Target-Aware Invertible Encoder with Reconstruction Guidance for Infrared Small Target Detection

Supplementary Material

6. Detailed Design of the GCTM

This appendix provides a comprehensive explanation of the GCTM, detailing the design rationale, empirical validation, and practical implementation of its gray-term component. We systematically address the limitations of geometry-only measures and demonstrate how radiometric consistency enhances detection robustness in infrared small target scenarios.

6.1. Discussion on Radiometric Descriptors

The fundamental limitation of purely geometric measures like IoU and TAM is their inability to distinguish between predictions that share similar spatial characteristics but capture radiometrically distinct patterns. In ISTD, this manifests as false alarms from background structures that happen to have similar sizes and locations as genuine targets. To address this, GCTM introduces a gray term that explicitly measures appearance consistency between prediction and ground truth.

We begin with a dataset-level correlation analysis on DUAB, evaluating four candidate radiometric descriptors:

- Global SNR ($\mathrm{SNR}_{\mathrm{glob}}$) and global entropy ($H_{\mathrm{glob}}$) computed over the full image
- Local SNR in the target neighborhood (LSNR) and local entropy (H_{loc}) computed within the ground-truth patch

Fig. 7(a) presents the Pearson correlation matrix, revealing critical insights for descriptor selection:

Global descriptors exhibit redundancy: The strong negative correlation between global SNR and global entropy ($\rho \approx -0.93$) indicates these measures capture the same underlying phenomenon—high-SNR images tend to have concentrated histograms and clean backgrounds, while low-SNR scenes show diffuse distributions and cluttered backgrounds. This redundancy makes either descriptor sufficient for characterizing overall image quality.

Local descriptors provide complementary information: The moderate correlation between LSNR and global entropy (used as a proxy for background entropy, $\rho \approx -0.35$) and the weak correlations involving local entropy (absolute correlations < 0.2) indicate that target-level saliency and background complexity are complementary rather than redundant aspects of detection difficulty. This non-redundancy suggests that neither descriptor alone can adequately capture the varied challenges in infrared scenes.

We also evaluated the signal-to-clutter ratio (SCR) but found its correlation pattern unstable across scenes due to its sensitivity to global mean intensity, particularly in non-uniform backgrounds or multi-target scenarios.

These findings motivate our selection of **local SNR** to quantify target saliency against immediate surroundings and **background entropy** to measure local scene complexity. This dual-descriptor approach, based on LSNR for target saliency and background entropy for scene complexity, adapts to varying detection difficulties while providing complementary information for robust evaluation.

6.2. Adaptive Tolerance Based on Scene Difficulty

Guided by the descriptor analysis, we design the gray term as a Bhattacharyya similarity modulated by a radiometry-aware tolerance:

$$\mathbb{S}_{\mathrm{gray}} = \frac{\mathrm{BC}(\mathcal{P}_{\mathrm{gt}}, \mathcal{P}_{\mathrm{pr}})}{t_{\mathrm{gray}}}, \qquad t_{\mathrm{gray}} = \frac{\mathrm{LSNR}(\mathcal{P}_{\mathrm{gt}})}{1 + H_{\mathrm{bg}}} + \varepsilon, \quad (12)$$

where $\mathrm{BC}(\cdot,\cdot) \in [0,1]$ denotes the Bhattacharyya coefficient between normalized histograms, $\mathrm{LSNR}(\mathcal{P}_{\mathrm{gt}})$ measures target saliency, and H_{bg} quantifies background complexity. In practice, we approximate H_{bg} by the global entropy computed over the whole image and apply clipping of t_{gray} to [0.05, 1.0] for numerical stability.

This formulation ensures adaptive behavior: for identical histogram similarity, larger t_{gray} yields stricter evaluation (smaller $\mathbb{S}_{\mathrm{gray}}$), while smaller t_{gray} provides more lenient scoring.

Fig. 7(b) characterizes the empirical behavior of t_{gray} on DUAB. The distribution after clipping shows most values in [0.05, 1.0] with mean 0.50 and median 0.46, indicating well-behaved dynamic range. The strong positive correlation with LSNR ($\rho \approx 0.94$) and moderate negative correlation with global entropy as a proxy for background entropy ($\rho \approx -0.53$) validate the intended design:

- **High-saliency, clean backgrounds**: When targets are radiometrically strong (large LSNR) and backgrounds are simple (small H_{bg}), t_{gray} becomes large, enforcing strict radiometric consistency—only proposals with histograms closely matching ground truth receive high scores.
- **Low-saliency, cluttered backgrounds**: When targets are weak (small LSNR) and backgrounds are complex (large H_{bg}), t_{gray} decreases, providing tolerance to avoid over-penalizing proposals in intrinsically difficult scenes.

