange risks discards these terms.

Swapping references j_1, j_2 between slots r, s preserves $F_{\mathcal{S}}$ but changes linearized outputs by $(J_r - J_s)(\xi_{j_1} - \xi_{j_2})$. Thus an exchange must specify its receiving slot; Section 3.3 tests assignment at fixed membership.

2.4 Bounded joint search

At allocation A_n, write $U_n = \mathrm{set}(A_n)$ and let $X_n = \{r : A_n[r] \neq A_0[r]\}$ be the modified-slot set (written X in Algorithm 1). Shortlist the $L_{\mathrm{out}} = 32$ slots outside $H \cup X_n$ with lowest

Algorithm 1 Bounded ordered exchange

Require: $A_0, R_0, \{\xi_j\}, F_{\mathcal{S}}, \mathcal{V}_s, H, \{g_b\}$
Parameters: $M_{\max}, \delta_F, \epsilon_{\text{proxy}}$
Guard: any non-finite search computation returns A_0
$A \leftarrow A_0$, $X \leftarrow \varnothing$, $\omega_b \leftarrow 0$
for $n = 1, \ldots, M_{\max}$ **do**
  Shortlist outgoing slots $r \notin H \cup X$
  Shortlist incoming $j \in \mathcal{V}_s \setminus (U_0 \cup \text{set}(A))$
  Keep pairs with $\Delta F_{rj} > \delta_F$ and $\widehat{\mathcal{R}}(A^{r \leftarrow j}) \le \epsilon_{\text{proxy}}$
  if no pair remains **then**
    break
  end if
  Choose largest gain; ties: lower risk, slot, index
  Commit $A[r] \leftarrow j$, $X \leftarrow X \cup \{r\}$, $\omega_b \leftarrow \omega_b + d_b(r, j)$
  Refresh affected coverage caches
end for
return A

removal loss $\ell_r = F_{\mathcal{S}}(U_n) - F_{\mathcal{S}}(U_n \setminus \{A_n[r]\})$. Construct up to $L_{\text{in}} = 128$ incoming candidates from $\mathcal{V}_s \setminus (U_0 \cup U_n)$ by cycling usable classes in ascending order through their positive-evidence lists (decreasing evidence, then index), skipping duplicates. Rebuild both lists after each commitment; no slot or incoming candidate is reused.

For each probe, maintain $\omega_b = \langle g_b, R(A_n) - R_0 \rangle_F$. Since an eligible slot still holds its native reference, a tentative exchange has

$$d_b(r, j) = g_b[r, :](\xi_j - R^{\top}_{0,r,:}),$$
$$\widehat{\mathcal{R}}(A_n^{r \leftarrow j}) = P^{-1} \sum_b (\omega_b + d_b(r, j))^2. \quad (9)$$

Thus each pair is judged with all committed changes, rather than under a reset risk budget. Algorithm 1 admits positive complete gain above numerical margin δ_F and cumulative risk below ϵ_{proxy}. Its feasible-prefix restriction can miss final plans whose interactions cancel; no global solution is claimed.

Each step considers at most 4096 pairs. Direct gain evaluation costs $O(M_{\max} L_{\text{out}} L_{\text{in}} N |\mathcal{C}^+|)$ for $N = |\mathcal{V}_s|$, excluding descriptors, VJPs and cache maintenance (Online Resource 1).

2.5 Verification and shared recognition

The unchanged decoder runs once on the complete nonempty plan $A^\star$. A separate actual-risk tolerance on the protected outputs determines deployment:

$$A_{\text{out}} = \begin{cases} A^\star, & \mathcal{R}_{\text{true}}(A^\star) \le \epsilon_{\text{true}}, \\ A_0, & \text{otherwise.} \end{cases} \quad (10)$$

On rejection or any bypass (empty plan, empty evidence or non-finite computation), all cached native PF-RPN outputs are restored without partial repair or further trials. Both accepted and restored proposals pass through the same recognition backend. The check addresses sketch/nonlinear error; derivative validity is checked separately.

