

Article

PromptScaleDINO: Prompt-Stabilized and Scale-Aware Adaptation of Grounding DINO for Infrared Small Target Detection

Chichi Huang ^{1,2,3} , Zefang Wang ^{1,2,3}, Yuanjun Chen ^{1,2,3}, Junqi Ji ^{1,2,3}, Yi Shen ^{1,2,3}, Changqing Lin ^{1,3} and Gaorui Liu ^{1,3,*}

- ¹ National Key Laboratory of Infrared Detection Technologies, Shanghai Institute of Technical Physics, Chinese Academy of Sciences, 500 Yutian Road, Shanghai 200083, China; huangchichi24@mailsucas.ac.cn (C.H.); wangzefang24@mailsucas.ac.cn (Z.W.); chenyanjun24@mailsucas.ac.cn (Y.C.); jijunqi23@mailsucas.ac.cn (J.J.); shenyi22@mailsucas.ac.cn (Y.S.); linchangqing@mail.sitp.ac.cn (C.L.)
- ² University of Chinese Academy of Sciences, Beijing 100049, China
- ³ Shanghai Institute of Technical Physics, Chinese Academy of Sciences, 500 Yutian Road, Shanghai 200083, China
- * Correspondence: liugaorui@mail.sitp.ac.cn

Highlights

What are the main findings?

- A lightweight PromptScaleDINO framework is proposed to adapt Grounding DINO for infrared small target detection by integrating anchor-centered prompt stabilization, scale-aware query refinement, and geometry-aware localization supervision.
- By combining textual context with task-specific query and localization adaptation, the proposed method improves detection robustness for weak, few-pixel infrared targets under cluttered remote-sensing backgrounds.

What is the implication of the main finding?

- The study shows that vision-language detectors can be effectively transferred to infrared small target detection without redesigning the full detection architecture.
- The proposed framework provides a practical foundation for language-assisted infrared remote-sensing detection, especially in scenarios where target appearance is weak, sparse, and easily confused with background clutter.

Abstract

Infrared small target detection (IRSTD) supports remote-sensing surveillance and early warning, but infrared targets often occupy only a few pixels, exhibit weak appearance cues, and resemble thermal clutter. Vision-language detectors such as Grounding DINO provide a flexible prompt-driven detection interface, but direct transfer toIRSTD faces three mismatches: a single prompt cannot describe target appearance diversity, generic decoder queries are not calibrated for weak scale-sensitive targets, and IoU-style localization losses are unstable for few-pixel boxes. We propose PromptScaleDINO as a lightweight adaptation framework for Grounding DINO. It introduces an Anchor Prompt Bank (APB) to enrich the text input while keeping supervision connected to a stable anchor phrase, a Scale-Aware Query Refinement Network (SQRN) to refine selected queries according to predicted box scale and semantic confidence, and a geometry-aware localization objective that combines Normalized Wasserstein Distance (NWD) with Generalized Intersection over Union (GIoU). Under a unified detection protocol, PromptScaleDINO achieves its largest gains on the challenging real-sceneIRSTD-1k benchmark while maintaining near-ceiling performance on SIRST and NUDT-SIRST. OnIRSTD-1k, it reaches 88.32% F1 and 87.40% mAP@0.5,



Academic Editor: Zhenming Peng

Received: 3 June 2026

Revised: 17 July 2026

Accepted: 21 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

improving the baseline by 2.56 and 4.00 percentage points, respectively. These gains mainly reflect improved weak-target recovery at comparable precision; accordingly, we position the framework as a target-sensitivity and localization adaptation rather than an explicit false-alarm suppression method.

Keywords: infrared remote sensing; infrared small target detection; vision-language detection; anchor prompt bank; scale-aware query refinement

1. Introduction

Infrared small target detection (IRSTD) aims to localize sparse, weak, and low-resolution targets in infrared remote-sensing imagery. It supports applications such as infrared search and tracking, maritime monitoring, and early warning, where a valid target may occupy only a few pixels and appear as a dim point, compact bright spot, or weak response embedded in clutter. Unlike common visible-image detection,IRSTD depends less on rich object semantics and more on preserving weak signal evidence under strong background interference. Clouds, waves, buildings, sensor noise, and other high-temperature structures can produce target-like responses, making the task sensitive to local contrast, target scale, background clutter, and one-pixel localization errors.

ExistingIRSTD methods have progressively shifted from handcrafted priors of target-background differences to learned contextual representations. Classical approaches characterize weak targets through low-rank and sparse decomposition, local contrast, multiscale patch contrast, or tensor nuclear norm formulations [1–5]. Although these methods are interpretable, their fixed assumptions can be unreliable when target appearances and background structures vary substantially. Deep models, including MDvsFA, ALCNet, ACM, ISNet, DNANet, and EFLNet, learn stronger contextual and multiscale representations for suppressing clutter and preserving weak targets [6–11]. Subsequent methods further improve pyramid context aggregation, feature propagation, attention mechanisms, and long-range dependency modeling [12–19]. These advances establish target scale, background context, and localization stability as central concerns inIRSTD. Nevertheless, most existing methods remain visually driven and do not explicitly use language to stabilize the target concept.

Vision-language foundation models provide a complementary route for introducing semantic priors intoIRSTD. CLIP demonstrates that natural language can define transferable visual concepts through large-scale image-text pretraining [20]. GLIP, MDETR, and Grounding DINO further connect language representations with phrase- or object-level localization [21–23]. Open-vocabulary detectors such as ViLD, OWL-ViT, and YOLO-World support flexible text-conditioned category specification [24–26], while caption supervision, dictionary-enriched pretraining, object-language alignment, and learned detection prompts further strengthen cross-modal localization [27–30]. Recent remote-sensing variants address the aerial-domain gap through large-scale data construction, visual-guided prompt learning, multiscale feature modeling, and language-guided query selection [31,32]. However, these methods mainly target multi-category open-vocabulary recognition, whereasIRSTD requires recovery and precise box localization of weak, few-pixel instances from a single target class. Consequently, direct transfer of a general-purpose vision-language detector may not preserve the evidence required for few-pixel targets.

Recent language-guidedIRSTD studies have begun to demonstrate the value of semantic information in this setting. Text-IRSTD introduces fuzzy semantic prompts and cross-modal interaction to distinguish weak targets from complex backgrounds [33]. MIR-

SAM and SAIST combine vision-language representations with SAM-based segmentation to improve semantic discrimination and target delineation [34,35]. LGNet and DGSPNet construct image-related semantic guidance through generated descriptions or image-to-text prompt inversion [36,37]. Collectively, these studies show that language can complement low-level infrared appearance cues and improve target-background discrimination. Their methods, however, are primarily developed for pixel-level segmentation and integrate semantic guidance within segmentation-oriented pipelines. They therefore leave open how a pretrained grounded detector should maintain stable one-class text supervision while producing reliable bounding boxes for few-pixel targets.

Adapting Grounding DINO to this setting involves three related mismatches. First, the text interface must provide sufficient semantic context without introducing prompt drift or inconsistent supervision between training and inference. Second, decoder queries designed for general objects may respond to clutter or overlook targets whose visual evidence is locally fragile. Third, conventional IoU-based objectives are sensitive to small coordinate shifts when a bounding box contains only a few pixels. These mismatches require coordinated adaptation of the text pathway, decoder queries, and localization objective while preserving the pretrained detector architecture.

To address these challenges, we propose PromptScaleDINO, a prompt-stabilized and scale-aware adaptation framework for Grounding DINO. Rather than replacing the original detection pipeline, the framework introduces three task-oriented components that correspond to the preceding mismatches. The Anchor Prompt Bank (APB) organizes training prompts around a stable class anchor and retains only the anchor phrase during inference. The Scale-Aware Query Refinement Network (SQRN) refines selected decoder queries using predicted scale cues and semantic confidence. A geometry-aware localization objective combines Normalized Wasserstein Distance (NWD) with GIoU and calibrates the NWD normalization constant for infrared small-target scales [38,39]. Together, these components adapt the text, query, and localization pathways to the weak semantics and small spatial extent of infrared targets.

Experiments onIRSTD-1k, SIRST, and NUDT-SIRST show the clearest gains onIRSTD-1k, where PromptScaleDINO reaches 88.32% F1 and 87.40% mAP@0.5, improving the baseline by 2.56 and 4.00 percentage points, respectively. These gains mainly arise from higher recall at comparable precision, while the ablations show complementary contributions from APB, SQRN, and the geometry-aware objective.

The main contributions of this work are summarized as follows.

- We formulate language-assistedIRSTD as a joint adaptation problem across the text pathway, query pathway, and localization objective, reflecting the task's weak semantics, tiny targets, and clutter sensitivity.
- APB provides an anchor-centered prompt grouping strategy for the text pathway. It enriches the training-time textual context while keeping supervision on a stable class anchor and preserving a single-anchor inference protocol.
- SQRN strengthens decoder queries through scale-aware refinement. It uses predicted box scale cues and semantic confidence to guide lightweight multiscale query updates for weak infrared small targets.
- The geometry-aware tiny-box localization objective combines NWD with GIoU and calibrates the normalization constant forIRSTD box scales, providing smoother optimization for few-pixel boxes where IoU-style losses are inherently unstable.

