TCPL: TARGET CONSENSUS-GUIDED PROTOTYPE LEARNING FOR OPEN-SCENE TARGET DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral small target detection in open scenes is challenged by variations in target spectra and backgrounds, which limit the transferability of fixed target priors and target knowledge derived from labeled scenes and varying references. To address this challenge, we propose **T**arget **C**onsensus-Guided **P**rototype **L**earning (**TCPL**), a framework that learns transferable target knowledge for detection without adaptation to new scenes. During densely supervised training, TCPL uses sparse target points and their neighborhoods to construct local spectral–spatial references. Its asymmetric text guidance combines target and background descriptions on the support side while retaining only target semantics on the query side. By matching features from support-point neighborhoods with query-pixel features, TCPL aggregates predictions for the same query image under different support conditions into a consensus, mitigating the effects of variations among support samples. Detection supervision and knowledge distillation transfer this consensus knowledge into a learnable prototype bank and target anchor, enabling query-only inference. Furthermore, we construct three datasets with pixel-level target annotations and establish a unified open-scene evaluation setting.

1 INTRODUCTION

Spectral target detection (STD) uses fine-grained spectral responses to identify targets of interest in complex backgrounds and produce pixel-level detections. Unlike conventional optical images that mainly rely on color, texture, and shape, hyperspectral images capture the spectral variations of ground objects through contiguous narrow bands, providing additional discriminative information for targets with similar appearances but different material properties. To exploit this information, recent studies have progressively strengthened target-prior matching, discriminative target–background representations, and prototype memory modeling, while further exploring the transfer of detection knowledge to new tasks and scenes through few-shot learning and cross-domain adaptation (Shen et al., 2025; Li et al., 2025; Yang et al., 2025; Luo et al., 2025; Feng et al., 2025).

However, seeing a target category during training does not ensure consistent spectra or appearance in unseen scenes. Sensor response, illumination, atmospheric conditions, and target–background mixing can alter measured spectra (Somers et al., 2011; Claverie et al., 2018), while spatial-resolution differences affect apparent target size, texture, and subpixel mixing. As a result, the same category may exhibit different spectral and appearance distributions across training and test scenes. Complex backgrounds may also contain distractors with locally target-like spectra. Classical target-prior STD (Fig. 1(a)) matches a pre-acquired, fixed target spectrum against the test image, but cross-scene changes in target responses cause prior-target mismatches. Cross-domain transfer methods (Fig. 1(b)) mitigate distribution differences using source-domain supervision or external data, yet require target-domain samples for adaptation, hindering direct application to entirely unseen, unlabeled query scenes (Luo et al., 2025; Feng et al., 2025; Zhang et al., 2026). Thus, the former cannot accommodate target variations with fixed references, while the latter still requires target-scene participation in knowledge transfer. This raises an unresolved question: **How can transferable target knowledge be distilled from support-dependent local matching for open-scene small target detection, without query-scene adaptation?** To address this challenge, we study open-scene STD for pixel-level detection (Fig. 1(c)), **using labeled training scenes to learn target representations that remain discriminative across unseen scenes.**

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.

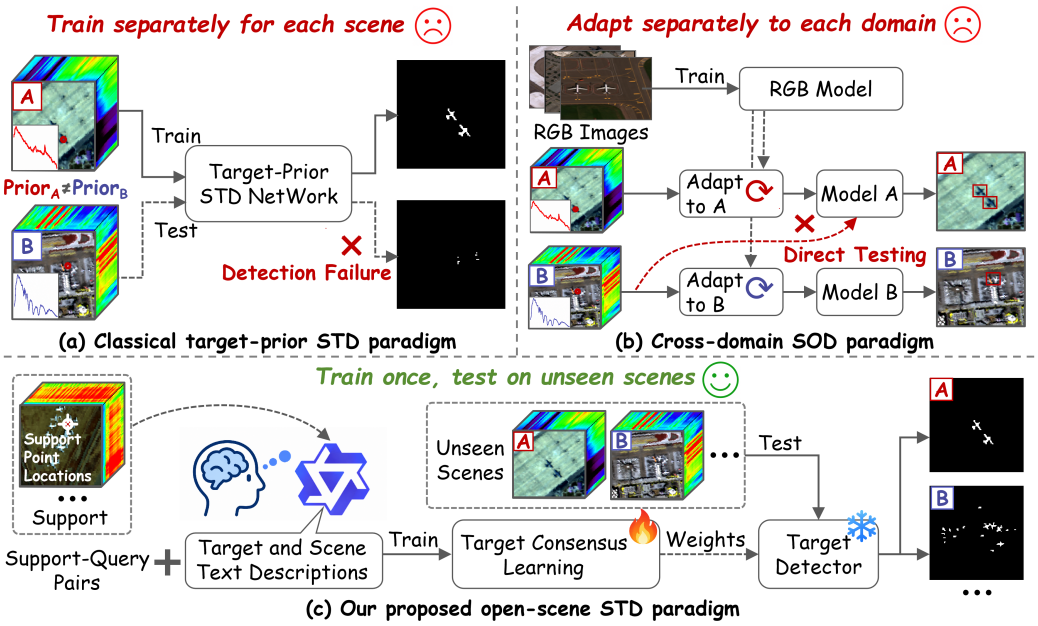


Figure 1: **Comparison of STD paradigms.** (a) Fixed target spectra in classical target-prior STD may mismatch targets in new scenes. (b) Cross-domain spectral object detection (SOD) requires per-domain adaptation, limiting direct application to unseen scenes. (c) TCPL combines sparse support-point references, semantic descriptions, and dense training-query supervision to learn transferable target representations for detection in unseen scenes.

Based on this analysis, we propose **T**arget **C**onsensus-guided **P**rototype **L**earning (**TCPL**), a framework that learns stable, transferable target representations from labeled scenes and varying references for open-scene spectral small target detection. First, we design the **Asymmetric Text-guided Knowledge Extraction (ATKE)** module to address open-scene variations in target appearance and spectra. It uses measured spectra at points inside targets as physical anchors. Target and background descriptions jointly guide discriminative knowledge extraction on the support side, while only target semantics guide the query side to reduce background interference. Our **target consensus under support variations (TCSV)** module reduces prediction inconsistencies across supports for the same query. It generates support-conditioned predictions through relational retrieval between support-point neighborhoods and query pixels and aggregates them into a consensus teacher. Finally, the **Consensus-guided Prototype Learning and Distillation (CPLD)** module uses the consensus prediction as a teacher with stopped gradients. Output distillation and detection supervision jointly optimize learnable target prototypes and a target anchor. This transfers discriminative knowledge across support conditions into prototypes that serve as references, reducing the need for additional reference information during detection.

The main contributions of this paper are as follows:

- **Open-scene spectral small target detection**: We propose **TCPL**, a framework that learns transferable target knowledge through support-conditioned training with dense query supervision for detection in unseen scenes without query-scene adaptation.
- **Target consensus and prototype learning**: We combine measured spectra with asymmetric target-background semantic guidance and aggregate predictions from support-query relational retrieval into a consensus teacher. Output distillation and detection supervision transfer this consensus into learnable target prototypes and anchors.
- **Datasets, benchmarking, and reproducibility**: We construct three STD datasets with pixel-level target annotations and establish a unified evaluation setting to assess performance under different support-point budgets in unseen scenes.

2 Related Work

Text-Guided Visual Localization Models. Vision–language pretraining provides reusable semantic representations for downstream visual tasks (Radford et al., 2021). Prior studies have introduced textual semantics into image fusion, infrared small target detection, and segmentation (Yi et al., 2026; Huang et al., 2025; Zhang et al., 2025). Semantic alignment and language–image pretraining are used to enhance visual representations and target discrimination. DGNet further uses target and background descriptions to modulate visual features, enhancing targets while suppressing backgrounds (Yu et al., 2026). These methods demonstrate the effectiveness of textual priors for target representation, but mainly focus on semantic enhancement within individual inputs rather than cross-scene transfer to unseen queries. We instead use text to learn transferable target representations from support-point spectra and adopt asymmetric guidance: target and background descriptions for support, but only target descriptions for query.

Few-Shot Visual Learning Models. Few-shot visual learning aims to construct category representations from limited support samples for query prediction. Early methods encode support samples into global vectors for feature reweighting or instance-level modulation (Kang et al., 2019; Yan et al., 2019). These global representations provide category-level priors but discard support-side spatial structure. This limits fine-grained support–query correspondence. Later methods explicitly model local correspondences through correlation and attention mechanisms to improve target matching in detection and segmentation (Fan et al., 2020; Min et al., 2021; Shi et al., 2022). With the development of visual foundation models, FSOD-VFM combines category prototypes, proposal matching, and graph diffusion for detection without additional training (Feng et al., 2026). Meanwhile, point prompts and pseudo-mask generation have been explored to reduce spatial annotation costs (Wei et al., 2026; Subhani, 2026). SPSD exploits spectral differences for remote-sensing target perception without dense annotations (Liu et al., 2025). For spectral small targets, annotated points provide location cues and measured spectral information from local neighborhoods. However, finer matching alone cannot ensure reliable knowledge transfer from these sparse and scene-sensitive observations. Such observations capture only limited target variability. Therefore, we focus on learning transferable target knowledge shared across support conditions.