This adaptive behavior encodes an inverse relationship between target saliency and detection difficulty, enforcing stricter radiometric consistency for strong targets in simple

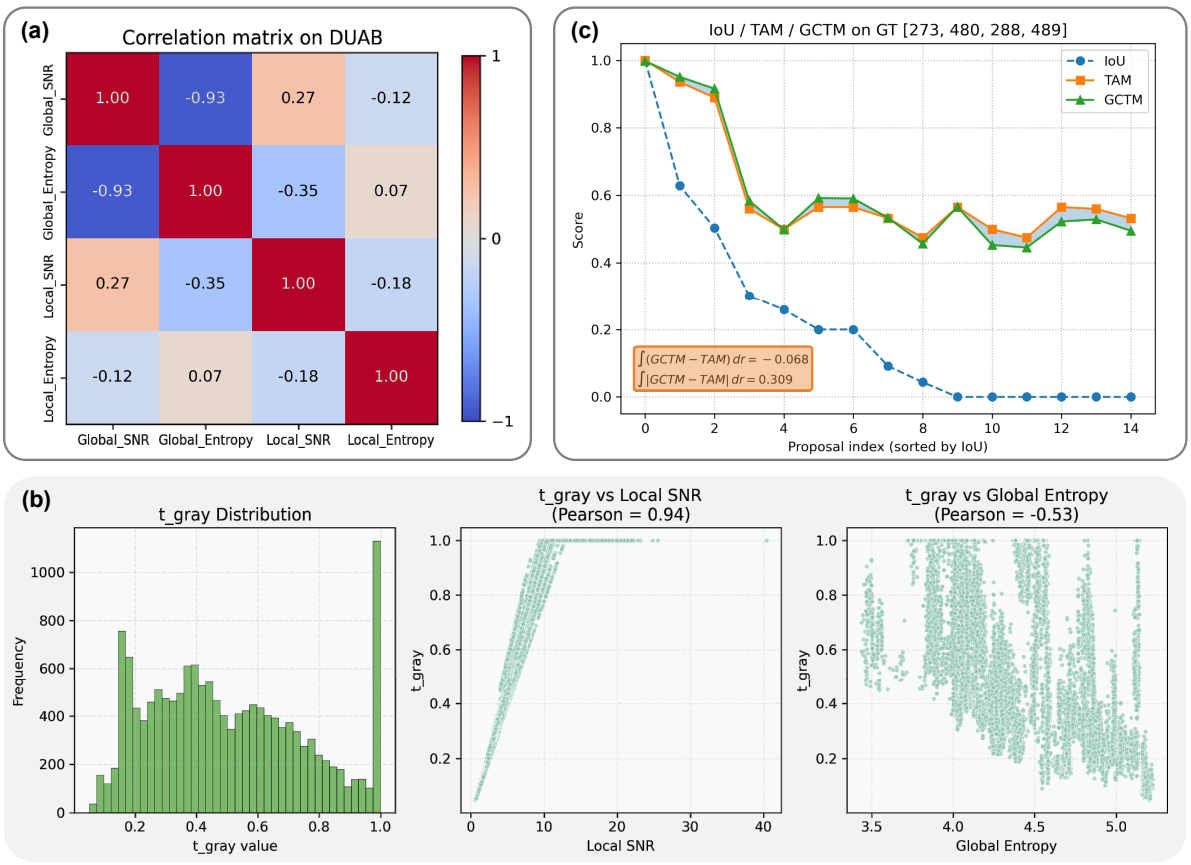

Figure 7. Comprehensive analysis of the GCTM design. (a) Pearson correlation matrix of radiometric descriptors on DUAB, showing the complementary relationship between LSNR and global entropy (used as a proxy for background entropy). (b) Empirical behavior of the radiometry-aware tolerance t_{gray}: distribution after clipping to [0.05, 1.0] (left), and its relationships with LSNR (middle) and global entropy as a proxy for background entropy (right). (c) Proposal re-ranking effect for a representative ground-truth box with coordinates [273, 480, 288, 489], with proposals sorted by IoU in descending order. The filled area between TAM and GCTM curves quantifies the cumulative adjustment, with signed area -0.068 and absolute area 0.309.

scenes and more tolerant scoring for weak targets in cluttered scenes.