2. Materials and Methods

2.1. Overall Framework

PromptScaleDINO adapts three parts of the Grounding DINO pipeline: the text input, decoder query update, and localization objective. The visual backbone, text encoder, multimodal transformer, and detection heads retain their roles in the original grounded detection framework [22]. This design preserves the pretrained detector while isolating the task-specific adaptations evaluated in the ablation studies.

Given an infrared image I , Grounding DINO extracts visual features and encodes the input text into language features [22]. The multimodal transformer then produces decoder queries for classification and box regression, following the general DETR-style query-decoding paradigm [40–42]. PromptScaleDINO introduces two architectural adaptations and one objective-level refinement. During training, APB replaces the isolated class phrase with an anchor-centered prompt group; during inference, it retains only the anchor phrase. SQRN refines selected decoder queries using predicted box scale and semantic confidence before the final prediction heads. The model also adds a geometry-aware NWD loss for tiny boxes during training [38]. Figure 1 illustrates the overall PromptScaleDINO framework.

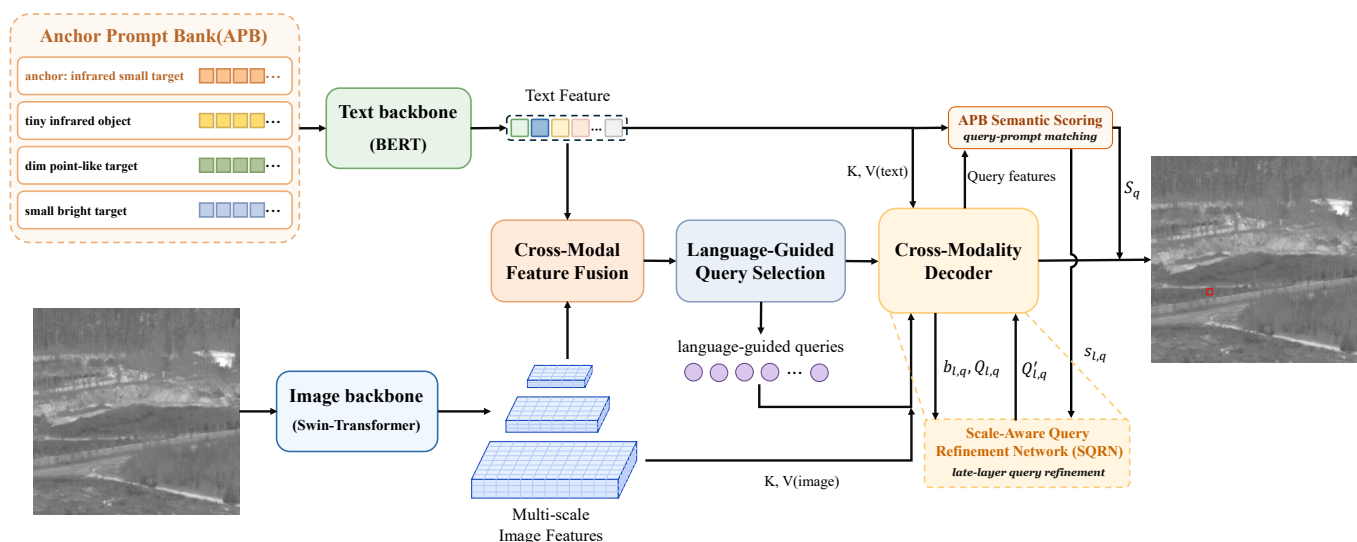


Figure 1. Overall framework of PromptScaleDINO.

2.2. Anchor Prompt Bank

TheIRSTD category is semantically narrow but visually diverse. A true target may be dim, bright, compact, blurred, or partially buried in clutter, yet all these cases belong to the same class. A single prompt such as “infrared small target” provides stable class identity, but it does not explicitly cover this appearance diversity. Treating many prompt variants as independent labels would create a different problem: the class boundary becomes blurred, and supervision may no longer match the one-class evaluation setting. APB is designed to enrich textual context while keeping the supervised class identity fixed.

For the single-classIRSTD setting, APB uses a compact four-prompt bank during training. The anchor phrase is “infrared small target”, and the three descriptive variants are “tiny thermal target”, “dim infrared target”, and “few-pixel bright target”. The bank is manually constructed under four constraints. First, class consistency requires every phrase to describe the same positiveIRSTD category rather than introduce a new category or a background concept. Second, attribute coverage uses the variants to represent complementary target properties, namely small-scale and thermal appearance, weak infrared intensity, and a few-pixel bright response. Third, compactness limits redundant wording that could

dilute the stable class identity. Fourth, dataset-agnostic wording excludes dataset names, sensor identifiers, and scene-specific background terms. The same fixed bank is used across datasets; the variants are not image-conditioned or optimized separately for each dataset.

Let $G = \{p_0, p_1, \dots, p_K\}$ denote a prompt group, where p_0 is the anchor phrase and $p_1, \dots, p_K$ are prompt variants. During training, the caption contains the full prompt group, but the ground-truth class label is mapped only to p_0 . If $E_t(\cdot)$ denotes the text encoder, the prompt group is encoded as

$$T_G = [E_t(p_0), E_t(p_1), \dots, E_t(p_K)]. \quad (1)$$

For a matched object query q , APB uses anchor-only supervision: the positive semantic target is attached only to the anchor phrase, while the variants remain contextual prompt descriptions rather than independent target categories. This can be summarized as

$$y_q^{\text{APB}} = p_0, \quad L_{\text{APB}}(q) = L_{\text{cls}}(q, p_0; G), \quad (2)$$

where $L_{\text{cls}}(q, p_0; G)$ denotes the standard text-conditioned classification loss for query q with the full prompt group G encoded as text context and p_0 used as the supervised class phrase. The variants therefore receive no independent positive class labels and do not create pseudo classes. Instead, they are jointly encoded with the anchor as training-time textual context while the target identity remains tied to p_0 .

$$G_{\text{test}} = \{p_0\}, \quad (3)$$

At inference, the variants are removed and only the anchor phrase “infrared small target” is supplied to the detector. If $a_{q,0}$ is the anchor logit for query q , the final class confidence is

$$S_q = s_{q,0} = \sigma(a_{q,0}). \quad (4)$$

No maximum, mean, or weighted prompt-score aggregation is performed at inference. This protocol preserves a fixed class identity and introduces no prompt-bank overhead during deployment. Figure 2 summarizes the revised APB workflow.

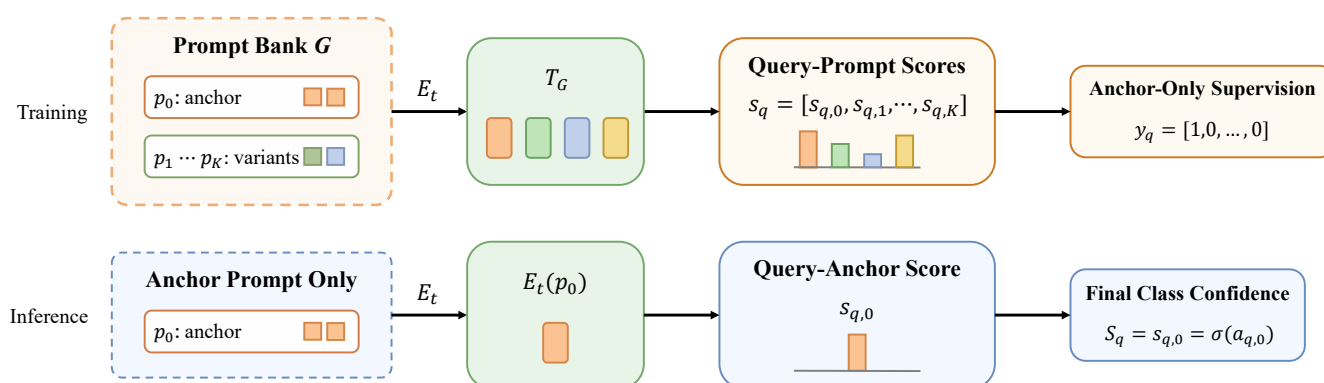


Figure 2. Training and inference workflow of the Anchor Prompt Bank (APB).

Unlike learned and conditional prompt optimization for adapting pretrained vision–language models [43,44], APB is a training-time prompt-context stabilization strategy that preserves a stable class target and the original single-anchor inference interface, rather than an inference-time prompt ensemble.

2.3. Scale-Aware Query Refinement Network

Grounding DINO decoder queries are designed for general object detection; related detectors have explored dynamic attention and learnable proposal refinement [45,46], whereas IRSTD depends on weak responses from extremely small targets. In cluttered infrared scenes, target-related and clutter-related queries may have similar confidence, and small localization errors can strongly affect evaluation. Query refinement should therefore be selective: it should strengthen queries likely to represent tiny targets without disturbing queries that are already reliable or clearly unrelated to targets. SQRN provides this scale- and confidence-aware refinement.

