# LEARNING WHAT TO PRESERVE: TARGET-DIRECTED INFORMATION BOTTLENECK FOR INFRARED SMALL TARGET DETECTION UNDER DEGRADATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Information preservation is essential for infrared small target detection, but indiscriminate preservation becomes a liability when degradation dominates the input. Conventional encoders make retention an implicit consequence of task optimization, whereas invertible encoders make information recoverable without deciding what is worth retaining. This work introduces Target-Directed Information Bottleneck (TDIB), which unifies selective information retention with inference-aligned detection. Target-Directed Selective Retention (TDSR) integrates a frozen invertible prefix and a variational bottleneck, confining information loss to the bottleneck and making its cost measurable through the fixed inverse. Its spatially weighted KL rate drives information compression, while localized inverse reconstruction protects fragile target evidence without requiring clean references. Inference-Aligned Ranking (IAR) further turns posterior variation into supervision for reliable detection, encouraging mean–sample prediction consistency and ranking candidates by localization quality discounted by posterior drift. Comparisons and ablations on the proposed IR-DegradeX benchmark show that TDIB consistently improves detection accuracy, while representation analyses confirm that it preserves target evidence and stabilizes predictions. These findings point to a principled pathway toward target-directed representation learning for degraded visual perception.

## 1 INTRODUCTION

An infrared small target may be distinguished from its surroundings by only a few low-contrast pixels. With little texture or shape available, detection depends on whether those sparse cues survive the transformations that build a useful representation. Imaging degradation makes this dependence more fragile: blur disperses target responses, while stripe noise, random noise, and nonuniformity obscure them within background clutter. Under such degradation, the evidence a detector needs is not only weak but also entangled with interference (Fig. 1, top).

To sustain such weak evidence, existing detectors preserve spatial resolution, strengthen cross-scale interaction, and enhance weak target responses (Dai et al., 2021; Li et al., 2023; Wu et al., 2023). Under degradation, restoration-guided and feature-purification approaches further improve the representation (Wang et al., 2026; Yuan et al., 2026). These advances improve feature quality, yet information retention in conventional encoders remains largely implicit: prediction supervision rewards a correct detection, while architecture and optimization determine which parts of the observation remain available to support it (Fig. 1, second row). Invertible representations offer a different starting point by keeping the observation recoverable from its features (Jacobsen et al., 2018), and InvDet exploits this property through an invertible encoder with target-aware reconstruction guidance (Yan et al., 2026). Under degradation, however, the recoverable observation contains target structure together with noise and imaging artifacts, all of which remain equally accessible (Fig. 1, third row). Recoverability ensures that weak evidence is not lost, but leaves open what a detector should selectively retain.

Inspired by the information bottleneck principle (Tishby et al., 1999), we formulate degraded infrared small-target detection as learning what to retain. The bottleneck constrains the information a



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.

representation inherits from its input while maintaining its predictive value for the task, and its variational form makes this trade-off trainable (Alemi et al., 2017). For degraded small targets, however, the cost of compression is strongly asymmetric: a slight change in a small target neighborhood can remove an essential cue, whereas a large background region need not receive the same protection. Relaxing compression around targets determines where retention should be less costly, but it does not specify which local structure should survive there. Answering the latter requires observing what selection does to the target. When a variational bottleneck follows a frozen invertible prefix, information loss can arise only at the bottleneck. The fixed inverse can then do more than recover the input: it can measure and constrain the consequences of discarding information.

Based on this insight, we propose Target-Directed Information Bottleneck (TDIB), which unifies selective information retention with inference-aligned detection. Its core, Target-Directed Selective Retention (TDSR), integrates a frozen invertible prefix and a variational bottleneck into a single retention mechanism (Fig. 1, bottom row). Within it, a spatially weighted KL rate drives information compression while reducing the pressure within target regions, and localized inverse reconstruction protects fragile target evidence from the damage that selection may cause. Spatial rate allocation thus determines where retention should be less costly, while local inverse fidelity determines what structure should survive there. The reconstruction reference is the degraded input itself, so learning requires no clean counterpart. All detection features, including finer-scale features recovered through a partial inverse, are derived from the selected representation, so the detector sees only what the bottleneck retains. To exploit the retained representation at inference, Inference-Aligned Ranking (IAR) turns posterior variation into supervision for the deterministic mean path. The bottleneck provides a posterior mean for inference and stochastic samples that reveal prediction sensitivity to posterior variation. IAR encourages mean–sample prediction consistency and ranks mean-path candidates by localization quality discounted by posterior drift, so that candidates supported across posterior variation are preferred at test time. Sampling and full-image reconstruction serve learning only, and inference requires a single forward pass.

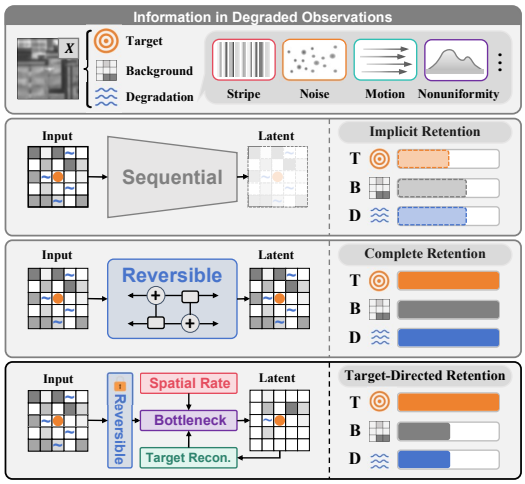


Figure 1: **From implicit to target-directed retention.** Bars schematically illustrate the retention of target (T), background (B), and degradation (D) information. A conventional encoder retains information implicitly, and weak target cues may be lost. An invertible encoder retains everything, including degradation. TDIB confines information loss to a bottleneck whose cost is measured by the fixed inverse, retaining target evidence while compressing the remaining observation.