Support adaptation jointly updates the native CD-ViTO heads, learnable instance features, instance reweighting and domain prompter on exposed support annotations; feature backbones and proposal generators stay frozen. One adapted state is reused across matched selectors. Query images and SCQE plans do not enter adaptation. At query inference, the backend supplies class scores without box refinement. It receives up to 900 PF-RPN proposals per image/chip. DOTA predictions are mapped to parents before shared class thresholding, class-wise NMS and a total cap of 100; there is no premature chip final cap. PF scores do not multiply target-class scores, and query GT never enters selection. Online Resource 1 gives the complete adaptation and postprocessing contract.

3 Experiments

3.1 Setup and external comparisons

Protocol K restricts target conditioning/adaptation to K exposed instances per class, with settings determined on non-target development data. Table 1 compares external detection systems and matched PF-RPN selectors on DIOR and DOTA-HBB at $K = 1, 5, 10$. Protocol definitions are given below; completed implementation and protocol checks are described in Section 3.2.

3.1.1 Data and metrics

The 20-class DIOR CDFSOD/CD-ViTO conversion [1, 10] contains 18,463 support-pool images and a 5,000-image test partition. Of these 5,000 test images, 100 are held out from formal evaluation and are not used for threshold tuning, model selection or architecture selection. All such choices are determined on non-target development data. All reported DIOR detection-accuracy results are computed on the remaining 4,900 query images. DOTA-v1.0 [2] assigns train-support/validation-query parents before 1024 × 1024 tiling with 200-pixel overlap. Oriented annotations become clipped enclosing horizontal bounding boxes (HBB); difficulty 0/1 instances are ordinary evaluation GT. Support uses fully containing difficulty-0 instances (smallest row-major chip ties). Predictions are merged to parents before thresholding/NMS/cap 100. This differs from official oriented-box DOTA evaluation.

The protocol uses five nested $K = 1, 5, 10$ supports, exactly K exposed IDs per class and image-disjoint roles (DOTA parent-disjoint). Manifests are fixed before inference; infeasible episodes cannot reduce K or borrow query labels. One non-target-calibrated configuration applies across datasets/shots. Primary evaluation is DIOR 5-shot global AP@[.50:.95], with mean±sample SD over five paired supports; AP75 supplements the ablations. Empty-GT images are retained and DOTA areas use parent coordinates. DOTA results use a parent-level detection cap of 100. Chips are not independent replicates.

3.1.2 Comparisons

External systems include Attention-RPN with its multi-relation detector [8], Meta-DETR [9], CD-ViTO, FT-FSOD and SI-ViTO [10–12]. CD-ViTO/SI-ViTO use ViT-L/14; FT-FSOD uses MMGroundingDINO-B/Swin-B. Native architectures, localization heads and source-training procedures are retained. Attention-RPN uses support-conditioned inference and Meta-DETR retains support fine-tuning. Protocol matching requires shared support IDs, queries, categories and evaluation units, plus disclosure of checkpoint/pretraining resources and implementation deviations. All target conditioning is restricted to exposed support annotations; literature values do not fill missing evaluations.

B0–B3/SCQE share proposal weights, learned contents, evidence, candidates, adapted recognition state and postprocessing. B0 is native selection. B1 ranks the fixed initial incoming shortlist by $\max_c \alpha_{jc}$ (index ties), pairing candidates with the initial low-removal-loss slots up to its calibrated budget, without gain/risk tests or gating. B2 greedily maximizes positive complete gain with rebuilt lists/one-use rules, but no risk/gating (slot/index gain ties); B1/B2 budgets are independently calibrated in $\{8, 16, 32\}$. B3 generates a separate B2 trajectory up to 32 exchanges and directly checks every distinct available prefix in $\{1, 4, 8, 16, 32\}$. It deploys the highest-coverage actual-feasible prefix (shorter ties), or B0; non-monotone risk requires all checks. No query GT/AP chooses a prefix. SCQE uses $Q = 900$, $P = 16$, $M_{\max} = 32$ and $\eta = 0.5$.