SQRN is inserted into selected decoder layers and operates on query features together with their current normalized box references. For each query, a four-dimensional box descriptor is mapped to query-specific weights over three one-dimensional convolution branches. The weighted branch outputs occupy complementary channel subsets, are concatenated, and are then fused into a single refinement term.

Let $Q_l \in \mathbb{R}^{N \times d}$ denote the query features at decoder layer l , where N is the number of queries and d is the feature dimension. For query q , let $b_{l,q} = (c_x, c_y, w, h)$ be its normalized box reference. Width and height are clamped to at least 10^{-6} , and the scale descriptor used by the implementation is

$$z_{l,q} = [w, h, wh, w/h]. \quad (5)$$

The box descriptor is detached before it enters a two-layer scale estimator. In the main setting, the estimator has dimensions $4 \rightarrow 64 \rightarrow 3$, uses ReLU between the two linear layers, and predicts normalized branch weights:

$$\omega_{l,q} = \text{softmax}(\text{MLP}_s(z_{l,q})), \quad \omega_{l,q} \in \mathbb{R}^3. \quad (6)$$

Let $\phi_m(Q_l)_q \in \mathbb{R}^{d_m}$ denote the output of refinement branch m for query q , where $\sum_m d_m = d$. The three branches use one-dimensional convolution kernels of sizes 3, 5, and 7. Their weighted channel-wise outputs are concatenated and fused as

$$\tilde{R}_{l,q} = \text{Concat}_{m=1}^3 [\omega_{l,q,m} \phi_m(Q_l)_q], \quad R_{l,q} = \text{MLP}_f(\tilde{R}_{l,q}), \quad (7)$$

where MLP_f is a $d \rightarrow d \rightarrow d$ fusion network with GELU activation. For $d = 256$, the three branch output dimensions are 86, 85, and 85, respectively.

SQRN further uses a semantic-confidence gate to control refinement strength. Let $s_{l,q}$ denote the anchor-phrase semantic confidence of query q at decoder layer l , and let $\bar{z}_{l,q} = \sqrt{wh}$ be the geometric-mean side of its box reference. Both the box reference and semantic confidence are detached when computing the scale weights and gate. The gate is a fixed-parameter product of a small-box term and a low-confidence term:

$$g_{l,q} = g_{\min} + (1 - g_{\min}) \sigma\left(\frac{\tau_z - \bar{z}_{l,q}}{\gamma_z}\right) \sigma\left(\frac{\tau_s - s_{l,q}}{\gamma_s}\right), \quad (8)$$

where σ is the logistic function, (τ_z, γ_z) and (τ_s, γ_s) are the center and temperature of the small-box and confidence terms, respectively, and $g_{\min}$ is a floor that guarantees a minimum refinement strength. Because both terms take the form $\sigma((\tau - \cdot)/\gamma)$, the gate is monotonically decreasing in both predicted box size and semantic confidence: small, low-confidence queries—the weak targets that SQRN is designed to rescue—receive the strongest refinement, whereas large or already-confident queries are gated down toward $g_{\min}$.

The updated query is then

$$Q'_{l,q} = \text{LN}(Q_{l,q} + \alpha_l g_{l,q} R_{l,q}), \quad (9)$$

where $LN(\cdot)$ is layer normalization and α_l is a layer-wise residual coefficient. The residual update preserves the original decoder representation and reduces the risk that refinement overwrites useful query information.

SQRN does not apply the same refinement to every query. In the main setting, it operates only in selected late decoder layers, where a semantic-confidence-guided gate controls the update strength. This design focuses refinement on queries most likely to benefit from scale-aware updates. Figure 3 summarizes the resulting workflow.

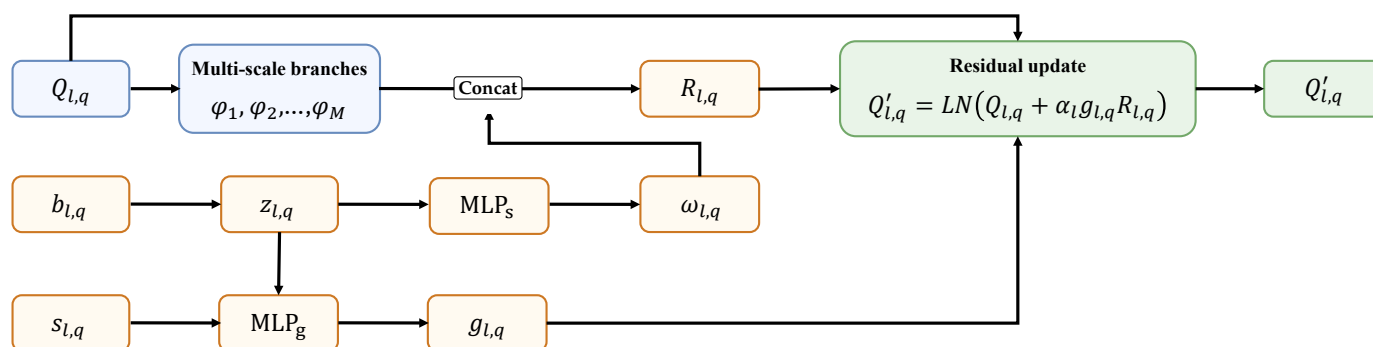


Figure 3. Workflow of the Scale-Aware Query Refinement Network (SQRN).

2.4. Geometry-Aware Tiny-Box Localization Objective

The third adaptation pillar of PromptScaleDINO targets the localization objective. Grounding DINO inherits the standard L1 + GIoU regression objective from DETR-style detectors [39,40]. Related IoU-family objectives can additionally incorporate center-distance geometry to improve general box regression [47]. Both terms become unstable for IRSTD: when a target box spans only a few pixels, a one-pixel center shift can cause a sharp drop in IoU, and L1 distance in absolute pixel coordinates does not reflect the geometric similarity of two near-identical Gaussian-like blobs. As a result, training signals on tiny boxes are noisy and the optimizer cannot reliably distinguish near-correct predictions from clearly wrong ones. We therefore design a geometry-aware tiny-box localization objective that explicitly addresses this instability.

A predicted box $B_p = (c_x^p, c_y^p, w_p, h_p)$ and a ground-truth box $B_g = (c_x^g, c_y^g, w_g, h_g)$ are first represented as two Gaussian distributions. Following the common NWD formulation for bounding boxes [38], their squared Wasserstein distance can be written as

$$W_2^2(B_p, B_g) = (c_x^p - c_x^g)^2 + (c_y^p - c_y^g)^2 + \frac{(w_p - w_g)^2 + (h_p - h_g)^2}{4}. \quad (10)$$

The normalized Wasserstein similarity is then defined as

$$\text{NWD}(B_p, B_g) = \exp\left(-\frac{\sqrt{W_2^2(B_p, B_g)}}{C}\right), \quad (11)$$

where C is a normalization constant. The corresponding loss is

$$L_{\text{NWD}} = 1 - \text{NWD}(B_p, B_g). \quad (12)$$

The overall training objective keeps the standard Grounding DINO classification and regression losses and adds NWD to the localization part:

$$L = L_{\text{cls}} + \lambda_{\text{L1}} L_{\text{L1}} + \lambda_{\text{GIoU}} L_{\text{GIoU}} + \lambda_{\text{NWD}} L_{\text{NWD}}. \quad (13)$$

Here, L_{cls} is the classification loss, L_{L1} and L_{GIoU} are the original box regression terms [39], and L_{NWD} measures box distance in a normalized Gaussian space [38]. Compared with IoU, NWD changes more smoothly when a tiny predicted box undergoes a small absolute displacement, which makes the optimization landscape better conditioned for few-pixel boxes.

A practical difficulty in applying NWD to IRSTD is that the normalization constant C controls the effective sensitivity range of the loss. A constant transferred from generic tiny-object detection does not match the box-size statistics of infrared small targets: too large a C flattens the loss and weakens supervision, while too small a C saturates the loss for moderate displacements. We therefore calibrate C for IRSTD box scales rather than reusing a default value. The candidate grid is centered on the default $C = 12.8$ and spans $0.5\times$ to $2\times$ that value, with intermediate multipliers of $0.75\times$ and $1.25\times$, thereby probing both faster and slower NWD similarity decay around the default. Specifically, we evaluate candidate constants $C \in \{6.4, 9.6, 12.8, 16.0, 25.6\}$ on the baseline validation setting and select $C = 25.6$ according to the COCO-style validation AP. Unless otherwise specified, $C = 25.6$ and $\lambda_{NWD} = 10.0$ are used in NWD-enabled PromptScaleDINO experiments so that the geometry-aware term provides a meaningful gradient alongside L1 and GIoU. The empirical effect of this calibration is reported in Section 3.6.4. Figure 4 visualizes the difference between IoU and NWD under center displacement.