We evaluate TDIB on IR-DegradeX, which composes seven degradation operators into single-, paired-, and multi-degradation observations of the same source images. A single model trained on mixed compositions achieves the highest accuracy in every composition setting, and its advantage over each ablated variant holds across three random seeds. Representation analyses further show that localized inverse fidelity preserves target structure and that IAR stabilizes predictions under posterior sampling, even when degradation remains identifiable after selection. Our main contributions are summarized as follows:

- We formulate degraded infrared small-target detection as learning what to retain, and propose TDIB, which unifies selective information retention with inference-aligned detection.
- We design TDSR, which integrates a frozen invertible prefix and a variational bottleneck so that information loss is confined to the bottleneck, and its consequences for target evidence are measured and constrained without clean references.
- We introduce IAR, which turns posterior variation into supervision for the deterministic mean path through mean–sample consistency and drift-discounted ranking.

- We construct IR-DegradeX and validate TDIB through comparisons, multi-seed ablations, and representation analyses, showing consistent accuracy gains together with preserved target evidence and stabilized predictions.

## 2 Related Work

### 2.1 Infrared Small Target Detection under Degradation

Deep infrared small target detectors sustain weak responses through contextual interaction, hierarchical aggregation, and resolution-preserving feature learning, as exemplified by ACM, DNANet, and UIU-Net (Dai et al., 2021; Li et al., 2023; Wu et al., 2023). Other designs introduce cascade refinement (Dai et al., 2023) or embed low-rank and sparse priors through deep unfolding (Wu et al., 2025). Recent work further addresses imaging degradation within the detection pipeline. UniCD jointly optimizes nonuniformity correction and infrared UAV detection, using a detection-guided loss to align features from corrected and clear images during joint training (Fang et al., 2025). JFD$^3$ restores blurred features under clear-image feature supervision and incorporates frequency-based structural guidance into detection (Wang et al., 2026). NS-FPN suppresses noise through low-frequency-guided feature purification and target-oriented sampling (Yuan et al., 2026). These methods improve the observations or features available to the detector. In contrast, TDIB explicitly learns which information to retain and constrains selection-induced target distortion using the degraded observation itself, without requiring paired clean references.

### 2.2 Invertible Representations and Information Bottleneck

Invertible networks make inputs recoverable from their representations through bijective mappings (Dinh et al., 2017; Kingma & Dhariwal, 2018). i-RevNet demonstrates invertible discriminative feature learning (Jacobsen et al., 2018), while IRN applies invertibility to image rescaling (Xiao et al., 2020). InvDet brings this principle to infrared detection through reconstruction-guided encoder learning, with target-aware modulation restricted to the inverse path (Yan et al., 2026). Reconstruction thereby constrains feature extraction rather than acting as a separate restoration task. Under degradation, however, such lossless representations preserve corruption as faithfully as target structure.

The information bottleneck offers a complementary principle by balancing compression with task relevance (Tishby et al., 1999), and Deep VIB provides a trainable variational formulation (Alemi et al., 2017). BiDet applies this principle to reduce redundancy in binarized detectors (Wang et al., 2020). Such objectives guide task-relevant compression but do not directly measure the resulting distortion of local target structure. More directly related, IB-INNs introduce controlled information loss into normalizing flows, using the bottleneck objective to balance generative modeling with classification (Ardizzone et al., 2020). TDSR instead uses the inverse to measure and constrain the cost of information selection for dense detection. Unlike reconstruction-guided encoder learning, the inverse remains fixed, providing a stable fidelity reference for the representation selected by the bottleneck rather than a moving reconstruction target. Reconstruction thus guides the bottleneck's selection rather than updating the invertible encoder.

### 2.3 Consistency Regularization and Quality-Aware Ranking

Consistency regularization encourages agreement between perturbed predictions. Mean Teacher enforces consistency with targets from weight-averaged models (Tarvainen & Valpola, 2017), and R-Drop minimizes bidirectional KL divergence between output distributions from dropout-based forward passes (Liang et al., 2021). In dense detection, IoU-Net predicts localization confidence to guide candidate selection (Jiang et al., 2018), GFL and VarifocalNet incorporate localization quality into classification scores (Li et al., 2020; Zhang et al., 2021), and RankDetNet explicitly learns IoU-guided candidate ordering (Liu et al., 2021). GFLv2 further estimates localization quality from bounding-box distribution statistics (Li et al., 2021). IAR differs in two respects. First, it draws its perturbations from TDSR's learned representation posterior rather than from dropout or parameter averaging. Second, it assesses reliability from the resulting changes in classification and localization rather than from output-box distribution statistics, ranking candidates by localization

quality discounted by posterior drift. Together, these turn posterior variation into supervision for the mean-path detector.

## 3 TARGET-DIRECTED INFORMATION BOTTLENECK

**Overview.** TDIB turns information preservation under degradation into explicit, target-directed selection over a recoverable representation. As shown in Fig. 2, TDSR integrates a frozen invertible prefix with a variational bottleneck whose mean $z_\mu$ and sample $z_s$ enter a shared detector $D_\psi$. IAR then converts the posterior-induced variation between the two paths into supervision for the mean path used at deployment.