3.1.3 Results on DIOR and DOTA-HBB

SCQE obtains the highest mean AP among the systems and selectors in Table 1 in all six dataset–shot settings. On DIOR, it achieves 24.86, 36.36 and 42.36 AP for 1-, 5- and 10-shot detection, improving native selection B0 by 7.11, 8.68 and 7.70 points. On DOTA-HBB, the corresponding scores are 21.47, 27.81 and 34.02 AP, with gains of 5.90, 4.93 and 6.27 points. Because B0 and SCQE share the frozen proposal network, adapted recognition state and postprocessing, these comparisons support the benefit of changing reference allocation within the common detection system.

Mean AP increases from B0 through B3 to SCQE in all settings. At 5-shot, SCQE exceeds B2 by 2.93 AP on DIOR and 2.55 on DOTA-HBB. Gains over B3 are 1.72, 1.89 and 1.31 AP on DIOR and 0.95, 1.42 and 1.36 on DOTA-HBB, supporting perturbation-aware search over post-hoc prefix verification. Section 3.3 isolates decision terms; Section 3.4 reports their computational cost.

FT-FSOD is the strongest external comparator across the six settings. SCQE exceeds it by 1.46, 2.32 and 2.65 AP points on DIOR, and by 2.88, 1.34 and 0.43 points on DOTA-HBB. These system-level comparisons involve different architectures and pretraining resources, whereas B0–B3 provide the controlled comparisons for reference selection. The reported means and SDs alone do not establish paired statistical significance.

Table 1 External systems and matched PF-RPN selectors on DIOR and DOTA-HBB. Values are AP@[.50:.95] on a 0–100 scale, reported as mean±sample SD over five support sets. DOTA-HBB uses enclosing horizontal boxes and parent-level evaluation after chip merging. Protocol and variability records were verified as described in Section 3.2.

	DIOR			DOTA-HBB		
System/selector	1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
Support-conditioned FSOD: fresh cross-domain evaluations						
Attention-RPN [8]	6.23 ± 0.51	14.02 ± 0.80	17.59 ± 0.62	3.38 ± 0.43	8.68 ± 0.72	12.91 ± 0.56
Meta-DETR [9]	11.75 ± 0.82	18.76 ± 2.77	19.38 ± 1.23	6.86 ± 0.52	11.32 ± 0.81	15.19 ± 0.72
CD-FSOD systems: fresh reproductions						
CD-ViTO	17.31 ± 2.34	26.17 ± 2.41	30.63 ± 0.75	14.78 ± 1.69	22.95 ± 1.74	26.16 ± 0.73
FT-FSOD	23.40 ± 2.11	34.04 ± 1.81	39.71 ± 0.66	18.59 ± 2.07	26.47 ± 1.79	33.59 ± 0.55
SI-ViTO	19.36 ± 2.42	27.31 ± 2.03	29.06 ± 0.73	13.69 ± 2.12	20.54 ± 1.47	25.86 ± 0.63
Matched PF-RPN selectors with shared recognition						
B0	17.75 ± 2.23	27.68 ± 1.89	34.66 ± 0.59	15.57 ± 1.72	22.88 ± 1.76	27.75 ± 0.81
B1	19.53 ± 2.01	29.71 ± 1.85	37.18 ± 0.91	17.10 ± 1.99	23.59 ± 1.77	29.05 ± 0.63
B2	22.67 ± 1.96	33.43 ± 2.11	40.02 ± 0.85	19.95 ± 2.12	25.26 ± 2.05	31.27 ± 0.54
B3	23.14 ± 2.25	34.47 ± 2.11	41.05 ± 0.82	20.52 ± 1.86	26.39 ± 1.28	32.66 ± 0.52
SCQE	24.86 ± 2.12	36.36 ± 1.94	42.36 ± 0.79	21.47 ± 2.21	27.81 ± 1.52	34.02 ± 0.50

3.2 Implementation and calibration

SCQE uses $Q = 900$, $P = 16$, $M_{\max} = 32$ and $\eta = 0.5$. Descriptors average normalized 7×7 bilinear samples from four frozen 256-channel encoder levels; coverage caches maintain the two largest values and their owners per atom. Proposal/functional/VJP inference uses FP32, with FP64 gain/risk accumulation. Native-equivalence, directional-derivative, cache-consistency, allocation and full-fallback checks passed.