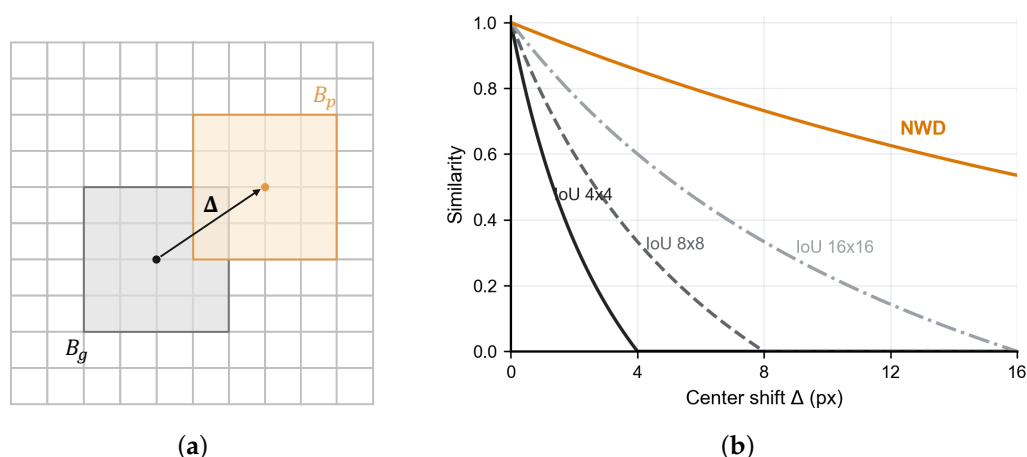


Figure 4. Comparison of IoU and NWD sensitivity to tiny-box center shifts: (a) center displacement between a predicted and ground-truth box; (b) similarity responses under center shift.

3. Results

3.1. Datasets

Following the benchmark split used in prior detection-based IRSTD studies [11,48], we evaluate on three publicly used infrared small target benchmarks and divide each dataset into training, validation, and test sets with an approximately 6:2:2 ratio. We use bounding-box annotations for detection evaluation; when the original benchmark provides pixel-level masks, the corresponding detection boxes are obtained from the target regions.

IRSTD-1k is a real infrared small target dataset containing approximately one thousand 512×512 images with pixel-level annotations and diverse scenes and target appearances [9]. We use it as the primary benchmark for the main comparison and ablation studies. SIRST, referred to as NUAA-SIRST in some prior detection studies, is a real single-frame infrared small target dataset constructed from representative frames selected from infrared sequences [8]. The dataset provides semantic segmentation and bounding-box annotations, and includes targets that are often dim, sparse, and embedded in cluttered backgrounds. We use SIRST as a complementary real-data benchmark for testing whether PromptScaleDINO

generalizes beyond IRSTD-1k. NUDT-SIRST is a synthetic infrared small target dataset containing 1327 images with a resolution of 256×256 [10]. Its main advantage is that it provides accurate annotations, controllable generation, rich target sizes, and diverse clutter backgrounds. We include NUDT-SIRST to evaluate robustness under a distribution that differs from the two real-image benchmarks. Table 1 summarizes the detection-format splits used in our experiments. The box counts refer to the COCO-style detection annotations used by our evaluation pipeline.

Table 1. Dataset statistics and train/validation/test splits used in this work.

Dataset	Source Type	Native Image Size	Train Images/Boxes	Val Images/Boxes	Test Images/Boxes	Total Images/Boxes
IRSTD-1k	Real	512×512	600/882	200/294	201/302	1001/1478
SIRST	Real	Variable	256/327	85/97	86/109	427/533
NUDT-SIRST	Synthetic	256×256	796/1110	265/371	266/380	1327/1861

3.2. Evaluation Metrics

Following the detection-style evaluation protocol of EFLNet [11], we report Precision, Recall, F1, and AP-based detection metrics as the main quantitative indicators. Our evaluator adapts the released detection-evaluation implementation to the Grounding DINO prediction format. Standard detection metrics use one-to-one class-consistent IoU matching. For a predicted box B_p and a ground-truth box B_g , IoU is defined as

$$\text{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}. \quad (14)$$

A prediction is counted as a true positive (TP) at IoU threshold τ if it is the highest-scoring unmatched prediction assigned to an unmatched ground-truth target of the same class and satisfies $\text{IoU}(B_p, B_g) \geq \tau$. Remaining predictions are false positives (FPs), and unmatched ground-truth boxes are false negatives (FNs). Precision, Recall, and F1 are computed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (15)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (16)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (17)$$

Average precision (AP) is computed from the precision-recall curve after sorting predictions by confidence. Following the evaluation code, the precision envelope is integrated using 101-point interpolation:

$$\text{AP}_\tau = \int_0^1 p_{\text{interp}}(r; \tau) dr. \quad (18)$$

We report mAP@0.5, which is AP at $\tau = 0.5$, and mAP@0.5 : 0.95, which averages AP over ten IoU thresholds $\tau \in \{0.50, 0.55, \dots, 0.95\}$:

$$\text{mAP@0.5 : 0.95} = \frac{1}{10} \sum_{\tau \in \{0.50, 0.55, \dots, 0.95\}} \text{AP}_\tau. \quad (19)$$

For threshold-dependent detection metrics generated in this work, including Precision, Recall, and F1, we use a fixed confidence threshold of 0.5. Predictions below this threshold are discarded before one-to-one IoU matching, and the same operating point is used for vanilla Grounding DINO and all PromptScaleDINO variants across the three datasets.

We do not report probability of detection (PD) or false alarm rate (FA) in the main tables because these IRSTD-specific metrics are often defined under pixel-level segmentation, connected-component, image-level, or frame-level protocols. In particular, FA is not directly equivalent to detection false positives or Precision because its normalization unit and target-background matching rule vary across IRSTD evaluation protocols. To avoid mixing incompatible metric definitions, we use Precision, Recall, F1, mAP@0.5, and mAP@0.5 : 0.95 as the primary metrics under the unified bounding-box detection protocol. AP and mAP are threshold-free curve metrics computed from the full ranked prediction list.

3.3. Implementation Details

PromptScaleDINO is implemented on top of Grounding DINO with the Swin-Tiny backbone pretrained at 224-pixel resolution. The text encoder is BERT-base-uncased and remains frozen during training. Unless otherwise specified, training uses the ODVG data pipeline with random horizontal flipping, multiscale resizing, and optional random size cropping. The short side is sampled from $\{480, 512, 544, 576, 608, 640, 672, 704, 736, 768, 800\}$, the maximum long side is 1333 pixels, and the crop range is 384–600 pixels. Validation and test images are resized with a short side of 800 and a maximum long side of 1333.

All models are optimized with AdamW, weight decay 1×10^{-4} , and gradient clipping at 0.1. The base learning rate is 1×10^{-4} , the backbone learning rate is 1×10^{-5} , and the learning rate is decayed after epoch 10. Single-stage baselines and module ablations are trained for 20 epochs. The main distributed runs use a batch size of 8 per GPU process on two NVIDIA V100 GPUs (32 GB each), giving an effective batch size of 16. The other detector parameters remain trainable. Unless otherwise specified, the main tables report the seed-42 run.

The full APB-SQRN model uses two sequential 10-epoch stages, for 20 optimization epochs in total. In Stage 1, the APB model is initialized from the pretrained Grounding DINO Swin-T checkpoint and trained for 10 epochs. Stage 2 is initialized from the validation-best APB checkpoint: matching parameters are inherited, and the new SQRN parameters keep their initialization. Stage 2 is a fresh 10-epoch run with a re-initialized optimizer and scheduler. During Stage 2, all non-frozen detector parameters continue to be fine-tuned, while SQRN uses a learning-rate multiplier of 0.25, corresponding to 2.5×10^{-5} .

APB uses the same compact four-prompt bank described in Section 2.2 and applies anchor-only positive supervision during training. All APB and APB-SQRN models use only “infrared small target” at inference. The six-layer decoder has 900 queries with dimension $d = 256$. A single shared SQRN module is applied from zero-based decoder index 3, i.e., to decoder layers 4–6, without an additional warm-up period. The scale estimator uses hidden dimension 64; the three convolution kernels are 3, 5, and 7; and the fusion MLP has dimensions $256 \rightarrow 256 \rightarrow 256$. The small-box gate uses $\tau_z = 0.08$ and $\gamma_z = 0.03$, the low-confidence gate uses $\tau_s = 0.60$ and $\gamma_s = 0.10$, the floor is $g_{\min} = 0.05$, and the residual coefficient is 0.35. NWD-enabled runs set $C = 25.6$ and $\lambda_{\text{NWD}} = 10.0$.

3.4. Comparison with Various Methods

Table 2 compares PromptScaleDINO with representative IRSTD methods and vanilla Grounding DINO. External IRSTD results are reproduced from the published sources [11,48]; Grounding DINO and PromptScaleDINO use the same partitions in Table 1, COCO-style boxes, preprocessing, one-to-one IoU@0.5 matching, and a fixed 0.5 confidence threshold. All entries are bounding-box results, although external file formats and mask-to-box conversions were not unified.

The cross-architecture comparison does not control input resolution, model capacity, or pretraining: Refs. [11,48] use 640×640 inputs, whereas our models use a short side of

800 and a maximum long side of 1333; Grounding DINO is also larger and vision-language pretrained. Table 2 is therefore a contextual benchmark comparison. The controlled evidence is vanilla Grounding DINO versus PromptScaleDINO, which share the base detector, initialization, data, preprocessing, and evaluator; later AP results and ablations provide threshold-free evidence.