6.3. Balanced Integration

The final GCTM score integrates geometric and radiometric information through adaptive interpolation:

$$\text{GCTM} = \lambda\,\mathbb{S}_{\text{geo}} + (1-\lambda)\,\mathbb{S}_{\text{gray}}, \qquad \lambda = \sigma\big(\mathbb{S}_{\text{geo}}/\tau\big), \quad (13)$$

where $\sigma(\cdot)$ is the logistic function and $\tau > 0$ is a temperature parameter.

This design ensures appropriate balance between spatial and appearance cues:

- When proposals are geometrically well-aligned ($\mathbb{S}_{\text{geo}}$ large), λ approaches 1 and GCTM is dominated by the geometry term, preventing over-penalization of well-localized targets even in cluttered scenes.
- When geometry is unreliable ($\mathbb{S}_{\text{geo}}$ moderate or small), λ decreases and the gray term gains influence, providing crucial discrimination among geometrically similar but radiometrically divergent proposals.

The geometry-driven weighting ensures that radiometric consistency serves as a refinement rather than a replacement for spatial alignment. In this way, GCTM maintains the fundamental importance of localization quality while adding the necessary discrimination capability.

6.4. Comparison on Proposal Re-ranking

From Fig. 7(c), the analysis reveals three key patterns:

Geometric baseline preservation: The IoU curve (dashed line) provides a purely geometric baseline, while TAM and GCTM curves (solid lines) both follow the overall IoU trend, maintaining high correlation with conventional localization quality. This confirms that GCTM preserves the fundamental geometric consistency principle.

Systematic radiometric discrimination: Despite the geometric correlation, GCTM systematically suppresses geometrically plausible proposals whose gray-level statistics

deviate from the ground-truth patch. The quantitative analysis shows a signed area of -0.068 between GCTM and TAM curves, indicating overall downward adjustment, and an absolute area of 0.309, revealing substantial re-ranking magnitude.

Targeted score redistribution: The re-ranking is non-uniform across the proposal spectrum. For proposals with moderate IoU values, GCTM introduces significant deviations from TAM, selectively penalizing candidates that are geometrically adequate but radiometrically inconsistent. This targeted suppression is particularly valuable in infrared small target detection, where multiple background structures often share similar spatial characteristics with genuine targets but exhibit different thermal signatures.

This controlled but substantial re-ranking effectively breaks the degeneracy among geometrically similar proposals while preserving the overall ordering induced by geometric quality, making GCTM particularly suited for infrared small target detection in cluttered environments.

7. Loss Formulation and Training Strategy

This section details the loss functions and optimization design that instantiate the gradient-decoupled training scheme outlined in the main text. The emphasis is on the target-aware reconstruction loss and its integration with the detection objective.

7.1. Implementation Configuration

Training employs distributed data parallelism with an effective batch size of 32 and automatic mixed precision. Core hyperparameters are summarized in Tab. 4 and Tab. 5. Data augmentation follows an SSD-style pipeline with photometric distortions and geometric transformations tailored for infrared imagery.

All experiments are conducted on a Linux server with two NVIDIA GeForce RTX 4090 GPUs (24 GB each) and dual Intel Xeon Gold 6426Y CPUs. The implementation uses PyTorch 1.14.0 with CUDA 11.8 and cuDNN 8.7.

7.2. Detection Objective

The detection branch employs a CenterNet-style anchor-free head with an enhanced focal loss that incorporates spatial weighting for infrared small targets. The classification loss $\mathcal{L}_{\text{cls}}$ extends standard focal loss through a focal-area weighting scheme:

- Standard focal loss reduces the weight of well-classified samples through $(1-p)^{\alpha}$ and p^{α} terms
- The focal-area enhancement adds a spatial prior: pixels far from target centers are treated as easy negatives and are further suppressed, forcing the model to focus on challenging regions such as target edges and blurred areas

In our implementation, the Gaussian heatmap serves as a soft spatial weight, producing smooth weight decay with distance from target centers. This provides finer handling of dense targets compared to hard boundary methods.

Table 4. Core hyper-parameters of InvDet.

Item	Value / setting
Optimization	
Optimizer	AdamW
Base learning rate	1×10^{-4}
LR schedule	Plateau on val mAP
Batch size	32
Mixed precision	Enabled
Reconstruction	
λ_{fit}	0.5
λ_{tv}	0.1
λ_{ce}	0
Reconstruction warmup	50 epochs
Reversible depth	2
InvBlocks per stage	[2,2,2,2]

Table 5. Data and augmentation settings for DUAB.