Settings are calibrated only on fixed non-target data (40 queries and two 5-shot supports), then fixed across target datasets and shots. Recognition adaptation/postprocessing precede selector calibration; mechanism ablations reuse full-SCQE tolerances. Implementation/protocol records and per-support means/SDs were verified. A research-code snapshot is available at https://github.com/sjtukng/SCQE; its README describes external dependencies and snapshot limitations. Online Resource 1 details sampling, numerical validation, calibration and shared recognition.

3.3 Ablation studies

Table 2 reports DIOR 5-shot ablations using the full query set and five-support protocol in Section 3.1. Controls share adapted backends, native outputs, evidence, probes, search limits, protection sets, numerical margins, postprocessing and calibrated tolerances unless explicitly changed. No target-data retuning is allowed. AP/AP75 follow the mean±sample SD convention; ΔAP is the difference of reported means relative to full SCQE.

3.3.1 Ordered exchange at fixed membership

The canonical control replays each pre-gate SCQE plan with identical incoming membership and receiving slots, assigning references in decreasing $\max_c \alpha_{jc}$ order (index ties) to ascending slots. Coverage and pre-gate exchange count are unchanged; zero/one-exchange plans are retained. The same actual gate/fallback is applied without rerunning search.

Canonical assignment loses 1.22 AP and 1.42 AP75 despite lower rejection (6.4% versus 7.1%). The fixed-membership comparison supports assignment sensitivity through the combined decoder-and-verification pipeline, without an optimality guarantee.

3.3.2 Complete gain and cumulative perturbation

The 2×2 design crosses complete/addition-only gain with cumulative/isolated perturbation.

Addition-only uses $F_{\mathcal{S}}(U_n \cup \{j\}) - F_{\mathcal{S}}(U_n)$ for admission/ranking, retaining the removal-loss shortlist without an extra positive-net-gain test. Isolated risk replaces Eq. (9) by $P^{-1} \sum_b \sum_{r \in X} d_b(r, A[r])^2$ for the tentative modified-slot set X, accumulating committed isolated contributions without cross terms or budget resets. Each cell uses its own search trajectory, the same rebuilding/one-use rules, numeric tolerances and actual gate.

Addition-only and isolated-risk controls lose 1.47 and 1.28 AP points; their combination loses 1.63. Complete gain improves AP by 1.47 points with cumulative risk but only 0.35 with isolated risk; cumulative risk improves AP by 1.28 points with complete gain but only 0.16 with addition-only gain. The positive four-cell AP contrast (1.12) describes dependence between the terms at the mean level. AP75 losses are approximately additive: 1.97 and 1.24 individually versus 3.22 jointly. Both terms therefore benefit stricter-IoU detection, without isolating localization from classification effects. The isolated-risk control has lower rejection than full SCQE (6.8% versus 7.1%) despite lower AP; rejection frequency alone does not measure selection quality.

3.3.3 Whole-plan verification

The no-gate control replays the identical SCQE plan and changed forward, removing only actual-threshold rejection; numerical/bypass guards remain. Rej. is pooled over finite nonempty proposed plans, whereas Over is the deployed over-tolerance fraction over all query–support pairs, measured before recognition/NMS after fallback. Bypasses and unavailable risk measurements are recorded separately. A dash in Rej. denotes a disabled gate.

Disabling the actual-risk gate loses 1.31 AP and 2.88 AP75 points and deploys over-tolerance outputs in 4.7% of query–support pairs. Thus proxy-feasible plans can violate the actual budget, and the actual-risk gate also improves stricter-IoU detection in this evaluation. All gated variants have zero deployed violations because infeasible plans restore native outputs; this enforcement property does not guarantee accuracy preservation on every query.

3.4 Efficiency evaluation

Experiments use an Intel Core i7-13700K Linux workstation and two RTX 4090 GPUs (24 GB each). GPU 0 runs proposal/selection with PyTorch 2.1.0/CUDA 12.1; GPU 1 runs recognition with PyTorch 1.13.1/CUDA 11.7. The two stages run in separate processes with separate software environments. Batch size is one; TF32 is disabled.