The results in Table 2 should be interpreted with the saturation level of each benchmark in mind. On IRSTD-1k, PromptScaleDINO reaches 90.31% Precision, 86.42% Recall, and 88.32% F1, improving the vanilla Grounding DINO baseline by 2.56 percentage points in F1 primarily through a 4.63-point Recall gain. Precision changes only slightly, from 90.14% to 90.31%. This pattern indicates that the proposed adaptation mainly recovers missed weak targets rather than acting as an explicit false-alarm filter. On SIRST and NUDT-SIRST, however, the vanilla Grounding DINO baseline already reaches 97.24% and 98.55% F1, respectively; PromptScaleDINO therefore provides only small additional gains and is better viewed as maintaining stable near-ceiling performance on these saturated benchmarks rather than producing a large margin.

Table 2. Comparison with existing IRSTD methods on NUAA-SIRST, NUDT-SIRST, and IRSTD-1k. A dash indicates that the corresponding results were not reported in the original paper.

Method	Year	Venue	NUAA-SIRST (%)			NUDT-SIRST (%)			IRSTD-1k (%)		
			Pre ↑	Rec ↑	F1 ↑	Pre ↑	Rec ↑	F1 ↑	Pre ↑	Rec ↑	F1 ↑
MDvsFA [6]	2019	ICCV	84.50	50.70	59.70	60.80	19.20	26.20	55.00	48.30	47.50
ALCNet [7]	2021	IEEE TGRS	84.80	78.00	81.30	86.80	77.20	81.70	84.30	65.60	73.80
ACM [8]	2021	WACV	76.50	76.20	76.30	73.20	74.50	73.80	67.90	60.50	64.00
ISNet [9]	2022	CVPR	82.00	84.70	83.40	74.20	83.40	78.50	71.80	74.10	72.90
DNANet [10]	2023	IEEE TIP	84.70	83.60	84.10	91.40	88.90	90.10	76.80	72.10	74.40
AGPCNet [12]	2023	IEEE TAES	39.00	81.00	52.70	36.80	68.40	47.90	41.50	47.00	44.10
EFLNet [11]	2024	IEEE TGRS	88.20	85.80	87.00	96.30	93.10	94.70	87.00	81.70	84.30
TY-RIST [48]	2025	ICCVW	92.90	92.10	92.50	96.80	95.80	96.30	-	-	-
Grounding DINO [22]	2024	ECCV	97.25	97.24	97.24	98.68	98.42	98.55	90.14	81.79	85.76
PromptScaleDINO (ours)	2026	-	99.08	99.08	99.08	99.47	98.42	98.94	90.31	86.42	88.32

To complement the literature-reported numerical comparison in Table 2, Figure 5 presents predictions generated for qualitative evaluation on the same displayed test images. ALCNet [7] and DNANet [10] are shown using their segmentation masks, whereas EFLNet [11], Grounding DINO [22], and PromptScaleDINO are shown using bounding boxes. Because these methods produce different spatial representations, we interpret the comparison at the target level, focusing on correct detections, false alarms, and missed targets rather than directly comparing mask-boundary quality with bounding-box localization accuracy.

Figure 5 shows that the compared methods can recover many isolated targets, but their differences become clearer in scenes containing weak targets or locally bright background structures. ALCNet, DNANet, and EFLNet produce false responses to background interference in several examples. In the fourth row, ALCNet, DNANet, EFLNet, and vanilla Grounding DINO miss the weak target, whereas PromptScaleDINO successfully detects it. This additional recovery is consistent with the recall-oriented improvement over Grounding DINO reported in Table 2. These examples provide qualitative illustrations of target recovery and background interference rather than an independent ranking of segmentation-boundary and bounding-box localization quality.

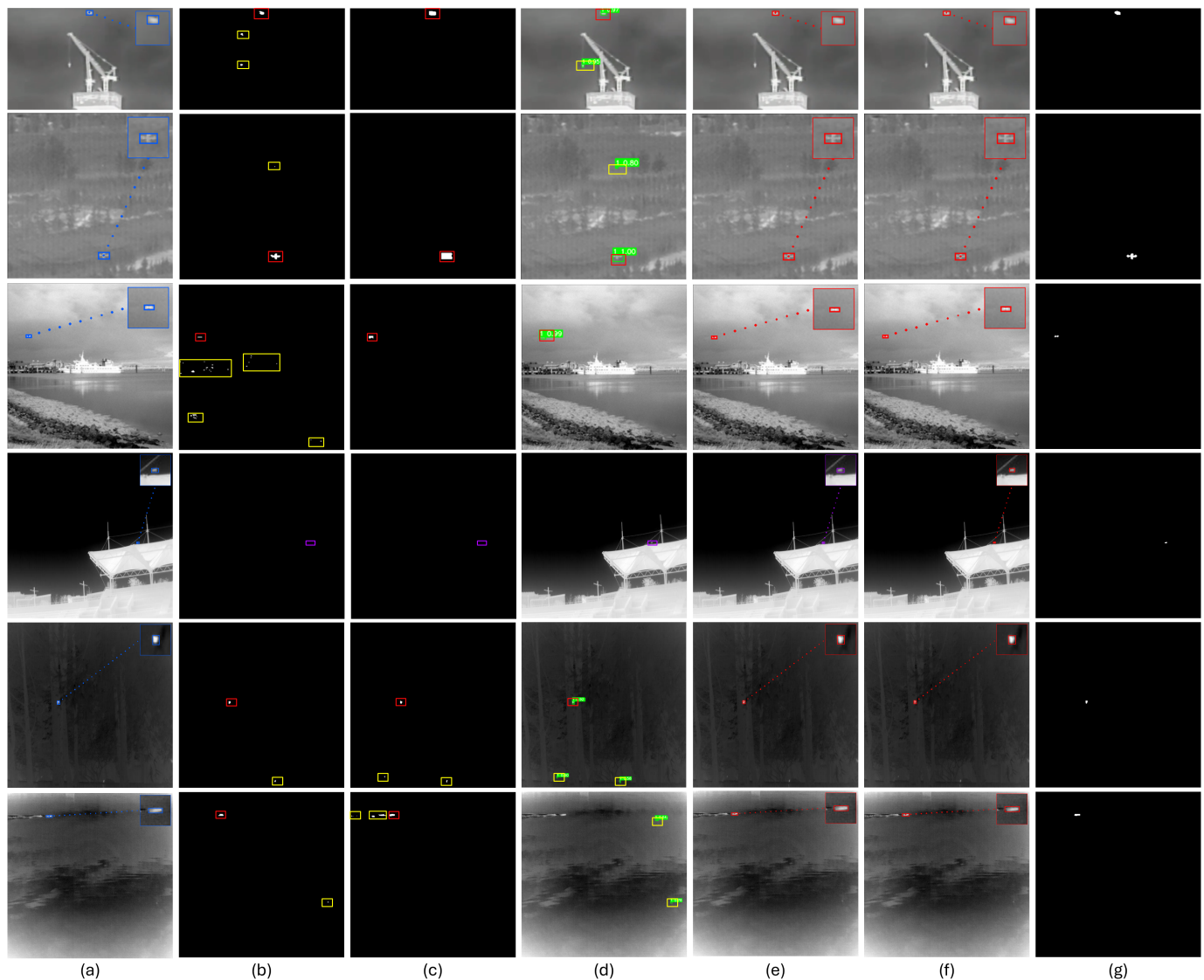


Figure 5. Representative qualitative results produced by different methods on the NUAA-SIRST, NUDT-SIRST, and IRSTD-1k datasets. Columns show (a) input images, (b) ALCNet, (c) DNANet, (d) EFLNet, (e) Grounding DINO, (f) PromptScaleDINO, and (g) ground truth. Blue, red, yellow, and purple boxes indicate ground-truth targets, correctly detected targets, false alarms, and missed targets, respectively.

3.5. Cross-Dataset Generalization

Cross-dataset evaluation tests whether PromptScaleDINO overfits to a single benchmark by evaluating a model trained on one dataset against another without additional fine-tuning. The results are direction-dependent. Training on IRSTD-1k yields F1 scores of 96.32% on SIRST and 87.39% on NUDT-SIRST, whereas models trained on SIRST and NUDT-SIRST obtain 76.83% and 80.75% F1 on IRSTD-1k. PromptScaleDINO therefore remains robust across related IRSTD datasets, but IRSTD-1k remains the harder domain.

Table 3 compares PromptScaleDINO with source-domain fine-tuned vanilla Grounding DINO under the same transfer protocol. PromptScaleDINO achieves higher F1 and mAP@0.5 in five of six directions and higher mAP@0.5:0.95 in four. The largest mAP@0.5 gains occur for NUDT-SIRST→IRSTD-1k (+9.12 percentage points) and IRSTD-1k→NUDT-SIRST (+8.46 percentage points). Grounding DINO performs better across all three metrics for SIRST→IRSTD-1k, confirming that the transfer advantage is direction-dependent rather than uniform.