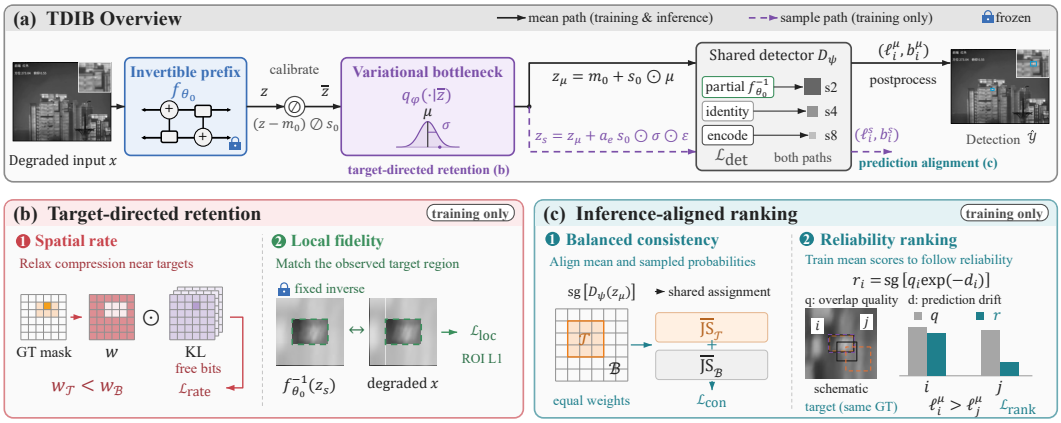


Figure 2: **Overview of TDIB.** (a) TDSR integrates a frozen invertible prefix with a variational bottleneck; the mean $z_\mu$ and sample $z_s$ enter a shared detector, and only $z_\mu$ is used at inference. (b) A spatially weighted rate lowers the compression penalty around targets, and local inverse fidelity constrains target distortion through the fixed inverse. (c) IAR uses posterior variation for mean–sample consistency and drift-discounted candidate ranking.

### 3.1 TARGET-DIRECTED SELECTIVE RETENTION

TDSR starts from a recoverable representation so that the consequences of subsequent information selection remain observable. Following InvDet (Yan et al., 2026), an invertible prefix with two stages, each comprising Haar downsampling and coupling blocks, maps the degraded input $x$ to $z = f_{\theta_0}(x)$, with $f_{\theta_0}^{-1}(z) = x$. The prefix and detector are initialized from a detector trained on the same degraded split, after which the prefix is frozen so that its inverse stays fixed while selection is learned.

Recoverability provides a stable pre-selection representation, but information selection itself is performed by the variational bottleneck. On the calibrated latent $\bar{z} = (z - m_0) \oslash s_0$, with fixed channel statistics $m_0$ and $s_0$, a residual mean predictor and a scale predictor give $\mu = \bar{z} + \Delta_\phi(\bar{z})$ and $\sigma = \mathrm{softplus}(S_\phi(\bar{z}))$. The resulting posterior yields the mean and sampled representations

$$q_\phi(\cdot \mid \bar{z}) = \mathcal{N}\big(\mu, \mathrm{diag}(\sigma^2)\big), \qquad z_\mu = m_0 + s_0 \odot \mu, \qquad z_s = m_0 + s_0 \odot (\mu + a_e\, \sigma \odot \epsilon), \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $a_e \in [0, 1]$ ramps in the sampling noise during training, while the rate always uses the nominal posterior. Both representations enter the detector $D_\psi$, which builds its finest features by partially inverting them, uses them directly at the intermediate scale, and encodes them further for coarser scales. The invertible prefix therefore preserves a recoverable reference before selection, while every detection scale is constructed after the bottleneck. Keeping the inverse fixed maps bottleneck-induced changes back to the observation domain, where their effect on target structure can be directly constrained.

The KL rate provides the pressure for information compression, but not its spatial allocation. TDSR therefore redistributes this pressure according to target location, weighting the local KL $k_{bcv} = \frac{1}{2}(\mu_{bcv}^2 + \sigma_{bcv}^2 - \log \sigma_{bcv}^2 - 1)$ at image $b$, channel $c$, and latent location $v$ by $w_{bv} \propto 1 - (1 - \rho) M_{bv}$

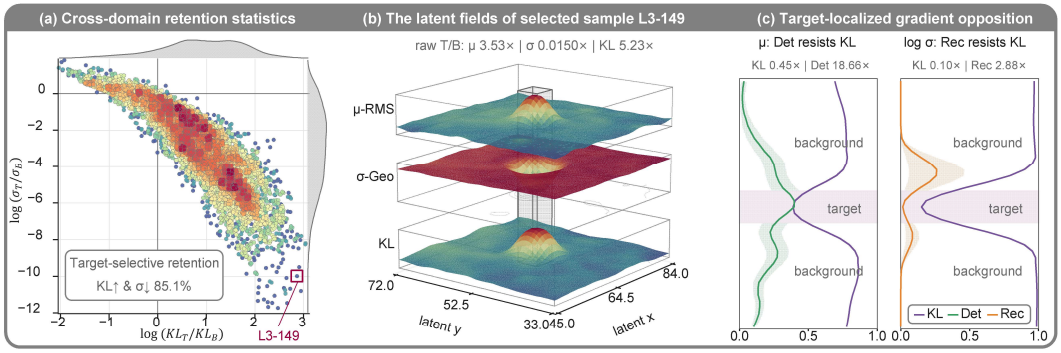


Figure 3: **Target-directed retention.** (a) Target/background ratios of raw KL and posterior standard deviation; most observations exhibit higher target KL and lower target variation. (b) Posterior mean magnitude, standard deviation, and raw KL around a representative target. (c) Detection and reconstruction oppose the weighted rate in $\mu$ and $\log \sigma$, respectively.