Item	Value / setting
Photometric distortion	
Brightness jitter	$\Delta \in [-32, 32]$
Contrast jitter	Scale factor $\in [0.5, 1.5]$
Saturation jitter	Scale factor $\in [0.5, 1.5]$
Hue jitter	Shift $\in [-18°, 18°]$
Geometric augmentation	
Random sample crop	SSD-style sampling
Expand	Scale $\in [1.0, 4.0]$
Horizontal flip	Probability 0.5
Resize	256×256
Normalization	$\mu = 0.343$, $\sigma = 0.155$

The overall detection loss combines:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{quality}}, \tag{14}$$

where $\mathcal{L}_{\text{reg}}$ uses Smooth-L_1 loss for center offsets and box sizes, and $\mathcal{L}_{\text{quality}}$ includes IoU regression and IoU-aware classification with small weights ($w_{\text{iou}} = w_{\text{aware}} = 0.25$).

7.3. Target-Aware Reconstruction Loss

The core innovation lies in the reconstruction loss design, which focuses supervision on semantically meaningful regions through GCTM-guided weighting.

Weight map generation. The GCTM-based weight generator produces multi-scale maps $\{W_s\}$ where high values concentrate around infrared targets. The full-resolution map $W_0 \in [0,1]^{H \times W}$ serves as primary guidance for reconstruction.

ROI-weighted fidelity term. The main reconstruction fidelity employs a weighted ℓ_1 loss:

$$\mathcal{L}_{\text{fit}} = \lambda_{\text{fit}} \cdot \frac{\|W_0 \odot (\mathbf{X}^{\text{rec}} - \mathbf{X})\|_1}{\|W_0\|_1 + \varepsilon} + \beta_{\text{img}} \|\mathbf{X}^{\text{rec}} - \mathbf{X}\|_1, \quad (15)$$

where $\beta_{\text{img}} = 10^{-4}$ prevents degenerate solutions when W_0 is sparse.

Background regularization. To suppress noise in background regions while preserving target edges, we apply anisotropic total variation on an eroded background mask:

$$\mathcal{L}_{\text{tv}} = \lambda_{\text{tv}} \cdot \frac{\|M_{\text{bg}} \odot (|\nabla_x \mathbf{X}^{\text{rec}}| + |\nabla_y \mathbf{X}^{\text{rec}}|)\|_1}{\|M_{\text{bg}}\|_1 + \varepsilon}, \quad (16)$$

where $M_{\text{bg}} = \text{erode}(1 - W_0)$ with a 3×3 kernel.

Progressive activation. The reconstruction loss is gradually activated through a cosine ramp synchronized with TARM:

$$\mathcal{L}_{\text{rec}}^{\text{eff}} = r_s \cdot (\mathcal{L}_{\text{fit}} + \mathcal{L}_{\text{tv}}), \quad (17)$$

where r_s starts from 0 and reaches 1 over 50 epochs, preventing conflicts during early training.

7.4. Gradient-Decoupled Optimization

The training strategy employs explicit gradient separation between detection and reconstruction pathways.

Parameter partitioning. Model parameters are divided into disjoint sets:

- Θ_{det}: detection neck, fusion modules, and prediction heads
- Θ_{rec}: invertible encoder and TARM components

Optimization setup. Two AdamW optimizers operate independently:

$$\Theta_{\text{det}} \leftarrow \text{AdamW}(\Theta_{\text{det}}, \nabla_{\Theta_{\text{det}}} \mathcal{L}_{\text{det}}) \quad (18)$$

$$\Theta_{\text{rec}} \leftarrow \text{AdamW}(\Theta_{\text{rec}}, \nabla_{\Theta_{\text{rec}}} \mathcal{L}_{\text{rec}}^{\text{eff}}) \quad (19)$$

Feature detachment. Features from the invertible encoder to the detection neck are detached:

$$\mathbf{F}_{\text{det}} = \text{stop\_grad}(\text{IRN}(\mathbf{X})), \quad (20)$$

ensuring $\mathcal{L}_{\text{det}}$ gradients do not propagate into Θ_{rec}.

Joint training protocol. A single forward-backward pass computes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{rec}}^{\text{eff}}, \quad (21)$$

with automatic gradient routing to respective parameter sets. This maintains computational efficiency while allowing reconstruction to enhance representation quality without destabilizing detection training.