Table 3. Cross-dataset generalization results.

Training Dataset	Test Dataset	Method	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
IRSTD-1k	NUDT-SIRST	Grounding DINO	83.59	83.56	28.88
		PromptScaleDINO	87.39	92.02	41.55
IRSTD-1k	SIRST	Grounding DINO	95.02	94.96	46.30
		PromptScaleDINO	96.32	98.38	47.85
NUDT-SIRST	IRSTD-1k	Grounding DINO	75.96	70.68	21.35
		PromptScaleDINO	80.75	79.80	28.29
NUDT-SIRST	SIRST	Grounding DINO	88.46	87.53	39.54
		PromptScaleDINO	91.78	92.25	39.48
SIRST	IRSTD-1k	Grounding DINO	79.71	78.53	30.29
		PromptScaleDINO	76.83	74.57	26.87
SIRST	NUDT-SIRST	Grounding DINO	91.43	94.09	40.29
		PromptScaleDINO	93.48	94.65	45.29

3.6. Ablation Studies

3.6.1. Structural Ablation

The structural ablation evaluates whether APB and SQRN each contribute independently and whether their combination improves over the stronger single-module variant on IRSTD-1k. Table 4 supports three conclusions. First, the APB-equipped variant improves F1 from 85.76% to 87.52% and mAP@0.5 from 83.40% to 85.84%, indicating that anchor-centered prompt enrichment improves text-side matching for weak infrared targets. Second, the SQRN-equipped variant improves mAP@0.5:0.95 from 35.88% to 39.67%, suggesting stronger high-IoU localization behavior for tiny boxes. Third, the full PromptScaleDINO model gives the best Recall, F1, and mAP@0.5, reaching 86.42%, 88.32%, and 87.40%, respectively. Its mAP@0.5:0.95 is slightly lower than that of the SQRN-only variant, indicating a trade-off: the full model favors recall, F1, and AP50-level detection quality, whereas SQRN alone gives the strongest strict high-IoU localization score in this ablation. Figure 6 provides qualitative examples of these structural-ablation differences. These results show that APB and SQRN are complementary but not identical in their metric effects.

Table 4. Structural ablation of APB and SQRN on IRSTD-1k. A ✓ indicates that the corresponding component is included in the model configuration; a blank cell indicates that it is not included.

Baseline	APB	SQRN	Precision (%)	Recall (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
✓			90.14	81.79	85.76	83.40	35.88
✓	✓		90.45	84.77	87.52	85.84	37.80
✓		✓	90.93	83.11	86.85	85.14	39.67
✓	✓	✓	90.31	86.42	88.32	87.40	39.25

Figure 6 also reveals a remaining limitation of the full model. In the second row, PromptScaleDINO identifies the target region, but the predicted box does not sufficiently overlap with the ground-truth box under the IoU@0.5 criterion and is therefore categorized as a localization error. This case illustrates the sensitivity of few-pixel target boxes to small coordinate deviations. Although SQRN and the geometry-aware objective improve localization overall, they do not completely eliminate pixel-level localization errors.

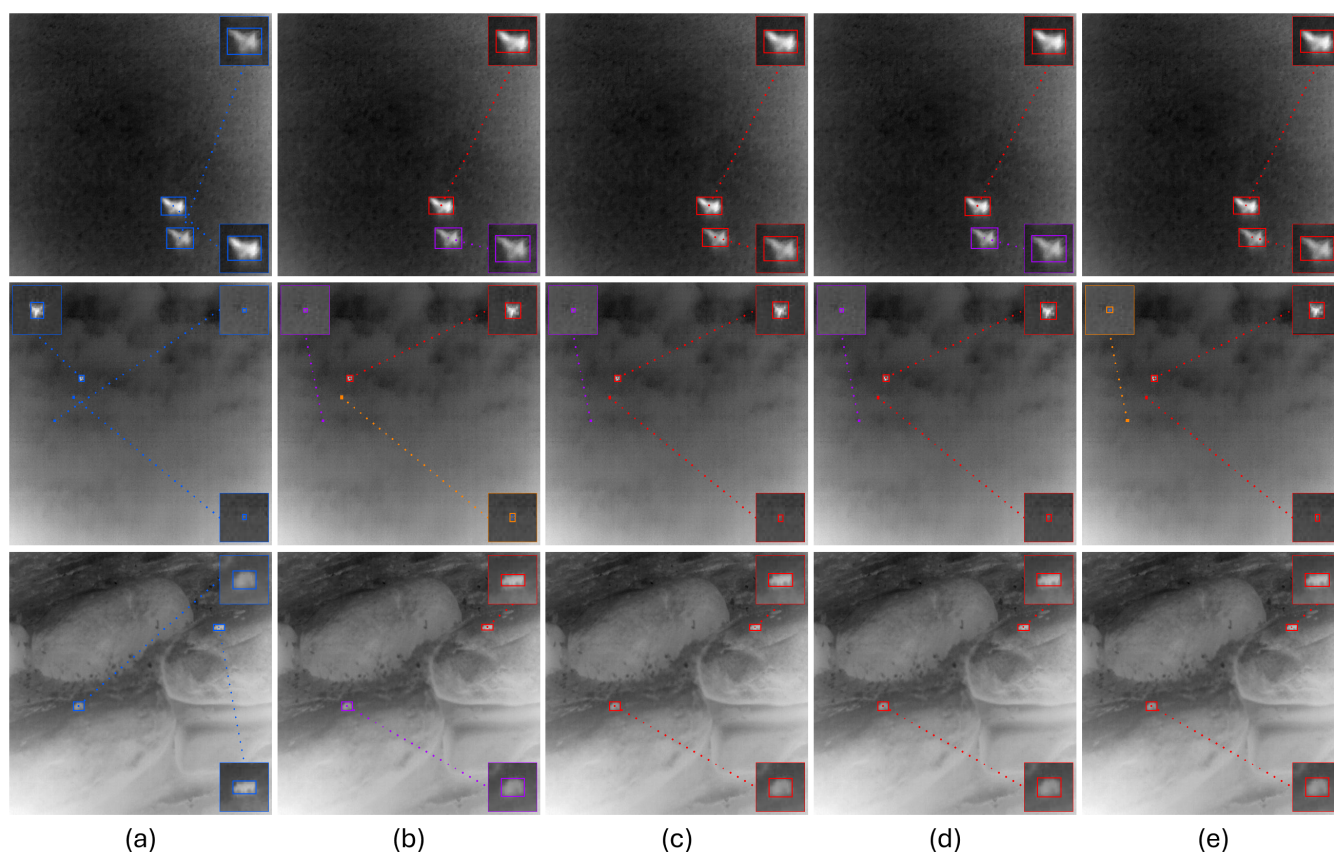


Figure 6. Representative qualitative results of the structural ablation onIRSTD-1k. Columns show (a) GT, (b) baseline, (c) baseline + APB, (d) baseline + SQRN, and (e) full PromptScaleDINO. Red, purple, blue, and orange boxes indicate TP, FN, GT, and localization errors (LOC), respectively.

3.6.2. APB Ablation

The APB ablation examines whether training-time prompt enrichment and anchor-only supervision improve detection under a single-anchor inference protocol. The baseline is trained with the anchor phrase alone, whereas APB introduces the four-prompt bank during training. All configurations use only “infrared small target” at inference. As shown in Table 5, compact-bank training improves F1 from 83.78% to 87.59%, mAP@0.5 from 81.76% to 86.52%, and mAP@0.5:0.95 from 34.24% to 37.94%. Because the inference prompt is identical, these gains isolate the contribution of the training-time prompt context.

Table 5. Training-time APB ablation onIRSTD-1k.

Training Configuration	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Single-prompt training	83.78	81.76	34.24
APB compact-bank training	87.59	86.52	37.94
APB without anchor-only supervision	76.21	78.05	32.29

The diagnostic setting without anchor-only supervision performs much worse, dropping to 76.21% F1 and 78.05% mAP@0.5. This result supports the core APB design: prompt variants should enrich the training context without becoming parallel supervised categories, while the semantic target remains anchored to one stable class phrase. The compact four-prompt bank with anchor-only supervision is therefore used for later experiments. This comparison tests the selected compact bank against single-prompt training under the same inference prompt; it does not exhaustively compare alternative prompt compositions or establish that these four phrases are optimal.

3.6.3. SQRN Ablation

Table 6 isolates the roles of scale estimation, multiscale query refinement, residual fusion, and semantic-confidence gating in SQRN. Removing the scale estimator causes the largest mAP@0.5 drop, from 87.14% to 84.74%, showing that scale-aware branch weighting is the main source of SQRN’s localization gain. Removing the multiscale branches produces a smaller degradation, suggesting that multiscale query transformation helps but is less critical than scale estimation.

Table 6. SQRN ablation on IRSTD-1k.