Spectral Detection Models and Domain Generalization. STD mainly exploits spectral differences between targets and backgrounds for pixel-level discrimination. Recent studies improve spectral discriminability through spatial modeling, diffusion-based augmentation, and convolutional gating (Shen et al., 2025; Li et al., 2025). PE-PMN constructs target and background prototype memory banks from region-level weak labels to capture target variations and background diversity (Yang et al., 2025). For knowledge transfer, self-supervised pretraining, few-shot fine-tuning, and domain adaptation have been used to reduce distribution shifts across scenes or domains (Luo et al., 2025; Feng et al., 2025). VHCDN further transfers supervision from visible remote-sensing imagery to hyperspectral target detection (Zhang et al., 2026). These methods improve spectral representations and cross-domain transfer, but mainly focus on adapting to inter-domain distribution shifts. Less attention has been paid to the stability of target knowledge across different support conditions. To address this issue, TCPL extracts target knowledge shared across support conditions and distills it into learnable prototypes and a target anchor.

3 Preliminaries

Let $X \in \mathbb{R}^{C\times H_{\text{img}}\times W_{\text{img}}}$ be a spectral image with C spectral channels and $Y \in \{0,1\}^{H_{\text{img}}\times W_{\text{img}}}$ its binary target mask. The g-th support condition $\mathcal{S}_g$ and training episode $\mathcal{U}_g$ are defined as:

$$\mathcal{S}_g = \big(X_s^{(g)}, \mathcal{P}_s^{(g)}, d_g^+, d_g^-\big), \qquad \mathcal{P}_s^{(g)} = \{p_{s,k}^{(g)}\}_{k=1}^{K}, \qquad \mathcal{U}_g \;= \big(\mathcal{S}_g, X_q, Y_q\big). \tag{1}$$

The subscripts s and q denote support and query images. Each support provides K target-internal points $p_{s,k}^{(g)}$ and target/background descriptions d_g^+ and d_g^-. Each training group contains G episodes with different supports sharing the same query and dense mask. Sparse points define compact spectral–spatial references, limiting reference information and matching cost for each support-conditioned prediction. Dense query masks supervise both prediction paths, while consensus distillation transfers knowledge across support conditions into the Bank for query-only inference. Thus, support sparsity constrains reference construction rather than training supervision. Here, "unseen" refers to held-out scenes.

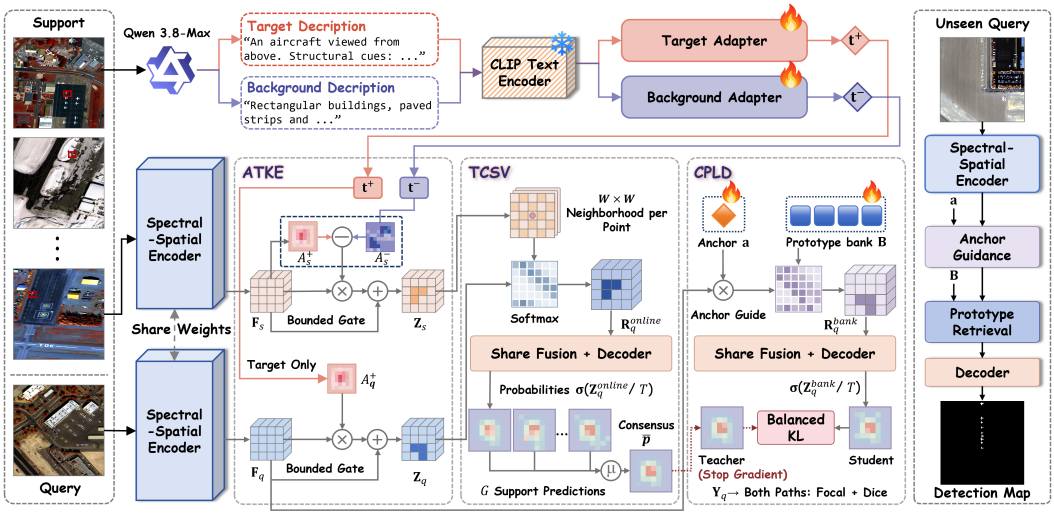


Figure 2: **Overview of the TCPL framework.** ATKE applies target–background text guidance to support features and target-only guidance to query features. TCSV retrieves local tokens from support-point neighborhoods and aggregates predictions for the same query across G support conditions into a stop-gradient consensus teacher. CPLD learns the target anchor $\mathbf{a}$ and prototype bank $\mathbf{B}$ through class-balanced output distillation and detection supervision. The rightmost pipeline performs query-only inference using the learned anchor and prototypes.

4 METHOD

Overview. As shown in Fig. 2, TCPL transfers knowledge from a support-conditioned Online path to a prototype-based Bank path for query-only inference. Both paths share a spectral–spatial encoder and detection decoder. During training, ATKE (Section 4.1) guides support features with target and background text, and query features with target text alone. TCSV (Section 4.2) matches support-point neighborhoods to query pixels, aggregating predictions for the same query across support conditions into a consensus teacher. CPLD (Section 4.3) distills this consensus into the Bank path by learning target prototypes and a target anchor. The framework is expressed as:

$$z_g^{\text{online}} = f_{\text{online}}(X_q; \mathcal{S}_g), \qquad z_q^{\text{bank}} = f_{\text{bank}}(X_q; \mathbf{B}, \mathbf{a}), \tag{2}$$

where $\mathcal{S}_g$ denotes the g-th support condition defined in Eq. 1. z denotes pixel-level logits, and σ is the sigmoid function. $\mathbf{B} \in \mathbb{R}^{M \times d}$ is a learnable prototype bank containing M target prototypes, where d is the feature dimension. $\mathbf{a} \in \mathbb{R}^d$ is a learnable target anchor.

Shared Representation. The shared spectral–spatial encoder $\mathcal{E}_\theta$ jointly models spectral shape, spectral variations, and their local spatial structure. It extracts support and query features as $\mathbf{F}_s = \mathcal{E}_\theta(X_s)$ and $\mathbf{F}_q = \mathcal{E}_\theta(X_q)$, respectively. App. A.1 details the architecture and retrieval complexity.

4.1 ASYMMETRIC TEXT-GUIDED KNOWLEDGE EXTRACTION (ATKE)

The support and query backgrounds differ. Directly applying support background semantics to the query may introduce a scene-specific negative-class bias. Therefore, ATKE uses target-background semantic contrast to extract discriminative knowledge on the support side, while retaining only target semantics on the query side to avoid reliance on background descriptions of the support scene.

Semantic Projection and Spatial Responses. We use a frozen CLIP text encoder $\mathcal{T}$ (Radford et al., 2021). Two trainable projectors, Π_+ and Π_-, map the target and background descriptions into the spectral feature space, respectively. A shared spatial feature mapping $\mathcal{G}_\theta$ transforms the support and query features for computing response maps:

$$\begin{aligned} \mathbf{t}^+ &= \Pi_+(\mathcal{T}(d^+)), & \mathbf{U}_s &= \mathcal{G}_\theta(\mathbf{F}_s), & A_s^{\pm}(u) &= \cos(\mathbf{U}_s(u), \mathbf{t}^{\pm}), \\ \mathbf{t}^- &= \Pi_-(\mathcal{T}(d^-)), & \mathbf{U}_q &= \mathcal{G}_\theta(\mathbf{F}_q), & A_q^{+}(u) &= \cos(\mathbf{U}_q(u), \mathbf{t}^{+}), \end{aligned} \tag{3}$$

where u denotes a spatial location in the feature map, and $\mathbf{U}_s(u)$ and $\mathbf{U}_q(u)$ are the local spectral–spatial features at that location. A_s^+ and A_s^- denote the target and background semantic response maps of the support image, respectively, and A_q^+ denotes the target semantic response map of the query image. The text projectors are trained jointly with the detection network to align text semantics with the spectral detection feature space.

Asymmetric Residual Guidance. On the support side, we construct a contrastive response $D_s = A_s^+ - A_s^-$ and modulate spatial features through bounded gating. On the query side, the gate is constructed using only the target response:

$$\begin{aligned} \widetilde{\mathbf{U}}_s &= \mathbf{U}_s \odot \left[1 + \eta \tanh(D_s/\tau_a)\right], \qquad \mathbf{Z}_s = \mathbf{F}_s + \alpha_s \mathcal{H}_s\left([\widetilde{\mathbf{U}}_s, A_s^+, A_s^-, D_s]\right), \\ \widetilde{\mathbf{U}}_q &= \mathbf{U}_q \odot \left[1 + \eta \tanh(A_q^+/\tau_a)\right], \qquad \mathbf{Z}_q = \mathbf{F}_q + \alpha_q \mathcal{H}_q\left([\widetilde{\mathbf{U}}_q, A_q^+]\right), \end{aligned} \tag{4}$$

where $[\cdot]$ denotes channel concatenation, and $\odot$ denotes element-wise multiplication. τ_a is the gating temperature, and $\mathcal{H}_s$ and $\mathcal{H}_q$ are guidance mappings comprising dilated convolution, normalization, and activation. $\alpha_s, \alpha_q \in (0, 1)$ are learnable residual coefficients parameterized by sigmoid functions. By default, $\eta = 0.5$, which bounds the gating multipliers within $[0.5, 1.5]$ to prevent semantic modulation from overriding the spectral information.

4.2 Target Consensus under Support Variations (TCSV)

ATKE provides target-related semantic guidance, but different support images may still lead the model to produce different predictions. TCSV establishes relations between point neighborhoods and query pixels for each support condition. It then aggregates multiple sets of predictions for the same query to construct a consensus teacher for the prototype path.