with $0 < \rho < 1$, where $M_{bv}$ marks target boxes projected onto the latent grid and dilated by one cell:

$$\mathcal{L}_{\text{rate}} = \frac{1}{C} \sum_{c=1}^{C} \max \left( \tau,\ \langle w_{bv}\, k_{bcv} \rangle_{b,v} \right), \tag{2}$$

where $\langle \cdot \rangle_{b,v}$ averages over the batch and latent grid. The spatial weights are normalized to unit mean, so relaxing target neighborhoods shifts pressure toward the background rather than uniformly weakening it. Target neighborhoods therefore receive a smaller relative rate penalty, leaving greater capacity for detection-relevant evidence. The free-bits threshold $\tau$ stops the rate term from further reducing channels below the allowance, while detection supervision determines which retained information is useful.

A lower cost of retention does not ensure that the observed target structure survives selection. The fixed inverse exposes the observation-space distortion introduced by the selected representation, allowing TDSR to constrain it within target regions. For a nonempty target-region mask $R_{bp}$,

$$\mathcal{L}_{\text{loc}} = \frac{\sum_{b,c,p} R_{bp} \left| f_{\theta_0}^{-1}(z_s)_{bcp} - x_{bcp} \right|}{C_x \sum_{b,p} R_{bp}}, \tag{3}$$

where $R_{bp}$ marks target boxes dilated by two pixels and $C_x$ is the number of input channels. Images without valid target regions contribute no local-fidelity loss. The reference is the degraded observation itself, so no clean counterpart is required. Evaluating fidelity on posterior samples requires target structure to remain stable under the variation allowed by the bottleneck. Background pixels do not enter the term, since whole-image fidelity would make retaining all background and corruption part of the objective. Together, the weighted rate determines where compression is relaxed, local inverse fidelity constrains what target structure survives there, and detection supervision determines which retained evidence is useful for the task. Figure 3 visualizes the resulting posterior: target regions typically exhibit higher raw KL and lower posterior variation than the background.

### 3.2 Inference-Aligned Ranking

The retained representation is stochastic during training, whereas inference uses only its mean. The two paths must therefore be compared under a common target correspondence. Let $(\ell_i^{\mu}, b_i^{\mu})$ and $(\ell_i^{s}, b_i^{s})$ denote the classification logit and box of candidate $i$ on each path. A task-aligned assignment (Feng et al., 2021), computed from detached mean predictions, supplies positives, matched boxes $g_i$, and IoU-based localization quality $q_i$ shared by both paths. Both paths receive the detection loss $\mathcal{L}_{\text{det}} = \frac{1}{2}\ell_{\text{det}}(D_\psi(z_\mu)) + \frac{1}{2}\ell_{\text{det}}(D_\psi(z_s))$, which combines Varifocal classification (Zhang et al., 2021) with quality-weighted Complete-IoU regression (Zheng et al., 2020).

Under this shared correspondence, IAR first constrains classification responses to remain stable under posterior variation. With $p_i^{\mu} = \text{sigmoid}(\ell_i^{\mu})$ and $p_i^{s} = \text{sigmoid}(\ell_i^{s})$, the disagreement $J_i =$

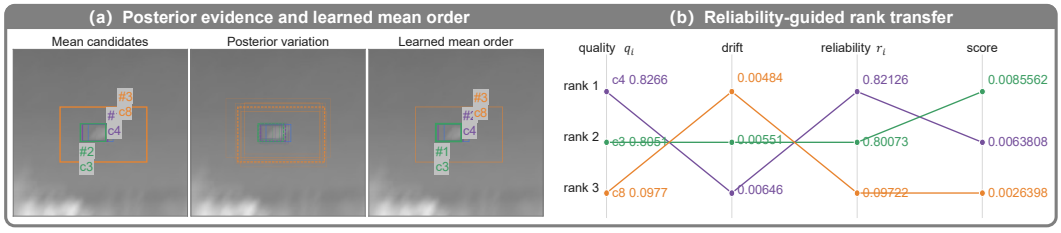


Figure 4: **Drift-discounted ranking.** (a) Candidates of one target on the mean path and across posterior samples. (b) Posterior drift discounts localization quality into reliability, which supervises the ordering of mean-path scores.

$\mathrm{JS}(\mathrm{Ber}(p_i^\mu) \,\|\, \mathrm{Ber}(p_i^s))$ is averaged separately over target-related candidates $\mathcal{T}$ and the remaining candidates $\mathcal{B}$:

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{|\mathcal{S}|} \sum_{\mathcal{U} \in \mathcal{S}} \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} J_i, \tag{4}$$

where $\mathcal{S}$ contains the nonempty sets among $\mathcal{T}$ and $\mathcal{B}$, and $\mathcal{T}$ covers assigned positives and candidates within boxes dilated by twice the detection stride. Balancing the two regions keeps abundant background responses from overwhelming sparse target responses.