Variant	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Full SQRN	87.80	87.14	38.59
Without scale estimator	86.93	84.74	37.92
Without multiscale branches	87.48	87.03	38.52
Without residual refinement	87.07	84.93	38.55
Without confidence/semantic gating	87.56	86.82	38.24

The full SQRN achieves the best F1, mAP@0.5, and mAP@0.5:0.95 among the SQRN variants. Removing confidence/semantic gating decreases F1 from 87.80% to 87.56%, mAP@0.5 from 87.14% to 86.82%, and mAP@0.5:0.95 from 38.59% to 38.24%, indicating that semantic-confidence-guided refinement helps target the query update more effectively. These results support the full SQRN design used in the main APB-SQRN model.

We analyze last-layer semantic-confidence gate values on the IRSTD-1k test set. Each ground-truth target is matched to its highest-IoU decoder query without confidence thresholding, retaining low-scoring dim-target queries; this yields 302 GT-matched queries. In Figure 7a, boxes show the interquartile range (IQR), whiskers the most extreme observations within $1.5 \times$ IQR, orange lines medians, green triangles means, and gray circles outliers. The mean gate decreases monotonically with target area, from 0.86 for targets ≤ 25 px² to 0.64 for those > 84 px². In Figure 7b, teal dots denote individual GT-matched queries and show a strong negative correlation between the low-confidence gate term and anchor semantic confidence (Pearson $r = -0.97$). Under the decreasing gate defined in Section 2.3, small, low-confidence queries—the weak targets SQRN is designed to rescue—receive the strongest refinement rather than being suppressed.

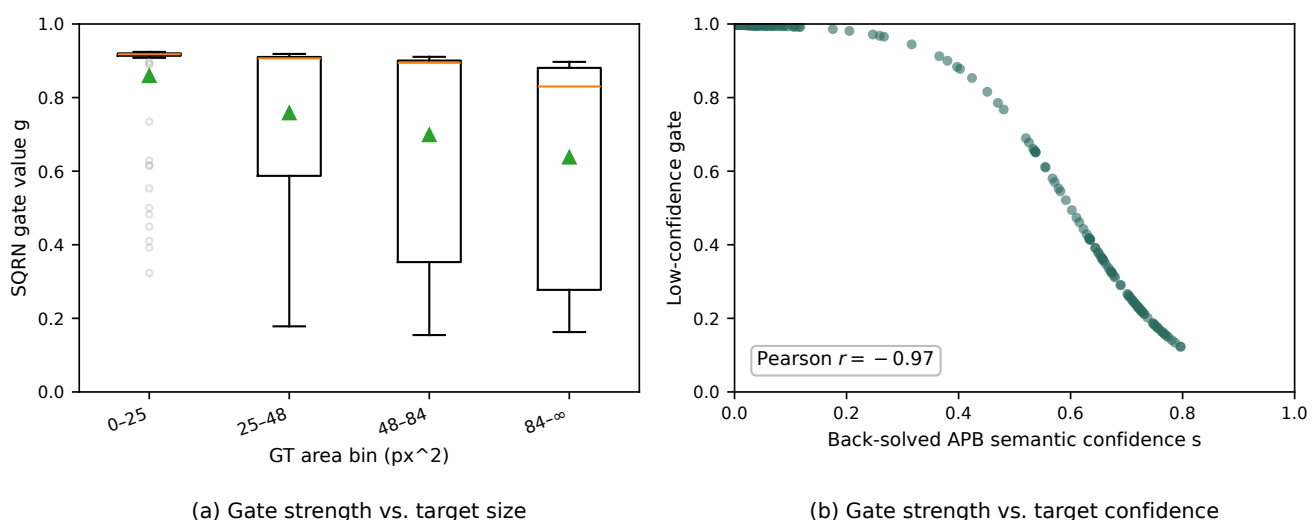


Figure 7. SQRN gate behavior on the IRSTD-1k test set in the last decoder layer (302 GT-matched queries). (a) Gate strength across target-area bins. (b) Low-confidence gate term versus back-solved anchor semantic confidence ($r = -0.97$).

3.6.4. Geometry-Aware Localization Objective Ablation

Table 7 compares GIoU only, NWD only, and GIoU + NWD using either the default $C = 12.8$ or the validation-calibrated $C = 25.6$ on theIRSTD-1k test split. AP is averaged over IoU thresholds from 0.50 to 0.95; AP50 and AP75 use individual thresholds, APs denotes small-object AP, and AR10/AR100 denote recall with at most 10/100 detections per image.

Table 7. Test-set ablation of the geometry-aware tiny-box localization objective on IIRSTD-1k, separating NWD integration at the default $C = 12.8$ from validation-based tuning to $C = 25.6$, fixed before testing.

Setting	AP (%)	AP50 (%)	AP75 (%)	APs (%)	AR10 (%)	AR100 (%)
GIoU only	39.46	86.65	29.61	39.47	48.84	49.11
NWD only	38.51	87.75	27.21	38.52	48.38	48.77
GIoU + NWD (default $C = 12.8$)	39.70	86.60	29.71	39.71	48.08	48.71
GIoU + NWD (calibrated $C = 25.6$)	40.34	88.85	29.42	40.35	48.77	49.47

Replacing GIoU with NWD alone reduces AP from 39.46% to 38.51% and AP75 from 29.61% to 27.21%, indicating that NWD cannot replace overlap-based regression as an independent localization objective. By contrast, combining NWD with GIoU using the default $C = 12.8$ increases AP to 39.70% and AP75 to 29.71%. This controlled comparison attributes a 0.24 percentage-point AP improvement to integrating the two losses, independently of dataset-specific calibration.

Calibrating the normalization constant from $C = 12.8$ to $C = 25.6$ further increases AP from 39.70% to 40.34%, AP50 from 86.60% to 88.85%, and AR100 from 48.71% to 49.47%. The calibration therefore contributes an additional 0.64 percentage points in AP, yielding a total gain of 0.88 percentage points over GIoU only. However, AP75 decreases slightly from 29.71% to 29.42%, and AR10 remains below the GIoU-only result. The benefit of calibration is therefore concentrated in overall small-target AP, AP50, and recall rather than uniformly improving strict high-IoU localization.

3.6.5. Reduced-Data Fine-Tuning

We assess PromptScaleDINO under reduced-data supervision on IIRSTD-1k to examine its robustness when training data are limited. We fine-tune PromptScaleDINO on the full training set and on random half and quarter subsets, evaluating all models with the same protocol. Table 8 reports the resulting sensitivity to training-set size. Reducing the training set from 100% to 50% decreases F1 from 88.32% to 86.97%, while the 25% subset retains 86.10% F1. The mAP@0.5 score follows the same trend, decreasing from 87.40% to 85.02% and 84.93%. We interpret this experiment conservatively as a split-level robustness check rather than a complete low-data learning study.

Table 8. Reduced-data fine-tuning results of PromptScaleDINO on IIRSTD-1k.

Training Data Scale (%)	F1 (%)	mAP@0.5 (%)
100	88.32	87.40
50	86.97	85.02
25	86.10	84.93

3.7. Efficiency Analysis

Table 9 evaluates the incremental overhead of PromptScaleDINO relative to its Grounding DINO backbone, rather than comparing efficiency across different architectures. APB

adds no parameters or inference-time GFLOPs because inference retains the single anchor prompt, while the geometry-aware localization objective is training-only. SQRN adds 0.46 M parameters (+0.26%) and 0.35 GFLOPs (+0.14%). The full model reduces throughput from 8.47 to 8.12 FPS (4.1%), indicating limited overhead relative to the reported accuracy gains.

Table 9. Efficiency comparison of PromptScaleDINO variants.

Method	Total Params (M)	Trainable Params (M)	GFLOPs	FPS
Grounding DINO	174.33	172.25	251.88	8.47
+APB	174.33	172.25	251.88	8.45
+SQRN	174.79	172.71	252.23	8.32
+APB-SQRN (full)	174.79	172.71	252.23	8.12

4. Discussion

PromptScaleDINO introduces complementary adaptations at three stages of Grounding DINO: APB stabilizes text-side supervision, SQRN refines scale-sensitive decoder queries, and the geometry-aware objective improves few-pixel localization. Because these components act on the text input, query update, and training loss, respectively, they can be combined while preserving the pretrained detector and the overall grounded-detection architecture.

We use two-stage APB–SQRN training because SQRN gates query updates using anchor-phrase confidence and predicted box scale, both of which can be unstable during early joint optimization. Training APB first provides a more stable confidence source and starting point for subsequent SQRN refinement. One-stage joint optimization remains possible but would require a separately designed training schedule and is not evaluated here.

The larger performance headroom and more difficult transfer into IRSTD-1k make it the most informative benchmark for evaluating the proposed adaptation, while SIRST and NUDT-SIRST provide complementary checks across real and synthetic distributions.

The method has three main limitations. First, its gains are recall-oriented: APB and SQRN improve target sensitivity, and the geometry-aware objective improves localization, but none explicitly suppresses false alarms or models background clutter. This recall-oriented behavior reflects the intended scope of the three adaptations rather than an unavoidable limitation of vision-language transfer; explicit clutter discrimination would require additional false-alarm-aware supervision. Second, APB uses a manually designed, dataset-agnostic prompt group, and Table 5 does not establish that its four phrases are optimal; dataset-specific or image-conditioned alternatives may introduce domain bias and would require separate validation under the same single-anchor deployment protocol. Third, SQRN depends on predicted boxes and confidence, and its scale descriptor includes only box geometry, so poor early predictions and missing appearance or scene context can limit refinement. The localization objective acts only on box regression and therefore cannot resolve semantic confusion.