Relational Retrieval from Support Points to Query Pixels. For each prompt point $p_{s,k}$, we extract a $W \times W$ neighborhood from the enhanced support features $\mathbf{Z}_s$, yielding a set of local tokens $\mathcal{N}_k = \{\mathbf{s}_{kj}\}_{j=1}^{W^2}$. For a query feature $\mathbf{q}$ and reference tokens $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^N$, define:

$$r_i = \cos(W_q\mathbf{q}, W_k\mathbf{v}_i), \qquad \omega_i = \frac{\exp(r_i/\tau_r)}{\sum_{j=1}^N \exp(r_j/\tau_r)}. \tag{5}$$

where W_q and W_k map query and reference features into the matching space, and τ_r is the softmax temperature. Both paths share the retrieval operator $\mathcal{R}$. Online applies it separately to each $\mathcal{N}_k$, with softmax over its W^2 tokens, then averages over K points. $r_{kj}(u)$ denotes r_j evaluated with $\mathbf{q} = \mathbf{Z}_q(u)$ and $\mathcal{V} = \mathcal{N}_k$:

$$\mathcal{R}(\mathbf{q}, \mathcal{V}) = W_r \sum_{i=1}^N \omega_i W_v \mathbf{v}_i, \qquad \mathbf{R}_q(u) = \frac{1}{K}\sum_{k=1}^K \mathcal{R}\big(\mathbf{Z}_q(u), \mathcal{N}_k\big), \tag{6}$$

where W_v and W_r are the value projection and retrieval output projection, respectively. We extract the maximum, top-3 mean, mean, and standard deviation from $r_{kj}(u)$ to form a four-dimensional statistical vector $\mathbf{h}_q(u)$. We then fuse the query features, retrieved features, and relational statistics to generate the Online prediction:

$$\mathbf{F}_q^{\mathrm{rel}} = \mathcal{F}_\theta\left([\mathbf{Z}_q, \mathbf{R}_q, |\mathbf{Z}_q - \mathbf{R}_q|, \mathbf{Z}_q \odot \mathbf{R}_q, \phi(\mathbf{h}_q)]\right), \quad z^{\mathrm{online}} = \mathcal{D}_\theta\left(\mathbf{F}_q^{\mathrm{rel}} + W_{\mathrm{skip}}[\mathbf{Z}_q, A_q^+]\right), \tag{7}$$

where ϕ embeds the relational statistics, and $\mathcal{F}_\theta$ and $\mathcal{D}_\theta$ are the relational fusion mapping and the shared detection decoder, respectively. W_{skip} maps the query features and target response into skip-connection features.

Consensus Teacher Construction for A Fixed Query. For a fixed query, z_g^{online} denotes the output of Eq. 7 under the g-th support condition. We apply temperature scaling and average the resulting probabilities over G support conditions to form the consensus teacher:

$$p_{g,T} = \sigma\big(z_g^{\mathrm{online}}/T\big), \qquad \underbrace{\bar{p}_T}_{\text{Teacher}} = \mathrm{sg}\left[\frac{1}{G}\sum_{g=1}^G p_{g,T}\right], \tag{8}$$

where T is the distillation temperature, and sg denotes the stop-gradient operation. The consensus distribution serves as the teacher signal for CPLD.

4.3 CONSENSUS-GUIDED PROTOTYPE LEARNING AND DISTILLATION (CPLD)

The consensus teacher aggregates predictions conditioned on different support samples during training. CPLD transfers this knowledge into learnable prototypes and a target anchor for query-only inference.

Learnable Prototypes and Target Anchor. To cover diverse support features, we initialize the prototype bank $\mathbf{B}$ using farthest point sampling (FPS) (Qi et al., 2017). The target anchor $\mathbf{a}$ is initialized from the projected target-text features of support samples. Both are subsequently optimized jointly. The details are described in App. A.1. In the Bank path, the normalized target anchor replaces the projected target-text features. The query guidance module shared with the Online path produces enhanced features and a target response:

$$\left(\mathbf{Z}_q^{\text{bank}}, A_q^{\text{bank}}\right) = \text{ATKE}_q\left(\mathbf{F}_q, \text{nrm}(\mathbf{a})\right), \tag{9}$$

where nrm denotes ℓ_2 normalization, $\mathbf{b}_m \in \mathbb{R}^d$ is the m-th prototype in $\mathbf{B}$ ($m = 1, \ldots, M$), and $\bar{\mathbf{b}}_m = \text{nrm}(\mathbf{b}_m)$. Let $\bar{\mathcal{B}} = \{\bar{\mathbf{b}}_m\}_{m=1}^M$, the Bank path applies the same retrieval operator to $\bar{\mathcal{B}}$, with softmax over all M prototypes:

$$\mathbf{R}_q^{\text{bank}}(u) = \mathcal{R}\big(\mathbf{Z}_q^{\text{bank}}(u), \bar{\mathcal{B}}\big). \tag{10}$$

Then, we extract the same types of relational statistics as in the Online path and reuse the fusion mapping, skip connection, and decoder in Eq. 7 to obtain z_q^{bank}.

Output Distillation of Multi-Support Consensus. Within the soft-target distillation framework (Hinton et al., 2015), CPLD uses the consensus teacher $\bar{p}_T$ constructed by TCSV in Eq. 8 to guide the Bank student path in learning target prediction without external references. Applying the same T to the Bank output yields the student probabilities:

$$\underbrace{p_T^{\text{bank}}}_{\text{Student}} = \sigma\big(z_q^{\text{bank}}/T\big). \tag{11}$$

Using a class-balanced aggregation operator, we define the distillation loss as:

$$\mathcal{L}_{\text{KD}} = T^2 \mathcal{A}_{Y_q}\left(D_{\text{B}}(\bar{p}_T \| p_T^{\text{bank}})\right), \tag{12}$$

where D_{B} denotes Bernoulli KL divergence, and $\mathcal{A}_{Y_q}$ equally weights the mean target and background errors. The Bank path distills the aggregated predictions across support conditions from the consensus teacher, through which gradients are stopped.

4.4 TRAINING OBJECTIVE AND INFERENCE

Algorithm 1 TCPL training and distillation

Input: $\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{val}}$, support points and target/background text, group size G, point count K, relation window size $W \times W$, prototype count M.

Output: Best weights θ^*, prototypes $\mathbf{B}^*$, and anchor $\mathbf{a}^*$.

Initialize $\theta, \mathbf{B}, \mathbf{a}$.
for each epoch **do**
  Group each query with G different supports.
  for each group **do**
    $\mathbf{F}_q \leftarrow \mathcal{E}_\theta(X_q)$ ▷ Once per query
    $\{z_g^{\text{online}}\}_{g=1}^G \leftarrow \text{Online}(\mathbf{F}_q; \{\mathcal{S}_g\}_{g=1}^G)$. ▷ Eqs. 3–7
    $z_q^{\text{bank}} \leftarrow \text{Bank}(\mathbf{F}_q; \mathbf{B}, \mathbf{a})$. ▷ Eqs. 9, 10, 7
    Compute Focal + Dice losses for both paths.
    Construct stop-gradient $\bar{p}_T$. ▷ Eq. 8
    Compute $\mathcal{L}_{\text{KD}}$. ▷ Eqs. 11, 12
    Update $(\theta, \mathbf{B}, \mathbf{a})$ with AdamW on $\mathcal{L}_{\text{total}}$. ▷ Eq. 13
  end for
  Evaluate Bank on $\mathcal{D}_{\text{val}}$. Save the checkpoint with the best validation AP.
end for
return $(\theta^*, \mathbf{B}^*, \mathbf{a}^*)$.

Joint Training Objective. We supervise both the Online and Bank paths with Focal loss (Lin et al., 2017) and Dice loss (Drozdzal et al., 2016), and combine them with consensus-teacher distillation:

$$\mathcal{L}_{\text{total}} = \frac{\mathcal{L}_{\text{sup}}^{\text{online}} + \mathcal{L}_{\text{sup}}^{\text{bank}}}{2} + \mathcal{L}_{\text{KD}}, \tag{13}$$

where $\mathcal{L}_{\text{sup}}^{\text{online}}$ averages the supervised loss over the G Online predictions, while $\mathcal{L}_{\text{sup}}^{\text{bank}}$ supervises the Bank prediction. See App. A.2 for loss definitions and Algo. 1 for training.

Model Inference. The default inference mode uses only the Bank path, as shown in the rightmost pipeline of Fig. 2. It encodes the query, applies target-anchor guidance and prototype retrieval, and decodes the resulting features to obtain $\widehat{Y}_q(u) = \sigma(z_q^{\text{bank}}(u))$. The inference procedure is detailed in App. A.3, Algo. 2.

5 EXPERIMENTS

Datasets and Evaluation. We construct three pixel-level annotated datasets for open-scene STD: Aircraft-14, Vehicle-74, and Plane-400. Aircraft-14 combines AVIRIS aircraft imagery with ABU-Airport scenes, Vehicle-74 is built from AeroRIT (Rangnekar et al., 2020), and Plane-400 adds manual pixel-level annotations to RarePlanes imagery (Shermeyer et al., 2021). They contain 14, 74, and 400 images with 192, 51, and 8 retained spectral bands, respectively, using fixed training, validation, and test splits. Details are provided in App. B.1. We compare against six few-shot segmentation methods, HSNet, DCAMA, IFA, DCDNet, FPTrans, and MPA (Min et al., 2021; Shi et al., 2022; Nie et al., 2024; Cong et al., 2026; Zhang et al., 2022; Nie et al., 2026), together with two HTD baselines, SelfMTL and TCFSL (Luo et al., 2025; Feng et al., 2025). Performance is measured by pixel-level AP and ROC-AUC (AUC), reflecting precision–recall performance under class imbalance and overall target–background ranking, respectively.