Table 3 uses 50 units per dataset (sampling seed 20261012, support seed 2023): a DIOR image or complete DOTA parent. Synchronized end-to-end wall-clock timing covers in-memory raw image to final detections, including preprocessing, transfers, feature extraction, decoding, descriptors/caches, VJPs, search, recognition and post-processing; DOTA includes all chips and parent merging. Timing accounts for completion on both GPUs and includes inter-process transfers. Model loading, disk I/O and once-per-support preparation are excluded. After five warm-ups, three cyclic-order rounds are averaged per unit. Mean and P95 are computed over these 50 unit means, not individual requests.

SCQE averages 1.391 s/DIOR image and 17.791 s/DOTA parent: 65.4% and 50.0% above B0. Relative to B3, the increases are 0.153 s (12.4%) and 2.324 s (15.0%). Table 3 also reports P95. Peak `max_memory_allocated` is recorded per GPU; the reported maximum spans units, repeats and devices, excludes reserved memory and does not sum device peaks. All selectors report 13.03 GiB on DIOR and 13.02 GiB on DOTA-HBB. Equal reported peaks do not imply zero intermediate-storage cost.

3.5 Qualitative comparison

Figure 2 compares B0 and SCQE on DOTA with support seed 2023 and shared display thresholds. At IoU .50, let $d_q = \Delta\mathrm{TP}_q - \Delta\mathrm{FP}_q$ relative to B0. Among accepted, nonempty actual-feasible plans, select the query nearest median d_q within benefit ($d_q > 0$) and loss ($d_q < 0$) strata, with query-ID ties. The cases recover a missed object and introduce a background false positive, respectively: protected-output feasibility does not ensure improved detection on every query. Case selection serves visualization only.

Table 2 DIOR 5-shot ablations. AP/AP75 follow the five-support mean±sample SD convention; ΔAP is variant minus full mean AP. Rej. uses finite nonempty plans; Over uses all query–support pairs. A dash denotes a disabled gate. Record verification is described in Section 3.2.

Variant	AP	ΔAP	AP75	Rej. (%)	Over (%)
SCQE (full)	36.36 ± 1.94	0.00	32.87 ± 2.05	7.1	0
Canonical assignment	35.14 ± 1.99	-1.22	31.45 ± 2.10	6.4	0
Addition-only gain	34.89 ± 2.01	-1.47	30.90 ± 2.23	10.2	0
Isolated perturbations	35.08 ± 2.12	-1.28	31.63 ± 1.88	6.8	0
Addition-only + isolated	34.73 ± 1.31	-1.63	29.65 ± 2.08	8.6	0
No actual gate	35.05 ± 1.86	-1.31	29.99 ± 1.71	—	4.7