Consistency constrains how each candidate changes between paths, but not how candidates of the same target should be ordered, and this ordering directly affects which candidates are retained during post-processing. IAR therefore derives a reliability target from posterior drift. For each positive candidate, drift combines normalized center displacement, log-area change, and classification divergence, and discounts localization quality:

$$d_i = \frac{\|c(b_i^\mu) - c(b_i^s)\|_2}{\mathrm{diag}(g_i)} + \left| \log \frac{A(b_i^\mu)}{A(b_i^s)} \right| + J_i, \qquad r_i = \mathrm{sg}[q_i \exp(-d_i)], \tag{5}$$

where $c(\cdot)$ and $A(\cdot)$ denote box center and area, $\mathrm{diag}(g_i)$ is the diagonal length of the matched target box, and $\mathrm{sg}$ denotes stop-gradient. A candidate thus receives high reliability only when it is both well localized and stable under posterior variation. Let $\mathcal{Q}$ contain pairs of positive candidates assigned to the same target with $r_i > r_j$. The mean-path ranking objective is

$$\mathcal{L}_{\mathrm{rank}} = \frac{\sum_{(i,j) \in \mathcal{Q}} (r_i - r_j)\, \log\big(1 + \exp(\ell_j^\mu - \ell_i^\mu)\big)}{\sum_{(i,j) \in \mathcal{Q}} (r_i - r_j)}, \tag{6}$$

in which larger reliability gaps receive greater weight. The ranking acts only on mean-path logits, supervising the deployed path to score stable, well-localized candidates above those that drift. Because $r_i$ is detached, localization itself is still optimized by $\mathcal{L}_{\mathrm{det}}$. Figure 4 illustrates how drift-discounted reliability supervises the ordering of candidates of the same target.

### 3.3 Training and Inference

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \beta_e \mathcal{L}_{\mathrm{rate}} + \lambda_{\mathrm{loc}} \mathcal{L}_{\mathrm{loc}} + a_e \big(\lambda_{\mathrm{con}} \mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{rank}} \mathcal{L}_{\mathrm{rank}}\big), \tag{7}$$

where $\beta_e$ ramps in the rate objective after warm-up, while $a_e$ schedules posterior sampling and the IAR objectives. Only the bottleneck parameters $\phi$ and the trainable parameters of $D_\psi$ are optimized; the invertible prefix, its reused partial inverse, and the latent calibration remain fixed. Coefficients, schedules, and numerical settings are given in Appendix 8.

At inference, $\widehat{\mathcal{Y}} = \mathrm{PostProcess}(D_\psi(z_\mu))$. Posterior sampling, full-image reconstruction, target masks, and reliability supervision are used only during training, while the partial inverse remains within $D_\psi$ to construct the finest-scale features. Deployment therefore follows a single deterministic mean path.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

We construct IR-DegradeX to study detection when several imaging effects act on the same weak target. Seven operators cover additive, multiplicative, and convolutional degradation: additive terms overlay interference on the target, multiplicative terms modulate its contrast, and convolutional terms spread its energy (Fig. 5). Each of 3,000 source images yields single-, paired-, and multi-family observations (L1–L3), and the three settings share one split of 2,400/300/300 source images for training, validation, and test, with manually reviewed validation and test annotations. Degradation composition therefore changes while the underlying scenes stay aligned, so the effect of increasingly compound degradation can be measured directly. A single model is trained on the 7,200 observations of the combined training set (LMix) and evaluated on each test set. We report COCO box AP over IoU 0.50:0.05:0.95, together with AP50 and AP75, each averaged over L1–L3. All baselines are retrained on LMix with their own configurations and evaluated on the same test sets. TDIB and its ablations share initialization, data, optimization, and validation-based checkpoint selection. Construction and implementation details are provided in the appendix.

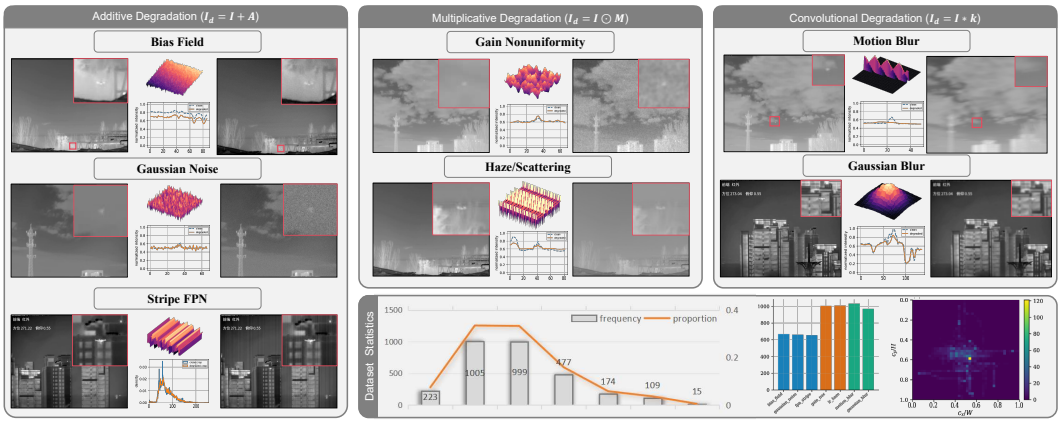


Figure 5: **IR-DegradeX.** Seven additive, multiplicative, and convolutional operators and their effects on weak target evidence. L1, L2, and L3 combine one, two, and three degradation families, respectively.