Experimental Setup. Training tasks use support and query images from the training set, with each support image providing K fixed prompt points inside the target. Before training, Qwen3.8-Max (Qwen Team, 2026) generates background descriptions from full support RGB images and target descriptions from the full images together with crops of the instances selected by support points. Target descriptions focus exclusively on the prompted instance to reduce semantic ambiguity caused by other instances and appearance differences, while background descriptions summarize the scene and potential distractors. Text is used only in the Online path during training. We use a frozen CLIP ViT-B/32 text encoder and optimize the model with AdamW, with both the learning rate and weight decay set to 10^{-4}. See App. B.2 for additional implementation and hyperparameter settings.

5.1 EXPERIMENTAL RESULTS

Comparative Experiments. Tables 1 and 2 report Aircraft-14 and Vehicle-74 results, respectively. App. B.3 adds Plane-400 comparisons and visualizations. TCPL achieves the highest mean AP across all evaluated support-point budgets on both datasets and the highest mean AUC on Vehicle-74. With 1 support point on Aircraft-14, its AP/AUC gains over the respective best baselines reach 20.58/15.88 percentage points (pp). Vehicle-74 AP gains are 24.71, 14.75, and 11.31 pp with 1, 3, and 5 support points, respectively. Modest few-shot segmentation performance may stem from sparse annotations limiting foreground representations and spectral differences hindering support–query matching. Despite stronger quantitative results, SelfMTL and TCFSL retain background responses on Vehicle-74 with 1 support point (Fig. 3). Some segmentation methods show diffuse responses or weak target activation, whereas TCPL focuses on true targets. With 3 support points on Aircraft-14, TCPL improves AP over SelfMTL by only 0.07 pp (AP std. dev.: 14.66), while AUC is 1.18 pp lower, indicating limited improvement.

Table 1: Aircraft-14 performance across support-point budgets (%, mean ± std over three seeds).

Method	1-point		3-point		5-point	
	AP ↑	AUC ↑	AP ↑	AUC ↑	AP ↑	AUC ↑
HSNet (Min et al., 2021)	5.52 ± 3.12	79.82 ± 2.91	16.78 ± 3.92	84.17 ± 2.15	9.29 ± 5.98	80.75 ± 4.83
DCAMA (Shi et al., 2022)	3.80 ± 0.63	81.81 ± 2.21	7.21 ± 5.33	82.14 ± 1.77	6.91 ± 5.22	78.81 ± 1.06
IFA (Nie et al., 2024)	1.59 ± 0.55	57.51 ± 14.57	1.63 ± 0.27	61.30 ± 6.52	3.48 ± 0.90	77.46 ± 3.67
DCDNet (Cong et al., 2026)	8.98 ± 5.55	68.99 ± 20.26	11.36 ± 8.22	77.58 ± 2.19	9.78 ± 4.64	81.42 ± 5.13
FPTrans (Zhang et al., 2022)	1.80 ± 0.24	62.75 ± 4.60	3.06 ± 0.72	61.01 ± 4.77	2.17 ± 0.42	63.08 ± 8.40
MPA (Nie et al., 2026)	2.64 ± 0.12	53.52 ± 7.12	2.50 ± 0.48	55.82 ± 10.35	2.69 ± 1.03	58.60 ± 5.63
SelfMTL (Luo et al., 2025)	12.60 ± 8.60	71.80 ± 12.68	30.80 ± 1.53	**94.72 ± 0.94**	32.22 ± 3.05	94.63 ± 0.89
TCFSL (Feng et al., 2025)	22.38 ± 4.55	77.88 ± 15.83	19.82 ± 5.51	65.51 ± 0.68	21.48 ± 7.97	72.13 ± 12.53
TCPL (ours)	**42.96 ± 4.80**	**97.69 ± 0.41**	**30.87 ± 14.66**	93.54 ± 3.79	**35.33 ± 3.07**	**96.10 ± 0.46**

Ablation Studies. Table 3 compares component variants, with configurations and further analysis in App. B.4. Removing ATKE or CPLD reduces AP by up to 8.62 and 10.16 pp, respectively, supporting the contributions of semantic guidance and prototype-based knowledge transfer. TCSV further improves AP on both datasets, indicating the benefit of attention retrieval with multi-support

Table 2: Vehicle-74 performance across support-point budgets (%, mean ± std over three seeds).

Method	1-point		3-point		5-point	
	AP ↑	AUC ↑	AP ↑	AUC ↑	AP ↑	AUC ↑
HSNet (Min et al., 2021)	11.24 ± 0.74	70.45 ± 1.74	12.41 ± 3.91	73.43 ± 6.59	12.89 ± 1.86	76.34 ± 3.42
DCAMA (Shi et al., 2022)	27.11 ± 1.84	90.67 ± 0.87	27.51 ± 2.76	89.51 ± 1.70	27.87 ± 3.06	89.41 ± 1.77
IFA (Nie et al., 2024)	12.08 ± 1.65	78.00 ± 0.51	10.58 ± 1.57	75.95 ± 2.92	11.17 ± 1.07	76.58 ± 1.39
DCDNet (Cong et al., 2026)	7.88 ± 7.54	58.76 ± 15.49	2.88 ± 0.17	52.38 ± 1.80	2.70 ± 0.21	53.59 ± 1.43
FPTrans (Zhang et al., 2022)	17.87 ± 1.05	86.64 ± 0.35	19.79 ± 0.92	88.33 ± 0.78	18.90 ± 1.00	89.36 ± 0.54
MPA (Nie et al., 2026)	15.23 ± 2.99	84.63 ± 3.85	18.38 ± 2.59	87.98 ± 2.84	20.60 ± 3.37	87.77 ± 1.10
SelfMTL (Luo et al., 2025)	37.81 ± 1.31	91.46 ± 1.32	31.91 ± 1.36	91.11 ± 0.48	32.91 ± 1.42	91.91 ± 0.73
TCFSL (Feng et al., 2025)	28.84 ± 15.58	82.71 ± 4.99	44.88 ± 5.72	92.46 ± 2.39	46.35 ± 2.68	89.42 ± 5.48
TCPL (ours)	**62.52 ± 1.50**	**93.73 ± 0.95**	**59.63 ± 3.67**	**93.67 ± 1.59**	**57.66 ± 1.80**	**94.79 ± 2.07**

consensus. Compared with query-only training using dense masks, full TCPL improves AP by 6.21 and 10.73 pp on Aircraft-14 and Vehicle-74, respectively (Table 4), supporting the effectiveness of the full framework. Consensus distillation complements dense detection supervision, as adding KD improves AP and AUC on both datasets (Table 5).

Table 3: Ablation study of the TCPL components.

	Components			Aircraft-14		Vehicle-74	
ID	ATKE	TCSV	CPLD	AP ↑	AUC ↑	AP ↑	AUC ↑
1	–	–	–	33.95 ± 11.33	91.52 ± 10.32	53.47 ± 4.88	92.75 ± 1.84
2	✓	–	–	33.84 ± 7.38	96.17 ± 1.76	56.21 ± 0.83	93.17 ± 0.72
3	–	✓	✓	35.50 ± 5.27	94.23 ± 2.26	53.90 ± 2.29	93.01 ± 0.16
4	✓	✓	–	32.80 ± 14.74	95.85 ± 2.97	55.72 ± 0.48	92.65 ± 0.97
5	✓	–	✓	42.01 ± 1.61	**97.76 ± 0.31**	55.40 ± 3.27	92.24 ± 2.22
6	✓	✓	✓	**42.96 ± 4.80**	97.69 ± 0.41	**62.52 ± 1.50**	**93.73 ± 0.95**

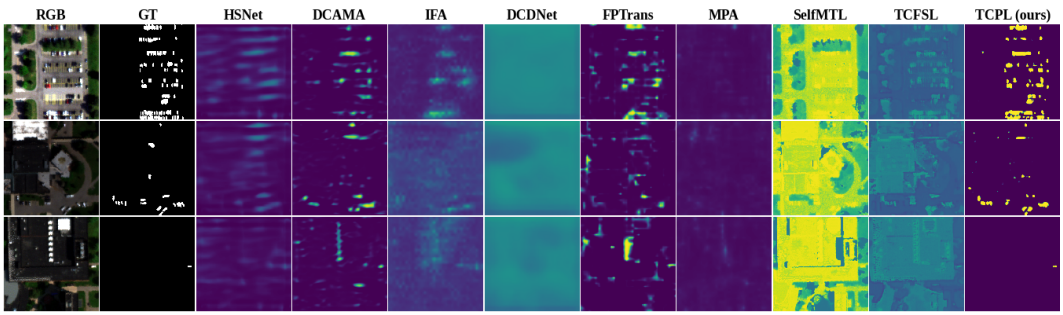


Figure 3: Qualitative comparison on Vehicle-74 in the 1-point setting.

Parameter Analysis. Fig. 4(a) varies the relation window with 1 support point. AP changes little from 3×3 to 7×7 and decreases at 9×9, possibly because larger neighborhoods include more background. Fig. 4(b) isolates the point budget using fixed support pools, support–query pairs, text, and evaluation queries, with nested sets of 1, 3, or 5 points per support image. Both the standard and controlled protocols fix points throughout training. Increasing the point budget improves AP in this controlled study, but neither AUC nor the main-table AP results improve monotonically. More points therefore do not guarantee better performance.