8. Quantitative & Qualitative Results

8.1. The mAP Analysis

Table 6. The mAP on five datasets. Mean column highlighted for overall comparison. Rec/Pre/F1 are reported in Tab. 1.

Method	[33]	[3]	[12]	[16]	[9]-P	[9]-S	[9]-E	Mean
MDFAcGAN	46.7	65.1	55.7	26.4	45.1	62.6	58.0	51.4
ACM	41.9	5.2	52.8	7.3	9.1	29.5	45.0	27.3
ObjectBox	86.7	84.6	54.7	69.1	42.2	19.0	58.6	59.3
YOLO-FR	72.5	2.6	46.9	97.9	58.7	30.0	52.6	51.6
UIU-Net	53.3	58.1	59.6	65.7	6.6	56.1	60.3	51.4
DNA-Net	44.6	72.6	54.7	64.4	63.5	67.3	94.4	65.9
RDIAN	9.9	37.2	18.9	67.2	44.3	32.0	34.9	34.9
OSCAR	39.9	68.1	69.0	92.5	73.1	73.5	86.3	71.8
MA-Net	86.5	85.1	94.2	96.9	95.5	87.1	97.5	**91.8**
Ours	78.0	80.6	80.1	95.3	98.8	86.9	97.9	*88.2*

Tab. 6 reports mAP for all competing methods on the five benchmarks used in the main paper. Overall, the method-wise trends are consistent with the Rec/Pre/F1 results in Tab. 1: methods that are strong in F1 also achieve high mAP, while weaker baselines (e.g., ACM and RDIAN) remain clearly behind across datasets. MA-Net attains the highest mean mAP (91.8%), with InvDet (Ours) obtaining the second-best mean score (88.2%) while maintaining competitive or superior F1 in the main tables. Concretely, InvDet achieves the best AP on DUAB-Point and DUAB-Extended and ranks among the top performers on NUDT-SIRST, IRSTD, and DUAB-Spot, demonstrating that the proposed invertible design improves not only the chosen operating point but also the overall precision–recall trade-off captured by mAP.

8.2. Evolution of Reconstruction and Detection

Figure 8 provides a fine-grained visualization of the interplay between feature reconstruction and detection optimization throughout the training process. The crucial features within the reconstruction path (LP, HP, $\mathbf{W}_s$, etc.) evolve systematically: in early epochs, the weight map $\mathbf{W}_s$ is diffuse, leading to a noisy reconstruction and low-confidence, misaligned predictions (e.g., Epoch 1: Conf=0.09, IoU=0.00). As training progresses, the GCTM-driven $\mathbf{W}_s$ becomes sharply concentrated on the target region, which in turn guides the LP and HP branches to reconstruct cleaner target signatures and suppress background textures. This is directly reflected in the improving quality of the detection results (shown in the zoomed-in insets), where both detection confidence and localization accuracy (IoU) rise significantly. Notably, the GCTM score, which fuses geometric and appearance cues, converges rapidly to a high value (e.g., 0.89 by Epoch 10) and remains stable, underscoring its effectiveness as a robust optimization objective compared to the more volatile IoU. This progression