## 4.2 COMPARISON WITH REPRESENTATIVE METHODS

Table 1: Comparison on IR-DegradeX (%). All methods are single runs trained on LMix, with TDIB using seed 0; AP50 and AP75 are macro averages over L1–L3. †Degraded-only adaptation: the clear-image branch and its paired-clean losses are disabled.

| | AP by composition | | | Macro | | |
|---|---|---|---|---|---|---|
| Method | L1 | L2 | L3 | AP | AP50 | AP75 |
| ACM (Dai et al., 2021) | 32.47 | 32.62 | 31.83 | 32.31 | 82.79 | 16.84 |
| DNANet (Li et al., 2023) | 55.39 | 54.73 | 55.35 | 55.16 | 96.67 | 52.12 |
| UIU-Net (Wu et al., 2023) | 64.46 | 64.66 | 65.65 | 64.92 | 96.87 | 74.01 |
| ObjectBox (Zand et al., 2022) | 57.35 | 57.53 | 57.58 | 57.49 | 98.28 | 64.43 |
| RPCANet++ (Wu et al., 2025) | 55.71 | 56.54 | 56.68 | 56.31 | 84.07 | 64.93 |
| OSCAR (Dai et al., 2023) | 54.43 | 54.61 | 55.46 | 54.83 | 97.35 | 54.77 |
| JFD$^3$ † (Wang et al., 2026) | 66.72 | 66.94 | 66.87 | 66.84 | **98.66** | 76.79 |
| NS-FPN (Yuan et al., 2026) | 65.75 | 66.97 | 67.48 | 66.73 | 98.43 | **80.24** |
| TDIB | **68.07** | **67.91** | **68.59** | **68.19** | 98.03 | 76.02 |

Table 1 shows a clear progression in how degraded evidence is handled. Among methods that sustain weak responses through detector architecture, UIU-Net is the strongest at 64.92% macro AP. JFD$^3$ and NS-FPN explicitly counter degradation through feature restoration and purification and raise this to about 66.8%, showing the benefit of treating degradation explicitly. TDIB instead learns

what to retain rather than what to repair, and reaches 68.19%, 1.34 and 1.46 points above the two degradation-aware baselines. It achieves the highest AP on L1, L2, and L3 individually, so the gain persists from single to compound degradation. The same LMix model also exceeds separately trained L1-, L2-, and L3-specific TDIB models by 1.34, 0.38, and 1.57 points, respectively. A shared target-directed retention mechanism therefore benefits from heterogeneous degradation rather than requiring a separate policy for each composition. The gain is not explained by parameter count alone: UIU-Net and ObjectBox use 50.5M and 86.1M parameters, respectively, compared with 16.0M for TDIB, yet both perform substantially worse. Moreover, the ablations in Sec. 4.3 share the same backbone and detection head, isolating the contribution of the proposed components.

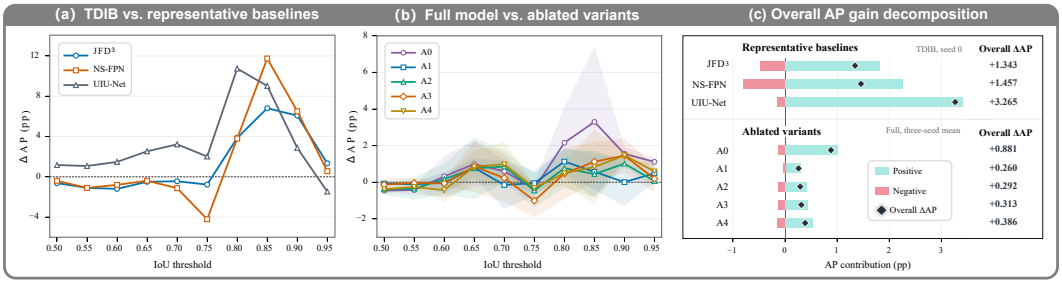


Figure 6: **AP differences across IoU thresholds.** (a) TDIB versus representative baselines. (b) Full versus ablated variants; lines are three-seed means and shading shows the standard deviation. (c) Decomposition of the overall AP difference into contributions from thresholds where the reference leads and trails.

JFD$^3$ and NS-FPN nevertheless attain the highest AP50 and AP75. Figure 6(a) resolves this tension over all ten COCO thresholds. TDIB trails both slightly through IoU 0.75 and overtakes them from 0.80 onward, with the largest margins around 0.85–0.90; against UIU-Net, it is better at nine of the ten thresholds. The overall advantage is the net of these contributions: gains at strict thresholds outweigh small losses at loose ones (Fig. 6(c)). For few-pixel targets, small boundary errors incur a disproportionately large IoU penalty, making strict thresholds particularly sensitive to how well target structure is retained. The shift toward strict-overlap gains is consistent with TDIB's design, which selects what to retain from a recoverable representation while constraining target-region distortion against the observation itself.

### 4.3 Ablation Study and Discussion

Table 2: Ablations and mechanism diagnostics. Values are means with standard deviations over three detector seeds. A0 removes the bottleneck and all objectives built on it. AP is measured on the test sets and diagnostics on the validation set; Recon. is measured against the degraded input, and Agr. denotes mean–sample assignment agreement.