Table 4: Ablation of text guidance.

Variant	Aircraft-14		Vehicle-74	
	AP	AUC	AP	AUC
Query-only	36.75±5.51	97.40±1.09	51.79±0.84	93.59±1.54
Query+Text	35.85±7.54	96.83±1.10	58.04±0.65	**94.25±1.03**
Dual-text	**42.96±4.80**	**97.69±0.41**	**62.52±1.50**	93.73±0.95

Table 5: Ablation of loss functions.

Loss	Aircraft-14		Vehicle-74	
	AP	AUC	AP	AUC
Focal+Dice	42.46±1.69	96.56±1.01	57.58±0.50	92.77±1.68
Full (+KD)	**42.96±4.80**	**97.69±0.41**	**62.52±1.50**	**93.73±0.95**

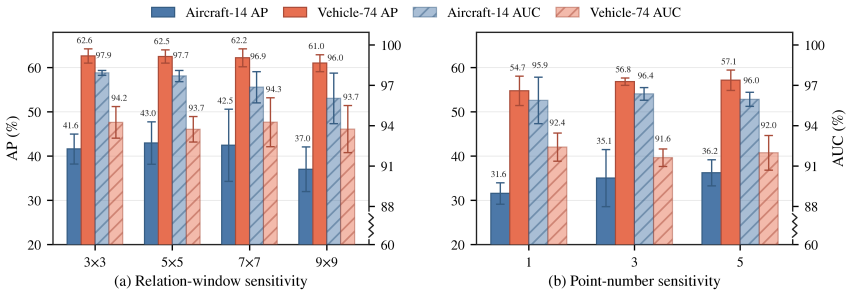


Figure 4: Parameter sensitivity analysis on Aircraft-14 and Vehicle-74.

Feasibility Analysis. Fig. 5 examines TCPL under support–query spectral variation, which causes clear fluctuations in support-conditioned predictions. TCSV aggregates them into a more reliable consensus teacher, which CPLD transfers to the spectral–spatial Bank. The resulting query-only path improves AP and yields coherent responses for both aircraft, with stronger prototype attention to A1 and A2 than to the background. These results demonstrate the conversion of support-dependent predictions into transferable target representations. The San Diego dataset further illustrates spectral and appearance differences across acquisition scenes despite a shared target category and sensor. Detection requires no test-scene adaptation (App. B.5).

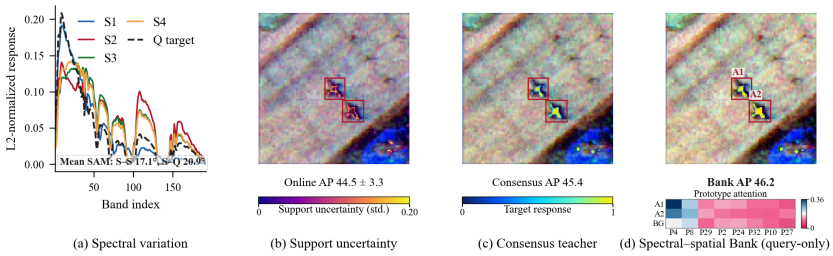


Figure 5: TCSV and CPLD under spectral variation on Aircraft-14. (a) Support–support and support–query spectral variations (mean spectral angle mapper (SAM): $17.1°/20.9°$). (b) Online predictions vary across supports (AP: 44.5 ± 3.3). (c) TCSV forms a consensus teacher (AP: 45.4). (d) CPLD enables query-only prototype detection (AP: 46.2).

6 Conclusion

We present TCPL for known-category open-scene STD, combining dense training supervision with sparse local references and semantic guidance for query-only detection in unseen scenes. ATKE enhances target representations through asymmetric semantic guidance, while TCSV aggregates predictions across support conditions into a consensus teacher. CPLD distills this consensus into a prototype bank and a target anchor, enabling default inference without support samples, text inputs, or test-time parameter updates. Overall, TCPL provides a new way to transform support-dependent matching knowledge into independently deployable target representations.

AI USE STATEMENT

In this work, we used generative AI tools to generate target and background textual descriptions with Qwen3.8-Max, which were used as textual inputs to the proposed method, as described in the experimental setup. We did not use generative AI tools for any other research tasks for which disclosure is required, and the remaining required-disclosure categories do not apply to this work. Additionally, we used an LLM to assist with literature search and organization and LaTeX formatting of the manuscript. We reviewed all AI-assisted work. The generated target and background descriptions were manually inspected before being used in our experiments, and literature-related information identified or organized with LLM assistance was verified against the sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The method and training procedure are described in Sec. 4 and App. A. Dataset descriptions and experimental settings are provided in App. B. An anonymized repository is available at `https://anonymous.4open.science/r/ICLR-TCPL-5EFE/`.

REFERENCES

Martin Claverie, Junchang Ju, Jeffrey G. Masek, Jennifer L. Dungan, Eric Vermote, Jean-Claude Roger, Sergii Skakun, and Christopher Justice. The harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219:145–161, 2018.

Runmin Cong, Anpeng Wang, Bin Wan, Cong Zhang, Xiaofei Zhou, and Wei Zhang. Divide-and-conquer decoupled network for cross-domain few-shot segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3416–3424, 2026.

Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, volume 10008 of *Lecture Notes in Computer Science*, pp. 179–187. Springer, 2016.

Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. 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.

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 *International Conference on Learning Representations*, 2026.

Shou Feng, Xueqing Wang, Rui Feng, Fengchao Xiong, Chunhui Zhao, Wei Li, and Ran Tao. Transformer-based cross-domain few-shot learning for hyperspectral target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

Feng Huang, Shuyuan Zheng, Zhaobing Qiu, Huanxian Liu, Huanxin Bai, and Liqiong Chen. Text-IRSTD: Leveraging semantic text to promote infrared small target detection in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10635–10644, 2025.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8420–8429, 2019.

Zhaokui Li, Fan Mo, Xiaobin Zhao, Cuiwei Liu, Xuewei Gong, Wei Li, Qian Du, and Bo Yuan. Hyperspectral target detection using diffusion model and convolutional gated linear unit. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Peifu Liu, Tingfa Xu, Guokai Shi, Jingxuan Xu, Huan Chen, and Jianan Li. Spectrum-oriented point-supervised saliency detector for hyperspectral images. *IEEE Transactions on Instrumentation and Measurement*, 74:1–15, 2025.

Fulin Luo, Shanshan Shi, Kai Qin, Tan Guo, Chuan Fu, and Zhiping Lin. SelfMTL: Self-supervised meta-transfer learning via contrastive representation for hyperspectral target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.

Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3380–3390, 2024.

Jiahao Nie, Guanqiao Fu, Wenbin An, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. Cross-domain few-shot segmentation via multi-view progressive adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41594–41604, 2026.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Aneesh Rangnekar, Nilay Mokashi, Emmett J. Ientilucci, Christopher Kanan, and Matthew J. Hoffman. AeroRIT: A new scene for hyperspectral image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):8116–8124, 2020.

Dunbin Shen, Xuanbing Zhu, Jiacheng Tian, Jianjun Liu, Zhenrong Du, Hongyu Wang, and Xiaorui Ma. HTD-Mamba: Efficient hyperspectral target detection with pyramid state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 2025. Article 5507315.

Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. RarePlanes: Synthetic data takes flight. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 207–217, 2021.

Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *Computer Vision – ECCV 2022*, volume 13680 of *Lecture Notes in Computer Science*, pp. 151–168. Springer, 2022.

Ben Somers, Gregory P. Asner, Laurent Tits, and Pol Coppin. Endmember variability in spectral mixture analysis: A review. *Remote Sensing of Environment*, 115(7):1603–1616, 2011.

M. Naseer Subhani. ReSAM: Refine, requery, and reinforce: Self-prompting point-supervised segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhaoyang Wei, Xumeng Han, Xuehui Yu, Xue Yang, Guorong Li, Zhenjun Han, and Jianbin Jiao. SAPNet++: Evolving point-prompted instance segmentation with semantic and spatial awareness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9577–9586, 2019.

Bo Yang, Jinjian Wu, and Changzhe Jiao. Proxy-enhanced prototype memory network for weakly supervised hyperspectral target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

Xunpeng Yi, Han Xu, Yibing Zhang, Hao Zhang, Linfeng Tang, and Jiayi Ma. GText-IF: Leveraging text-driven semantics for degradation-aware image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. In press.

Chenglong Yu, Mingzhu Xu, Jing Wang, Tongtong Wang, Pingping Miao, and Liqiang Nie. DGNet: Dual-knowledge guided network for infrared small target detection. In *Proceedings of the ACM International Conference on Multimedia*, 2026.

Hongqi Zhang, He Sun, Hongmin Gao, Lianru Gao, and Bing Zhang. Hyperspectral remote sensing object detection via cross-domain learning from visible images. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–18, 2026.

Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mingjin Zhang, Xiaolong Li, Fei Gao, Jie Guo, Xinbo Gao, and Jing Zhang. SAIST: Segment any infrared small target model guided by contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9549–9558, 2025.

PROMPT CONSTRUCTION FOR TEXTUAL DESCRIPTIONS

We use Qwen3.8-Max to construct separate target and background descriptions from training-support RGB images. The instance crops are defined by boxes enclosing the target instances selected by the fixed support points. The target request receives the full support image and the selected crops, and the background request receives only the full image. The following prompts summarize the semantic constraints used for description generation.