The evaluation also has boundaries. Threshold-dependent metrics use a fixed confidence threshold of 0.5, but deployment on new sensors or scenes may require operating-point calibration. The main tables report the seed-42 run; auxiliary seed sweeps show that absolute mAP@0.5 varies, although the adapted models remain competitive with the APB and SQRN single-module baselines. Cross-dataset results further show direction-dependent robustness and persistent domain shift.

The NWD normalization constant is domain-dependent: $C = 25.6$ is selected by validation for the present IRSTD-1k protocol rather than serving as a sensor-independent default. Its appropriate value depends on the target-box scale after image preprocessing.

Recalibration may not be necessary when target-size distributions are similar but should be considered when sensor resolution, imaging distance, resizing policy, or target definition shifts this distribution. In that case, a small validation set can select C from a limited candidate range guided by robust statistics of $\sqrt{wh}$, after which it should be fixed before test-set evaluation.

Future work will investigate learned or image-conditioned prompt contexts, stronger prompt-group supervision, explicit false-alarm modeling through negative prompts or hard-background supervision, adaptive query refinement, and evaluation across broader sensor types.

5. Conclusions

This paper presents PromptScaleDINO, a lightweight framework for adapting Grounding DINO to infrared small target detection in remote-sensing imagery. The framework aligns the pretrained detector along three complementary axes. APB stabilizes the text side through anchor-centered prompt enrichment during training while retaining a single anchor phrase at inference. SQRN improves the decoder side by refining queries according to predicted box scale and semantic confidence. A geometry-aware tiny-box localization objective integrates NWD with GIoU and calibrates the NWD normalization constant for IRSTD box scales, providing smoother supervision for few-pixel boxes than the original Grounding DINO regression objective.

Under the adopted detection protocol, PromptScaleDINO achieves its clearest gains on IRSTD-1k while maintaining strong performance on SIRST and NUDT-SIRST. The ablations show that APB, SQRN, and the geometry-aware objective provide complementary improvements in weak-target recovery and localization robustness. PromptScaleDINO therefore offers a practical basis for language-assisted infrared small target detection, while explicit false-alarm-aware modeling remains a direction for future work.

Author Contributions: Conceptualization, C.H.; methodology, C.H. and Y.S.; software, C.H.; validation, C.H., Y.C. and Y.S.; formal analysis, C.H.; investigation, C.H. and Z.W.; resources, C.H., Z.W., C.L. and G.L.; data curation, C.H. and J.J.; writing—original draft preparation, C.H.; writing—review and editing, C.H., Z.W., Y.C., C.L. and G.L.; visualization, C.H.; supervision, C.H., J.J., C.L. and G.L.; project administration, C.H. and G.L.; funding acquisition, C.H. and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB0980000).

Data Availability Statement: The SIRST dataset supporting this research can be obtained at the following website: <https://github.com/YimianDai/sirst> (accessed on 19 November 2025). The IRSTD-1k dataset supporting this research can be obtained at the following website: <https://github.com/RuiZhang97/ISNet> (accessed on 20 November 2025). The NUDT-SIRST dataset supporting this research can be obtained at the following website: <https://github.com/YeRen123455/Infrared-Small-Target-Detection> (accessed on 24 November 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gao, C.; Meng, D.; Yang, Y.; Wang, Y.; Zhou, X.; Hauptmann, A.G. Infrared patch-image model for small target detection in a single image. *IEEE Trans. Image Process.* **2013**, *22*, 4996–5009. [CrossRef] [PubMed]
2. Chen, C.L.P.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 574–581. [CrossRef]
3. Wei, Y.; You, X.; Li, H. Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recognit.* **2016**, *58*, 216–226. [CrossRef]

4. Zhang, L.; Peng, L.; Zhang, T.; Cao, S.; Peng, Z. Infrared small target detection via non-convex rank approximation minimization joint l2,1 norm. *Remote Sens.* **2018**, *10*, 1821. [[CrossRef](#)]
5. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [[CrossRef](#)]
6. Wang, H.; Zhou, L.; Wang, L. Miss detection vs. false alarm: Adversarial learning for small object segmentation in infrared images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 8509–8518.
7. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional local contrast networks for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [[CrossRef](#)]
8. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2021; pp. 949–958.
9. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 867–876.
10. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **2023**, *32*, 1745–1758. [[CrossRef](#)] [[PubMed](#)]
11. Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; Pi, Y. EFLNet: Enhancing feature learning network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5906511. [[CrossRef](#)]
12. Zhang, T.; Li, L.; Cao, S.; Pu, T.; Peng, Z. Attention-guided pyramid context networks for detecting infrared small target under complex background. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4250–4261. [[CrossRef](#)]
13. Chen, G.; Wang, W.; Tan, S. IRSTFormer: A hierarchical vision transformer for infrared small target detection. *Remote Sens.* **2022**, *14*, 3258. [[CrossRef](#)]
14. Hou, Q.; Zhang, L.; Tan, F.; Xi, Y.; Zheng, H.; Li, N. ISTDU-Net: Infrared small-target detection U-Net. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7506205. [[CrossRef](#)]
15. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior attention-aware network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002013. [[CrossRef](#)]
16. Liu, F.; Gao, C.; Chen, F.; Meng, D.; Zuo, W.; Gao, X. Infrared small and dim target detection with transformer under complex backgrounds. *IEEE Trans. Image Process.* **2023**, *32*, 5921–5932. [[CrossRef](#)] [[PubMed](#)]
17. Sun, H.; Bai, J.; Yang, F.; Bai, X. Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset IRDST. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5000513. [[CrossRef](#)]
18. Yuan, S.; Qin, H.; Yan, X.; Akhtar, N.; Mian, A. SCTransNet: Spatial-channel cross transformer network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5002615. [[CrossRef](#)]
19. Liu, Q.; Liu, R.; Zheng, B.; Wang, H.; Fu, Y. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 17490–17499.
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, Virtual, 18–24 July 2021; pp. 8748–8763.
21. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 10955–10965.
22. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 38–55.
23. Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; Carion, N. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 1780–1790.
24. Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2022.
25. Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. Simple open-vocabulary object detection with vision transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2022.
26. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 16901–16911.
27. Zareian, Y.; Rosa, K.D.; Hu, D.H.; Chang, S.-F. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 14393–14402.

28. Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; Xu, H. DetCLIP: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*; Curran Associates Inc.: Red Hook, NY, USA, 2022; pp. 9125–9138.
29. Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; Cai, J. Learning object-language alignments for open-vocabulary object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2023.
30. Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; Li, G. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 14084–14093.
31. Pan, J.; Liu, Y.; Fu, Y.; Ma, M.; Li, J.; Paudel, D.P.; Gool, L.V.; Huang, X. Locate anything on Earth: Advancing open-vocabulary object detection for remote sensing community. *Proc. AAAI Conf. Artif. Intell. (AAAI)* **2025**, *39*, 6281–6289. [\[CrossRef\]](#)
32. Ju, M.; Feng, Y.; Diao, W.; Liu, C. Addressing dense small-object detection in remote sensing: An open-vocabulary object detection framework. *Remote Sens.* **2026**, *18*, 851. [\[CrossRef\]](#)
33. Huang, F.; Zheng, S.; Qiu, Z.; Liu, H.; Bai, H.; Chen, L. 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 (ICCV)*; IEEE: New York, NY, USA, 2025.
34. Zhang, M.; Xu, Q.; Wang, Y.; Li, X.; Yuan, H. MIRSAM: Multimodal vision-language segment anything model for infrared small target detection. *Vis. Intell.* **2025**, *3*, 4. [\[CrossRef\]](#)
35. Zhang, M.; Li, X.; Gao, F.; Guo, J.; Gao, X.; Zhang, J. 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 (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 9549–9558.
36. Chib, P.S.; Singh, P. Leveraging language prior for infrared small target detection. *Neurocomputing* **2026**, *684*, 133512. [\[CrossRef\]](#)
37. Wang, Z.; Sun, H.; Lu, J.; Wang, W.; Huang, Z.; Zhang, D.; Qian, X.; Han, J. Dual-granularity semantic prompting for language guidance infrared small target detection. *arXiv* **2025**, arXiv:2511.19306.
38. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. *arXiv* **2022**, arXiv:2110.13389.
39. Rezaatofghi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 658–666.
40. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
41. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2021.
42. 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 *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA, 2023.
43. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [\[CrossRef\]](#)
44. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 16795–16804.
45. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic DETR: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 2988–2997.
46. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 14454–14463.
47. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell. (AAAI)* **2020**, *34*, 12993–13000. [\[CrossRef\]](#)
48. Atrash, A.; Ertekin, S.; Ugur, O.; Moured, O.; Chen, Y.; Zhang, J. TY-RIST: Tactical YOLO tricks for real-time infrared small target detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE: New York, NY, USA, 2025; pp. 2222–2231.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.