| | Components | | | | | Target retention | | Posterior stability | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Wt. rate | Loc. fid. | Con. | Rank | AP↑ (%) | Recon.↓ ($10^{-3}$) | $\sigma$ ↓ ($10^{-2}$) | JS↓ ($10^{-6}$) | Jitter↓ ($10^{-2}$ px) | Agr.↑ (%) |
| A0 | | | | | $67.32_{\pm 0.67}$ | – | – | – | – | – |
| A1 | | ✓ | ✓ | ✓ | $67.94_{\pm 0.19}$ | $1.99_{\pm 0.01}$ | $1.36_{\pm 0.06}$ | $5.64_{\pm 0.21}$ | $1.40_{\pm 0.06}$ | $97.1_{\pm 0.1}$ |
| A2 | ✓ | | ✓ | ✓ | $67.91_{\pm 0.13}$ | $63.40_{\pm 6.11}$ | $6.76_{\pm 0.36}$ | $12.91_{\pm 0.34}$ | $2.40_{\pm 0.22}$ | $95.0_{\pm 0.4}$ |
| A3 | ✓ | ✓ | | | $67.89_{\pm 0.28}$ | $2.31_{\pm 0.04}$ | $1.67_{\pm 0.26}$ | $35.39_{\pm 13.93}$ | $2.56_{\pm 0.60}$ | $95.5_{\pm 1.1}$ |
| A4 | ✓ | ✓ | ✓ | | $67.82_{\pm 0.29}$ | $2.04_{\pm 0.05}$ | $1.64_{\pm 0.39}$ | $11.03_{\pm 7.95}$ | $1.83_{\pm 0.67}$ | $96.7_{\pm 1.3}$ |
| Full | ✓ | ✓ | ✓ | ✓ | $\mathbf{68.20}_{\pm 0.07}$ | $\mathbf{1.97}_{\pm 0.02}$ | $\mathbf{1.32}_{\pm 0.04}$ | $5.65_{\pm 0.18}$ | $\mathbf{1.36}_{\pm 0.01}$ | $\mathbf{97.3}_{\pm 0.2}$ |

Table 2 removes one design element at a time and pairs detection with diagnostics of what each element is meant to change: target reconstruction and posterior scale for TDSR, and prediction stability under posterior sampling for IAR. All variants are reported over three detector seeds, and Full improves on every variant for each matched seed; reading the relatively small detection differences together with these diagnostics shows how each component changes the representation and the predictions.

A0 retains the frozen invertible prefix and detector but removes the bottleneck and all objectives built on it, and therefore passes on everything the prefix preserves. Introducing explicit selection

over the recoverable representation improves AP by 0.88 points and makes the result markedly more consistent across seeds. The advantage over every variant is more apparent at stricter IoU thresholds, and against A0 it holds for all three seeds (Fig. 6(b)). Recoverability therefore provides the reference from which information can be selected, but does not by itself determine what should be retained for detection.

A1 and A2 isolate the two constraints of TDSR. A1 replaces the target-relaxed rate with uniform weights while keeping local inverse fidelity. Its target reconstruction, posterior scale, and prediction stability remain essentially those of Full, yet AP drops. The benefit of spatial weighting is thus not a by-product of better reconstruction or lower posterior variation; it lies in where compression pressure is allocated, which lets detection supervision spend more representation capacity around targets.

A2 reveals the converse. With the target-relaxed rate still active, removing local inverse fidelity raises the target reconstruction error more than thirtyfold and the target posterior scale fivefold. The effect reaches the detector: predictions under posterior sampling become markedly less stable, and assignment agreement between the two paths is the lowest among all variants. Relaxing compression around a target makes its information cheaper to retain, but cannot ensure that the retained evidence remains faithful under posterior variation. The fixed inverse supplies that missing constraint. Spatial weighting thus determines where retention capacity is allocated, local inverse fidelity constrains the target structure preserved there, and detection supervision determines which retained evidence is useful.

This fidelity concerns target evidence rather than the removal of degradation. Probes trained on frozen representations show essentially unchanged target readability across the bottleneck, while degradation-family information remains readily decodable and is not consistently reduced after selection. TDIB therefore does not require degradation information to disappear; its gain comes from preserving task-useful target evidence within a representation that may still contain identifiable degradation.

Finally, A3, A4, and Full show how IAR uses posterior variation. Adding consistency to A3 reduces classification divergence by about two thirds and box jitter by nearly a third, yet AP does not rise: consistency stabilizes predictions under posterior variation but does not determine how candidates should be ordered. Adding drift-discounted ranking then raises AP by 0.39 points to that of Full. The two objectives thus play distinct roles: consistency suppresses unwanted variation between the posterior paths, whereas ranking converts posterior-derived reliability into relative supervision for the mean-path scores used at inference.

## 5 Conclusion

This work introduces TDIB, which reframes degraded infrared small-target detection as learning what to retain from a recoverable representation rather than what to restore. TDSR confines information loss to an explicit bottleneck: spatially weighted compression allocates retention capacity toward fragile target regions, while the fixed inverse constrains the target-region distortion introduced by selection without requiring paired clean references. IAR further uses posterior variation to align training with deterministic mean-path inference, with prediction consistency stabilizing candidate responses and reliability-aware ranking supervising their relative ordering. Experiments on IR-DegradeX show that TDIB outperforms both full preservation and the compared restoration- and purification-based alternatives, with its advantage particularly visible under strict localization criteria. The ablations reveal complementary roles: the weighted rate determines where retention capacity is allocated, inverse fidelity determines how faithfully and stably target evidence survives there, and detection supervision determines which retained evidence is useful. Representation probes further show that such evidence can be preserved without requiring identifiable degradation to disappear.

These results position recoverability not as the objective itself, but as a reference that makes task-directed information loss controllable. Because the fidelity reference is the degraded observation itself, the same principle can be extended to sensor-specific real degradations and temporal infrared observations, where what should be retained may vary across domains and frames. More broadly, degraded visual perception stands to benefit less from preserving or repairing everything than from learning which evidence the task truly needs to retain.

## AI USE STATEMENT

Generative AI tools assisted with code drafting, manuscript planning and editing, literature lookup, interpretation of existing experimental summaries, and preparation of table-generation and LaTeX code. The research workflow was primarily human-led, including method development, data preparation, experiments, and figure preparation. The authors take responsibility for the final content and artifacts.