Target-structure prompt

```
Extract coarse structural attributes only for the selected
targets.  Use the full image for context and the tight instance
crops with limited surrounding context for structural evidence.
For aircraft, assess the fuselage, wings, tail, outline, and
bilateral symmetry.  For vehicles, assess the footprint,
longitudinal axis, roof/body structure, and bilateral symmetry.
Use predefined attribute categories and mark unresolvable
structures as unknown.  Do not infer invisible details or
include background, color, brightness, absolute orientation,
coordinates, counts, subtypes, materials, or spectral
properties.
```

Background-description prompt

```
Describe the visible non-target surroundings, focusing on
land-cover structures, textures, and their relative layout.
Ignore all targets and no-data borders.  Do not infer scene
elements from the target category or include target locations,
colors, brightness, counts, coordinates, materials, or spectral
properties.  Produce a concise English description within 32
CLIP tokens.
```

Text construction. Target attributes are converted into a fixed category sentence ("An aircraft viewed from above." or "A road vehicle viewed from above.") followed by at most three structural phrases. For conflicting attribute values across selected instances, the most frequent value is retained, with a fixed priority order for ties. Complete phrases are omitted as needed to keep the combined target and background text within 77 CLIP tokens. The background description is generated once per support image and reused across point budgets.

A Appendix: Method Details

A.1 Network Architecture and Prototype Initialization

The shared encoder takes a spectral vector and its differences between adjacent channels as input. For the spectral vector $\mathbf{x}(u) \in \mathbb{R}^C$ at location u, we define:

$$\begin{aligned} \bar{\mathbf{x}}(u) &= \frac{\mathbf{x}(u)}{\max(\|\mathbf{x}(u)\|_2, \epsilon)}, \\ \Delta\bar{\mathbf{x}}(u) &= [\bar{x}_2(u) - \bar{x}_1(u), \ldots, \bar{x}_C(u) - \bar{x}_{C-1}(u), 0]^{\mathsf{T}}, \end{aligned} \tag{A.1}$$

where $\bar{x}_c(u)$ denotes the c-th component of the normalized spectrum. The sequence branch extracts spectral responses from $\bar{\mathbf{x}}(u)$, while the shape branch jointly encodes $\bar{\mathbf{x}}(u)$ and $\Delta\bar{\mathbf{x}}(u)$. The two types of features are fused and then passed through spatial residual blocks to produce d-dimensional features. The main network architecture is listed in Table 6.

Table 6: Main network architecture of TCPL

Component	Architecture
Spectral sequence branch	Conv1D with kernel length 1, depthwise Conv1D with kernel length 3, GELU, Conv1D with kernel length 1, and average pooling along the spectral dimension.
Spectral shape branch	1×1 Conv2D, GroupNorm, GELU, and 3×3 depthwise Conv2D.
Spatial encoding	Two depthwise convolutional residual blocks with dilation rates of 1 and 2, respectively.
Text projection	Separate two-layer projectors with dimensions $512 \to 16 \to 16$ for the target and background.
ATKE	A shared 3×3 spatial mapping, and support-side and query-side guidance mappings with a dilation rate of 3.
Relation module	8-dimensional query–key matching, 16-dimensional value projection, 4-dimensional relational statistics, and relational fusion.
Detection decoder	A spatial residual block and a detection head with dimensions $16 \to 8 \to 1$.

The gating temperature of ATKE is $\tau_a = 0.1$, and the gating magnitude is $\eta = 0.5$. The relational retrieval temperature is $\tau_r = 0.1$. The Online path computes attention within a $W \times W$ neighborhood of each prompt point and then averages the retrieval results across prompt points. The four-dimensional relational statistics comprise the maximum, top-3 mean, mean, and standard deviation. These statistics are computed jointly from the raw cosine similarities across the neighborhoods of all prompt points. The Bank path performs the same relational retrieval over all prototypes and shares the relational fusion module and detection decoder with the Online path.

Prototype initialization. Prototype initialization uses only support images from the training set. We collect candidate features from the 3×3 neighborhoods centered on prompt points in up to 32 support images and retain at most 4096 candidates. After normalizing the candidate features, we select M prototypes using farthest point sampling (Qi et al., 2017) based on cosine distance. The target anchor is initialized from the mean of the target text projections of these support images:

$$\begin{aligned} \mathbf{B}^{(0)} &= \mathrm{FPS}_{\cos}(\{\mathrm{nrm}(\mathbf{c}) : \mathbf{c} \in \mathcal{C}_{\mathrm{cal}}\}, M)\,, \\ \mathbf{a}^{(0)} &= \mathrm{nrm}\left(\frac{1}{N_{\mathrm{cal}}} \sum_{n=1}^{N_{\mathrm{cal}}} \mathbf{t}_n^{+}\right), \end{aligned} \tag{A.2}$$

where nrm denotes the ℓ_2 normalization used in Eq. A.1. $\mathcal{C}_{\mathrm{cal}}$ is the set of candidate features, N_{cal} is the number of support images used for initialization, and $\mathbf{t}_n^{+}$ is the target text projection of the n-th support image. After initialization, both $\mathbf{B}$ and $\mathbf{a}$ are trained.

Retrieval complexity. TCPL reduces retrieval cost by restricting references to support-point neighborhoods during training and to learned prototypes at inference. Let Q be the number of query pixels

and d_r the matching dimension. For each support condition, local retrieval involves only KW^2 reference tokens, yielding $O(QKW^2d_r)$ matching computation and $O(QKW^2)$ score memory. In contrast, dense matching over S support pixels requires $O(QSd_r)$ computation and $O(QS)$ score memory. Thus, local retrieval is substantially cheaper when $KW^2 \ll S$. Reference aggregation further requires $O(QKW^2d_v)$ operations per support condition, where d_v is the value-feature dimension. Hence, the overall training cost scales linearly with the number of support conditions G. At inference, the Bank contains only M learned prototypes, yielding $O(QMd_r)$ matching computation and $O(QM)$ score memory, independent of the support-image size and the number of support conditions.

A.2 Details of the Loss Functions

Eq. 12 in the main text uses class-balanced aggregation. For a per-pixel error $E(u)$ and a binary annotation $Y(u)$, it is defined as:

$$\mathcal{A}_Y(E) = \frac{1}{2}\left[\frac{\sum_u Y(u)E(u)}{\max(\sum_u Y(u), 1)} + \frac{\sum_u [1-Y(u)]E(u)}{\max(\sum_u [1-Y(u)], 1)}\right]. \tag{A.3}$$

The Bernoulli KL divergence is defined as:

$$D_{\mathrm{B}}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, \tag{A.4}$$

where p and q denote the target probabilities of the teacher and student, respectively. Eq. A.3 first computes the mean errors over the foreground and background regions and then sums the two terms with equal weights.

Both the Online path and the Bank path use focal loss and Dice loss. Let $p(u) = \sigma(z(u))$, $p_t(u) = Y(u)p(u) + [1-Y(u)][1-p(u)]$, and $\alpha_t(u) = \alpha_f Y(u) + (1-\alpha_f)[1-Y(u)]$. The focal loss is:

$$\mathcal{L}_{\mathrm{Focal}}(z, Y) = -\frac{1}{N}\sum_u \alpha_t(u)[1-p_t(u)]^{\gamma_f}\log p_t(u). \tag{A.5}$$

The Dice loss is:

$$\mathcal{L}_{\mathrm{Dice}}(z, Y) = 1 - \frac{2\sum_u p(u)Y(u) + \epsilon}{\sum_u p(u) + \sum_u Y(u) + \epsilon}, \tag{A.6}$$

where N is the number of pixels included in the computation, $\alpha_f = 0.75$, and $\gamma_f = 2$. The supervised loss for a single path is $\mathcal{L}_{\mathrm{sup}} = \mathcal{L}_{\mathrm{Focal}} + \mathcal{L}_{\mathrm{Dice}}$. Substituting the supervised losses of the two paths and the distillation loss in Eq. 12 into Eq. 13 yields the final training objective.

A.3 Query-Only Inference

Algo. 2 details inference with the trained parameters held fixed. The encoder first extracts query features from X_q. The normalized target anchor $\mathrm{nrm}(\mathbf{a}^*)$ replaces the projected target text in Eq. 9, producing the guided features $\mathbf{Z}_q^{\mathrm{bank}}$ and target response A_q^{bank}. Each guided query feature then retrieves from the M normalized prototypes using Eq. 10, with softmax applied over the prototype bank.

The maximum, top-3 mean, mean, and standard deviation of the query–prototype cosine similarities form the relational statistics. Following Eq. 7, the fusion module combines their embedding with the query features, retrieved features, and their element-wise absolute differences and products. A projected skip connection from $[\mathbf{Z}_q^{\mathrm{bank}}, A_q^{\mathrm{bank}}]$ is added before decoding. The sigmoid-transformed logits yield the target probability map $\widehat{Y}_q$. All target references are supplied by the learned anchor and prototypes, so inference requires neither support inputs nor text encoding.

B Datasets and Experimental Details

B.1 Dataset Descriptions

All three datasets treat targets of interest as foreground and all other land-cover objects as background. Their statistics are provided in Table 7.

Algorithm 2 Query-only open-scene STD

Input: Unseen query image X_q and trained parameters $(\theta^*, \mathbf{B}^*, \mathbf{a}^*)$.

Output: Pixel-level target probability map $\widehat{Y}_q$.