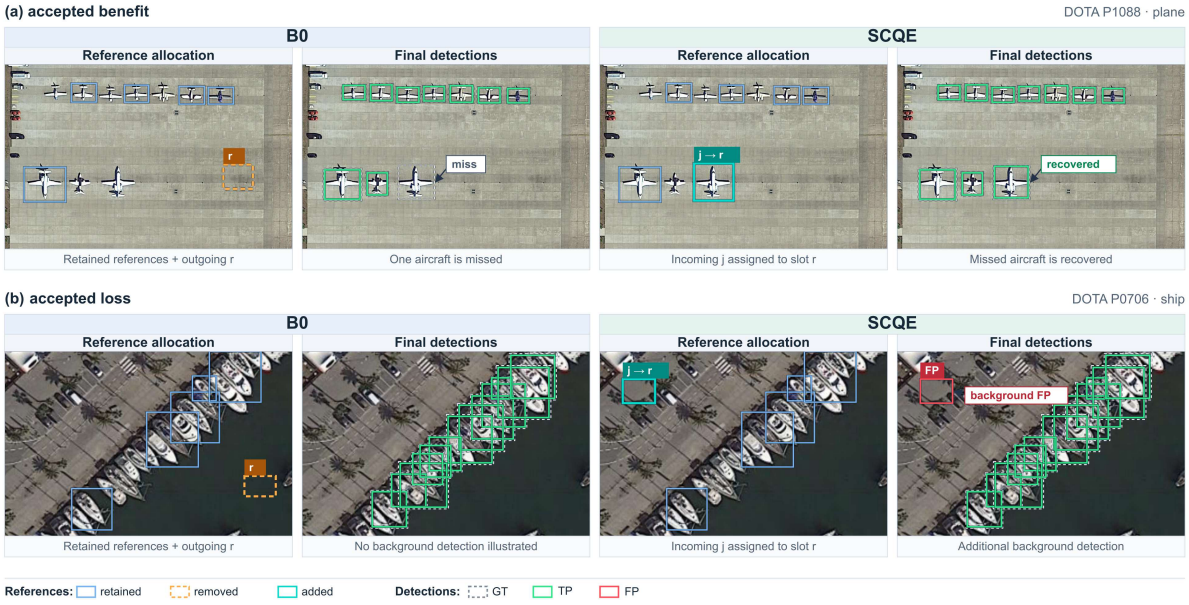


Fig. 2 Measured qualitative comparison of B0 and SCQE on DOTA images [2] under shared support and display settings. (a) An accepted exchange recovers a missed object. (b) An accepted exchange introduces a background false positive. Reference allocation and final detections are shown separately. Only a subset of references is displayed; the query count Q remains unchanged.

Table 3 5-shot query-inference efficiency comparison. Latency is mean/P95 across 50 evaluation units after averaging three repeats per unit; memory is the maximum per-device peak allocated GPU memory across all units and repeats. DOTA includes all chips and parent merging. Model loading, disk I/O and once-per-support preparation are excluded.

Selector	DIOR s/image mean/P95	DOTA s/parent mean/P95	Peak GiB DIOR/DOTA
B0	0.841/0.845	11.863/12.016	13.03/13.02
B3	1.238/1.315	15.467/15.987	13.03/13.02
SCQE	1.391/1.427	17.791/18.231	13.03/13.02

4 Conclusion

SCQE couples ordered reference allocation with complete gain and cumulative protected-output perturbation while fixing proposal weights, learned contents and query count. It improves native selection on both datasets, with 5-shot gains of 8.68 AP on DIOR and 4.93 on DOTA-HBB. Ablations support assignment and both decision terms. Whole-plan verification enforces the output tolerance, without per-query accuracy or optimality guarantees. Mean 5-shot latency exceeds B3 by 12.4% and 15.0%, with unchanged

reported peak memory. Candidate-density weighting, aggregate protection and feasible-prefix search remain limitations.

Statements and Declarations. The authors declare that they have no financial or non-financial conflicts of interest relevant to this work.

Funding. This study was supported by the Qing Lan Project and Jiangsu Province's 333 Talent Project.

Author contributions. Feng Jiang contributed to the study design and methodology and wrote the original draft. Zhipeng Ye conducted the experiments and analyzed the results. Lili Zhou supervised the study and reviewed the manuscript. All authors reviewed and approved the final manuscript.

Data Availability Statement. The DIOR dataset is publicly available from https://gcheng-nwpu.github.io/, and DOTA-v1.0 is available from https://captain-whu.github.io/DOTA/, subject to their respective access and use conditions. This study evaluates DIOR on 4,900 images after excluding 100 held-out images from the 5,000-image test partition. The excluded images are not used for threshold tuning, model selection or architecture selection. The research-code snapshot is publicly available at https://github.com/sjtukng/SCQE. The support/query manifests and the list of the 100 excluded DIOR images are available at https://github.com/sjtukng/SCQE/tree/main/manifests.

References

[1] Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing 159, 296–307 (2020). https://doi.org/10.1016/j.isprsjprs.2019.11.023.

[2] Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: DOTA: A large-scale dataset for object detection in aerial images. In: CVPR, pp. 3974–3983 (2018). https://doi.org/10.1109/CVPR.2018.00418.

[3] Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J.: Towards large-scale small object detection: Survey and benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(11), 13467–13488 (2023). https://doi.org/10.1109/TPAMI.2023.3290594.

[4] Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: ICCV, pp. 8420–8429 (2019). https://doi.org/10.1109/ICCV.2019.00851.

[5] 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: ICCV, pp. 9577–9586 (2019). https://openaccess.thecvf.com/content_ICCV_2019/html/Yan_Meta_R-CNN_Towards_General_Solver_for_Instance-Level_Low-Shot_Learning_ICCV_2019_paper.html.

[6] Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 119, pp. 9919–9928 (2020). https://proceedings.mlr.press/v119/wang20j.html.