## REPRODUCIBILITY STATEMENT

Section 3 describes the model and its design. Appendix 6 specifies the construction, annotation, and split of IR-DegradeX together with the common evaluation protocol, and Appendix 7 gives the complete objectives with their numerical conventions. Appendix 8 describes the architecture, training schedules, coefficient calibration, and the training settings and interfaces of the external baselines. Appendix 9 reports seed-level, per-family, training-scope, and validation-locked operating-point results, and Appendices 10 and 11 specify the diagnostic aggregation, uncertainty intervals, and representation readouts.

## REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

Lynton Ardizzone, Radek Mackowiak, Carsten Rother, and Ullrich Köthe. Training normalizing flows with the information bottleneck for competitive generative classification. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/593906af0d138e69f49d251d3e7cbed0-Abstract.html.

Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 950–959, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Dai_Asymmetric_Contextual_Modulation_for_Infrared_Small_Target_Detection_WACV_2021_paper.html.

Yimian Dai, Xiang Li, Fei Zhou, Yulei Qian, Yaohong Chen, and Jian Yang. One-stage cascade refinement networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. doi: 10.1109/TGRS.2023.3243062. URL https://arxiv.org/abs/2212.08472.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1605.08803.

Houzhang Fang, Xiaolin Wang, Zengyang Li, Lu Wang, Qingshan Li, Yi Chang, and Luxin Yan. Detection-friendly nonuniformity correction: A union framework for infrared UAV target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11898–11907, 2025. doi: 10.1109/CVPR52734.2025.01111.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. TOOD: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3510–3519, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Feng_TOOD_Task-Aligned_One-Stage_Object_Detection_ICCV_2021_paper.html.

Bo Huang, Jianan Li, Junjie Chen, Gang Wang, Jian Zhao, and Tingfa Xu. Anti-UAV410: A thermal infrared benchmark and customized scheme for tracking drones in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2852–2865, 2024. doi: 10.1109/TPAMI.2023.3335338. URL https://ieeexplore.ieee.org/document/10325629/.

Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.07088.

Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Computer Vision – ECCV 2018*, pp. 784–799, 2018.

Durk P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1807.03039.

Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2023. URL https://arxiv.org/abs/2106.00487.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/f0bda020d2470f2e74990a07a607ebd9-Abstract.html.

Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11632–11641, 2021.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5a66b9200f29ac3fa0ae244cc2a51b39-Abstract.html.

Ji Liu, Dong Li, Rongzhang Zheng, Lu Tian, and Yi Shan. RankDetNet: Delving into ranking constraints for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 264–273, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Liu_RankDetNet_Delving_Into_Ranking_Constraints_for_Object_Detection_CVPR_2021_paper.html.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999. URL https://www.princeton.edu/~wbialek/our_papers/tishby%2Bal_99.pdf.

Xiaolin Wang, Houzhang Fang, Qingshan Li, Lu Wang, Yi Chang, and Luxin Yan. Blur-robust detection via feature restoration: An end-to-end framework for prior-guided infrared UAV target detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10181–10189, 2026. doi: 10.1609/aaai.v40i12.37986. URL https://ojs.aaai.org/index.php/AAAI/article/view/37986.

Ziwei Wang, Ziyi Wu, Jiwen Lu, and Jie Zhou. BiDet: An efficient binarized object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2049–2058, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_BiDet_An_Efficient_Binarized_Object_Detector_CVPR_2020_paper.html.

Fengyi Wu, Yimian Dai, Tianfang Zhang, Yixuan Ding, Jian Yang, Ming-Ming Cheng, and Zhenming Peng. RPCANet++: Deep interpretable robust PCA for sparse object segmentation. *arXiv preprint arXiv:2508.04190*, 2025. URL https://arxiv.org/abs/2508.04190.

Xin Wu, Danfeng Hong, and Jocelyn Chanussot. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing*, 32:364–376, 2023. doi: 10.1109/TIP.2022.3228497. URL https://arxiv.org/abs/2212.00968.

Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *Computer Vision – ECCV 2020*, pp. 126–144, 2020.

Shule Yan, Zetian Zhang, Xiao Ma, and Zexuan Ji. Target-aware invertible encoder with reconstruction guidance for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32714–32723, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yan_Target-Aware_Invertible_Encoder_with_Reconstruction_Guidance_for_Infrared_Small_Target_CVPR_2026_paper.html.

Maoxun Yuan, Duanni Meng, Ziteng Xi, Tianyi Zhao, Shiji Zhao, Yimian Dai, and Xingxing Wei. Seeing through the noise: Improving infrared small target detection and segmentation from noise suppression perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27783–27792, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yuan_Seeing_Through_the_Noise_Improving_Infrared_Small_Target_Detection_and_CVPR_2026_paper.html.

Mohsen Zand, Ali Etemad, and Michael Greenspan. ObjectBox: From centers to boxes for anchor-free object detection. In *Computer Vision – ECCV 2022*, pp. 390–406, 2022. doi: 10.1007/978-3-031-20080-9_23. URL https://link.springer.com/chapter/10.1007/978-3-031-20080-9_23.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. VarifocalNet: An IoU-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8514–8523, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VarifocalNet_An_IoU-Aware_Dense_Object_Detector_CVPR_2021_paper.html.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12993–13000, 2020. doi: 10.1609/aaai.v34i07.6999. URL https://ojs.aaai.org/index.php/AAAI/article/view/6999.