1: $\mathbf{F}_q \leftarrow \mathcal{E}_{\theta^*}(X_q)$.
2: $(\mathbf{Z}_q^{\mathrm{bank}}, A_q^{\mathrm{bank}}) \leftarrow \mathrm{ATKE}_q(\mathbf{F}_q, \mathrm{nrm}(\mathbf{a}^*))$.
3: Compute query–prototype relations and $\mathbf{R}_q^{\mathrm{bank}}$ using Eq. 10.
4: Extract the four-dimensional relational statistics; reuse the fusion module in Eq. 7 to obtain $\mathbf{F}_q^{\mathrm{rel,bank}}$.
5: $z_q^{\mathrm{bank}} \leftarrow \mathcal{D}_{\theta^*}\big(\mathbf{F}_q^{\mathrm{rel,bank}} + W_{\mathrm{skip}}[\mathbf{Z}_q^{\mathrm{bank}}, A_q^{\mathrm{bank}}]\big)$.
6: $\widehat{Y}_q(u) \leftarrow \sigma(z_q^{\mathrm{bank}}(u))$.
7: **return** $\widehat{Y}_q$.

Table 7: Dataset statistics and training, validation, and test splits.

Dataset	Target	Images	Bands	Dimensions	Train/Val/Test
Aircraft-14	Aircraft	14	192	128×128	9/2/3
Vehicle-74	Vehicles	74	51	256×256	32/10/32
Plane-400	Aircraft	400	8	256×256	264/39/97

Aircraft-14. This dataset is designed for pixel-level aircraft target detection in hyperspectral remote sensing. Its data originate from the AVIRIS airborne hyperspectral imaging spectrometer developed by NASA's Jet Propulsion Laboratory in the United States, and incorporate publicly available ABU-Airport data. During construction, 10 images were cropped from two AVIRIS scenes containing aircraft targets, and 4 airport scenes from the ABU dataset were added, yielding a total of 100 aircraft instances. To prevent data leakage caused by scenes from the same source, the dataset is split by grouping samples according to their original acquisition scenes, ensuring that the training, validation, and test subsets have independent scene sources.

Vehicle-74. This dataset is reconstructed from the AeroRIT airborne hyperspectral data. The original data were collected using a Headwall Photonics Micro Hyperspec E-Series CMOS hyperspectral sensor mounted on an airborne platform and include a hyperspectral flight-line image covering a campus area, with annotations for multiple land-cover classes, including roads, buildings, vegetation, water bodies, and vehicles. We crop the original image into non-overlapping 256×256 patches, retain 51 valid spectral bands within the 400–900 nm range, and binarize the multiclass annotations by treating vehicles as foreground and all other land-cover objects as background, thereby creating a pixel-level vehicle target detection dataset.

Plane-400. This dataset is reconstructed from the RarePlanes multispectral imagery acquired by the Maxar WorldView-3 satellite. The original data provide only bounding-box annotations for aircraft and cannot be used directly for pixel-level segmentation. Therefore, we select and crop the original imagery and manually annotate the aircraft in the cropped regions pixel by pixel to generate instance-level masks, annotating a total of 6472 aircraft instances. The dataset is split by original image to prevent samples from the same source from appearing simultaneously in the training, validation, and test subsets.

B.2 Implementation Details and Hyperparameter Settings

Runtime environment and model configuration. All experiments are run on a single NVIDIA GeForce RTX 4090 GPU. The feature dimension is $d = 16$, and the query–key matching dimension is 8. The default configuration uses $M = 32$ target prototypes, a $W \times W$ neighborhood with $W = 5$ around each support point, and a distillation temperature of $T = 2$. All datasets use $G = 4$ support conditions per training group. The training configurations for each dataset are provided in Table 8.

Table 8: Training configurations for each dataset

Dataset	Training epochs	Batch size
Aircraft-14	100	2
Vehicle-74	100	4
Plane-400	50	16

Input adaptation for comparison methods. HSNet (Min et al., 2021), DCAMA (Shi et al., 2022), IFA (Nie et al., 2024), DCDNet (Cong et al., 2026), FPTrans (Zhang et al., 2022), and MPA (Nie et al., 2026) are adapted to 192, 51, and 8 input channels for Aircraft-14, Vehicle-74, and Plane-400, respectively, using full support images and fixed K-point maps. HSNet, DCAMA, IFA, DCDNet, FPTrans, and TCPL use dense training-query supervision. MPA instead trains on augmented support views with synchronously transformed point maps, without dense labels from independent training queries. SelfMTL (Luo et al., 2025) learns embeddings from spectral triplets sampled with dense training masks and scores query pixels by cosine similarity to the embedding of the mean support-point spectrum. TCFSL (Feng et al., 2025) forms source and target episodes from paired training-query and training-support images using spectral–spatial patches. Dense training masks supply source foreground/background and target background samples, while target foreground samples come only from the designated K support points. All methods share the same data splits and support-point sets, with parameters fixed during evaluation. Support-conditioned methods average query score maps over eligible training supports, while TCPL uses only its learned Bank.

Support-point sampling and inference. Both the standard K-point experiments and the strict point-budget study use nested support-point sets that remain fixed throughout training. For each support image eligible at all three point counts, the sets satisfy $P_1 \subset P_3 \subset P_5$. In the standard experiments, support eligibility depends on K. The strict study fixes the support pool, the pairs of support and query images, text descriptions, and evaluation queries across K to isolate the effect of point count. At test time, support-conditioned comparison methods use training-set support images with their fixed support points.

B.3 Additional Quantitative and Qualitative Comparisons

Table 9 reports quantitative comparisons on Plane-400, derived from RarePlanes imagery. Fig. 6 provides additional qualitative comparisons of target localization and background suppression across methods. TCPL achieves the highest mean AP and AUC at all three support-point budgets on Plane-400. Its AP exceeds the strongest baseline by 4.12, 3.27, and 4.29 percentage points with 1, 3, and 5 points, respectively. Fig. 6 shows localized target responses on Aircraft-14 and Vehicle-74, whereas the Plane-400 responses retain more activation in complex backgrounds. Plane-400 has only eight spectral bands, which may limit target–background discrimination and make these background regions harder to suppress.

Table 9: Plane-400 performance across support-point budgets (%, mean ± std over three seeds).

Method	1-point		3-point		5-point	
	AP ↑	AUC ↑	AP ↑	AUC ↑	AP ↑	AUC ↑
HSNet (Min et al., 2021)	22.33 ± 2.63	95.17 ± 0.53	20.72 ± 1.21	95.00 ± 0.35	19.76 ± 0.92	95.09 ± 0.48
DCAMA (Shi et al., 2022)	32.47 ± 1.02	95.49 ± 0.56	30.87 ± 1.35	95.66 ± 0.23	31.74 ± 0.52	95.60 ± 0.27
IFA (Nie et al., 2024)	20.56 ± 0.61	94.34 ± 0.17	21.01 ± 0.13	94.53 ± 0.28	20.51 ± 0.14	94.46 ± 0.00
DCDNet (Cong et al., 2026)	18.54 ± 0.13	92.58 ± 0.41	18.45 ± 0.52	93.44 ± 1.08	17.99 ± 0.99	94.78 ± 0.33
FPTrans (Zhang et al., 2022)	16.47 ± 0.07	91.73 ± 0.05	15.07 ± 0.80	90.58 ± 0.17	12.22 ± 2.40	91.11 ± 0.45
MPA (Nie et al., 2026)	9.77 ± 0.67	90.77 ± 2.08	14.20 ± 0.31	94.35 ± 0.60	16.00 ± 0.17	92.76 ± 2.19
SelfMTL (Luo et al., 2025)	14.99 ± 1.84	90.00 ± 5.02	15.95 ± 3.52	90.53 ± 1.69	15.50 ± 3.69	90.47 ± 1.58
TCFSL (Feng et al., 2025)	28.27 ± 9.82	92.85 ± 2.18	33.06 ± 8.60	71.99 ± 6.23	31.47 ± 5.53	68.31 ± 4.92
TCPL (ours)	**36.59 ± 2.27**	**96.79 ± 0.30**	**36.33 ± 3.50**	**96.60 ± 0.85**	**36.03 ± 3.33**	**96.64 ± 0.85**