[7] Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: FSCE: Few-shot object detection via contrastive proposal encoding. In: CVPR, pp. 7352–7362 (2021). https://doi.org/10.1109/CVPR46437.2021.00727.

[8] Fan, Q., Zhuo, W., Tang, C.-K., Tai, Y.-W.: Few-shot object detection with Attention-RPN and multi-relation detector. In: CVPR, pp. 4013–4022 (2020). https://doi.org/10.1109/CVPR42600.2020.00407.

[9] 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(11), 12832–12843 (2023). https://doi.org/10.1109/TPAMI.2022.3195735.

[10] 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: Computer Vision–ECCV 2024, pp. 247–264 (2025). https://doi.org/10.1007/978-3-031-73636-0_15.

[11] Huang, Y., Mei, J., Wu, Z., Yang, Y., Zhao, H., Jiu, M., Sahbi, H.: Less is better: Sparse instance learning for cross-domain few-shot object detection. Proceedings of the AAAI Conference on Artificial Intelligence 40(7), 5176–5184 (2026). https://doi.org/10.1609/aaai.v40i7.37432.

[12] Yu, X., Sha, Y., Liu, L., Shen, X., Yang, D.: A closer look at cross-domain few-shot object detection: Fine-tuning matters and parallel decoder helps. In: CVPR, pp. 26593–26603 (2026). https://openaccess.thecvf.com/content/CVPR2026/html/Yu_A_Closer_Look_at_Cross-Domain_Few-Shot_Object_Detection_Fine-Tuning_Matters_CVPR_2026_paper.html.

[13] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020, pp. 213–229 (2020). https://doi.org/10.1007/978-3-030-58452-8_13.

[14] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: ICLR (2021). https://openreview.net/forum?id=gZ9hCDWe6ke.

[15] Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: ICLR (2022). https://arxiv.org/abs/2201.12329.

[16] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: ICLR (2023). https://arxiv.org/abs/2203.03605.

[17] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Computer Vision–ECCV 2024, pp. 38–55 (2025). https://doi.org/10.1007/978-3-031-72970-6_3.

[18] Tang, Q., Liu, C., Zhang, S., Li, W., Fan, Q., Gao, Y.: Prompt-free universal region proposal network. In: CVPR, pp. 13080–13090 (2026). https://openaccess.thecvf.com/content/CVPR2026/html/Tang_Prompt-Free_Universal_Region_Proposal_Network_CVPR_2026_paper.html.

[19] Xu, L., Wu, Z., Qiu, W., Pang, S., Bai, X., Mei, K., Xue, J.: Redundant queries in DETR-based 3D detection methods: Unnecessary and prunable. Proceedings of the AAAI Conference on Artificial Intelligence 40(13), 11323–11331 (2026). https://doi.org/10.1609/aaai.v40i13.38113.

[20] Shi, Z., Ma, B.: Reading decoder trajectories: Training-free counterfactual query-trajectory reliability for small-object detection. arXiv:2609.06581 (2026). https://arxiv.org/abs/2609.06581.

[21] Oquab, M., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024). https://openreview.net/forum?id=a68SUt6zFt.

[22] Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., Gao, J.: Grounded language-image pre-training. In: CVPR, pp. 10965–10975 (2022). https://openaccess.thecvf.com/content/CVPR2022/html/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.html.

[23] Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions–I. Mathematical Programming 14, 265–294 (1978). https://doi.org/10.1007/BF01588971.

[24] He, S., Liang, Y., He, R., Zeng, J., He, J., Wu, C., Zheng, L.: QCPruner: Query-conditioned population coverage for visual token pruning. arXiv:2609.19990 (2026). https://arxiv.org/abs/2609.19990.

[25] Huang, Y., Yang, H., Dong, Z., Gudovskiy, D., Okuno, T., Nakata, Y., Du, Y., Zhang, S., Keutzer, K.: Output sensitivity-aware DETR quantization. In: 2nd International Workshop on Practical Deep Learning in the Wild (Practical-DL), AAAI 2023, extended abstract (2023). https://practical-dl.github.io/2023/extended_abstract/4/CameraReady/4.pdf.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM.tex