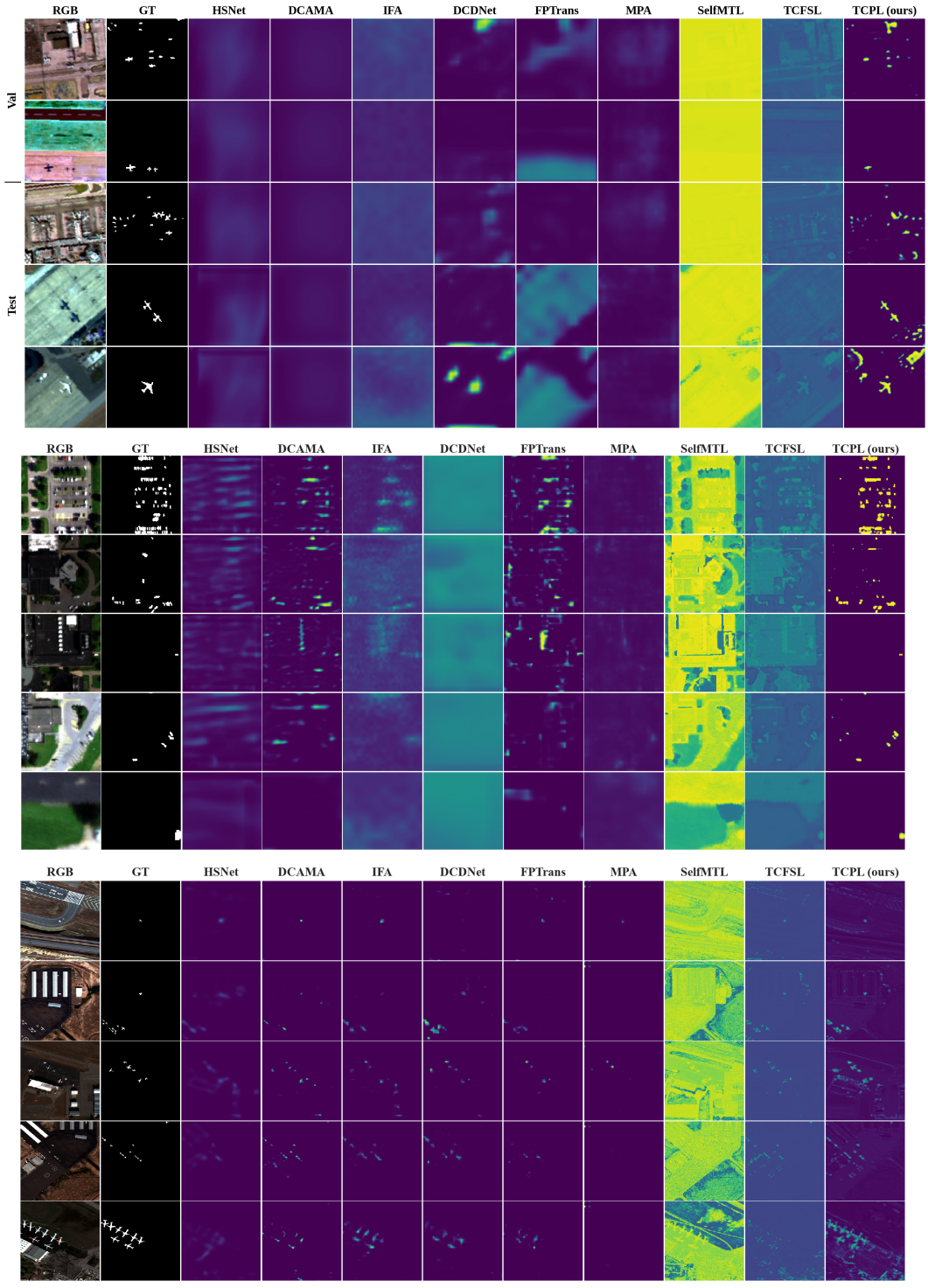


Figure 6: Additional qualitative comparisons on Aircraft-14, Vehicle-74, and Plane-400.

B.4 Additional Ablation Details

Table 10 summarizes the variants evaluated in Table 3, using the same IDs. All variants share the spectral–spatial encoder, detection decoder, and supervised losses. Bank guidance uses the learned target anchor to modulate query features.

Table 10: Component configurations of the ablation variants.

ID	Variant	Components				Retrieval and distillation	
		Text	ATKE	CPLD	Bank guidance	Retrieval	KD teacher
1	Base	–	–	–	–	Mean pooling	–
2	Base + ATKE	✓	✓	–	–	Mean pooling	–
3	w/o ATKE	–	–	✓	–	Attention	Mean teacher
4	w/o CPLD	✓	✓	–	–	Attention	–
5	w/o TCSV	✓	✓	✓	✓	Mean pooling	Single teacher
6	**Full TCPL**	✓	✓	✓	✓	**Attention**	**Mean teacher**

Component ablations. Base averages features within support-point neighborhoods and across points. Base + ATKE adds text guidance. Without ATKE, attention retrieval and prototype distillation remain, but text and target-anchor guidance are omitted. Removing CPLD excludes prototype learning, Bank guidance, and KD. Removing TCSV replaces attention retrieval and multi-support consensus with mean pooling and a single teacher. Removing ATKE reduces AP by 7.46/8.62 percentage points on Aircraft-14/Vehicle-74, supporting semantic guidance within the full framework. Removing CPLD lowers AP by 10.16/6.80 percentage points, demonstrating the collective value of prototype-based learning and distillation. Compared with mean pooling and a single teacher, TCSV improves Vehicle-74 AP/AUC by 7.12/1.49 percentage points and Aircraft-14 AP by 0.95 percentage points, with similar Aircraft-14 AUC (97.69% versus 97.76%). These results favor attention retrieval with multi-support consensus in mean AP. Full TCPL achieves the highest mean AP on both datasets, while removing any component lowers AP, supporting the design for target detection.

Text guidance and distillation. Table 4 compares three configurations. Query-only uses densely annotated queries without support points or text, randomly initializing prototypes and a target anchor. Query+Text guides features with a projected concatenated target–background prompt. Dual-text separately encodes both descriptions to guide support features, using only target semantics for query features. Both retain support-point retrieval, TCSV, and CPLD, with prototypes initialized from support-point neighborhood features. Query+Text and Dual-text initialize anchors from mean projected combined-text and target-text embeddings, respectively. Dual-text achieves the highest AP on both datasets, although its Vehicle-74 AUC is lower than Query+Text's (93.73% versus 94.25%), indicating metric-dependent benefits. KD improves AP/AUC by 0.50/1.13 percentage points on Aircraft-14 and 4.94/0.96 on Vehicle-74 (Table 5), supporting consensus distillation beyond dense detection supervision, particularly for Vehicle-74 AP. KD improves mean performance but not consistency across seeds in these experiments, as AP standard deviations increase from 1.69 to 4.80 and from 0.50 to 1.50, respectively.

B.5 Additional Qualitative Analysis

San Diego 1 and San Diego 2 are hyperspectral airport scenes containing aircraft targets. They were acquired by the same AVIRIS sensor as Aircraft-14 but originate from different acquisition scenes. Aircraft-14 contains 14 images with 192 retained bands and a scene-based 9/2/3 training, validation, and test split (App. B.1). The two San Diego scenes are outside this split and are used only for additional testing, without participation in training or validation. These scenes evaluate same-category transfer under changes in acquisition conditions.

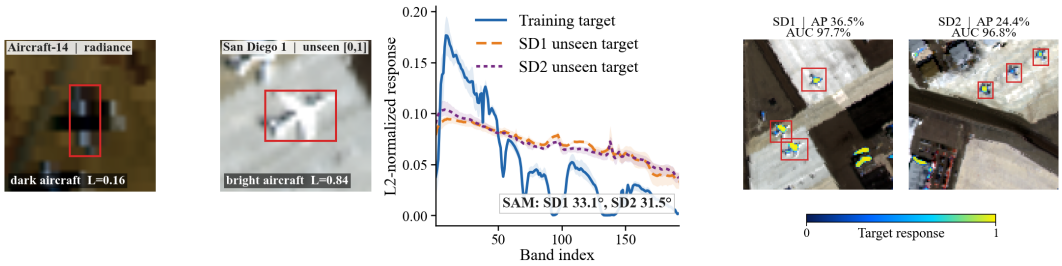
Fig. 7 compares a dark aircraft from the Aircraft-14 training set with bright aircraft in San Diego 1 and San Diego 2. Their spectral angles relative to the training example are 33.1° and 31.5°, respectively. Thus, familiarity with the aircraft category does not imply that the training data cover

the spectral signatures and visual appearances encountered in these test scenes. The trained Bank detects aircraft in both scenes without support inputs or fine-tuning.

We measure spectral differences using the spectral angle mapper (SAM). For two nonzero spectra $\mathbf{x}$ and $\mathbf{y}$,

$$\mathrm{SAM}(\mathbf{x}, \mathbf{y}) = \frac{180}{\pi} \arccos\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} \right). \tag{B.1}$$

Larger angles indicate greater spectral disagreement. The query-only predictions use the trained Bank without test-scene adaptation or support inputs.



(a) Dark train aircraft (b) Bright unseen aircraft (c) Cross-domain spectral shift (d) Query-only open-scene detection

Figure 7: Open-scene aircraft detection under cross-scene spectral variation. (a,b) A dark aircraft from a training scene and a bright aircraft from an unseen San Diego scene. (c) Their target spectra exhibit a large discrepancy (SAM: 33.1°). (d) The query-only model detects aircraft in both San Diego scenes without support inputs or test-time parameter updates.

Prediction with test-query point references. During training, TCPL learns to predict query targets by matching query features to references from support-point neighborhoods. Table 11 examines this reference-based prediction mechanism at inference by replacing the learned prototype bank with features derived from L target points in the test query. These points provide support references for predicting the same image. This additional setting requires test-time point annotations, whereas default query-only inference uses the learned Bank.

Table 11: Replacing learned prototypes with test-query point references.

Dataset	L	AP (%) ↑		AUC (%) ↑	
		Bank	Point references	Bank	Point references
Aircraft-14	1	42.96	39.86	97.69	97.31
Aircraft-14	3	42.96	39.92	97.69	97.36
Aircraft-14	5	42.96	40.07	97.69	97.42
Vehicle-74	1	62.52	61.97	93.73	93.26
Vehicle-74	3	62.52	64.25	93.73	93.33
Vehicle-74	5	62.52	66.29	93.73	93.58

Using test-query point references instead of Bank prototypes modestly reduces Aircraft-14 AP/AUC by 2.89–3.10/0.27–0.38 pp. On Vehicle-74, using 1 point lowers AP/AUC by only 0.55/0.47 pp. With 3 and 5 points, AP reaches 64.25% and 66.29%, respectively, while AUC remains within 0.40 pp of the Bank. These results indicate that TCPL retains effective reference-based prediction with few annotated points, even without learned prototypes as retrieval references. The support–query matching learned during training thus remains useful when references are supplied by the test image.