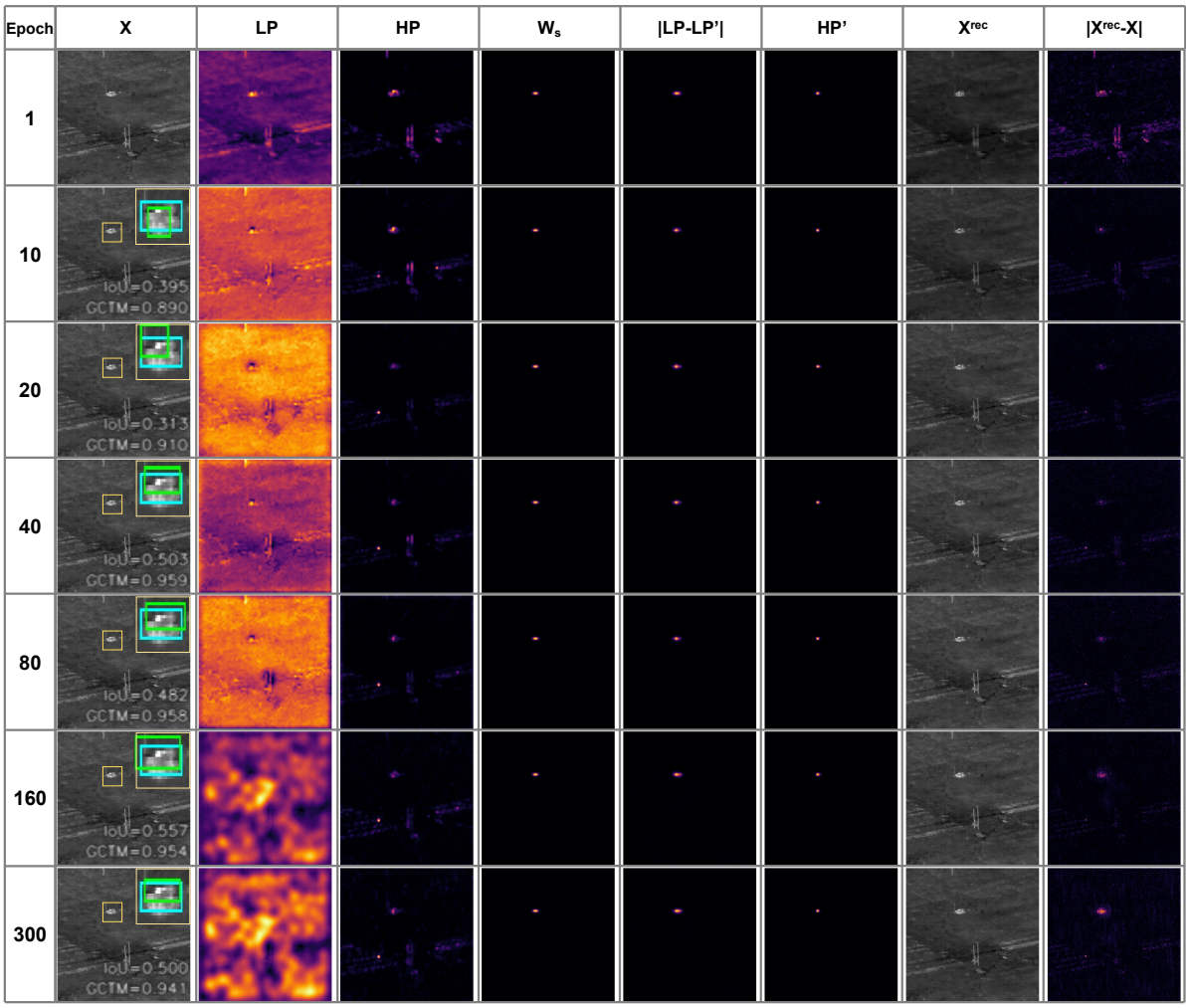

Figure 8. Evolution of reconstruction features and detection confidence across training epochs (1, 10, 20, 40, 80, 160, 300). From left to right: LP, HP, weight map $\mathbf{W}_s$, LP reconstruction error $|\text{LP} - \text{LP}'|$, HP reconstruction HP′, reconstructed image $\mathbf{X}^{\text{rec}}$, and final error $|\mathbf{X}^{\text{rec}} - \mathbf{X}|$. The zoomed regions (top-right) in image $\mathbf{X}$ show progressive refinement of detection results, with confidence scores, IoU, and GCTM scores annotated (bottom-right). The convergence of $\mathbf{W}_s$ and improving reconstruction fidelity correlate with detection accuracy.

empirically validates that the reconstruction path acts as a powerful regularizer, enabling the feature encoder to learn a representation that is both information-preserving for reconstruction and discriminative for detection.

8.3. Performance on Challenging Scenarios

The robustness of InvDet is further stressed-tested on a curated set of challenging scenarios in Figure 9. These samples exhibit complex backgrounds (e.g., structured cloud layers and sea clutter), strong interfering objects, and targets with extremely low signal-to-noise ratios. Despite these adversities, our method successfully localizes the targets with high precision and minimal false alarms. The qualitative results demonstrate the practical value of the proposed framework: the target-aware reconstruction guidance enforced by TARM and GCTM equips the model with a strong prior to distinguish faint target cues from deceptive background structures. This capability is crucial for deploying reliable infrared search and track systems in real-world environments where such challenging conditions are prevalent.

9. Cross-Dataset Generalization

We evaluate the generalization of InvDet through cross-dataset experiments. Three benchmarks with diverse characteristics are selected: IRSTD-1K (512×512, 1001 real images), NUAA-SIRST (256×256, 427 real images), and NUDT-SIRST (256×256, 1327 synthetic images). These datasets differ substantially in resolution, scale, imaging source (real vs. synthetic), and target/background distributions.

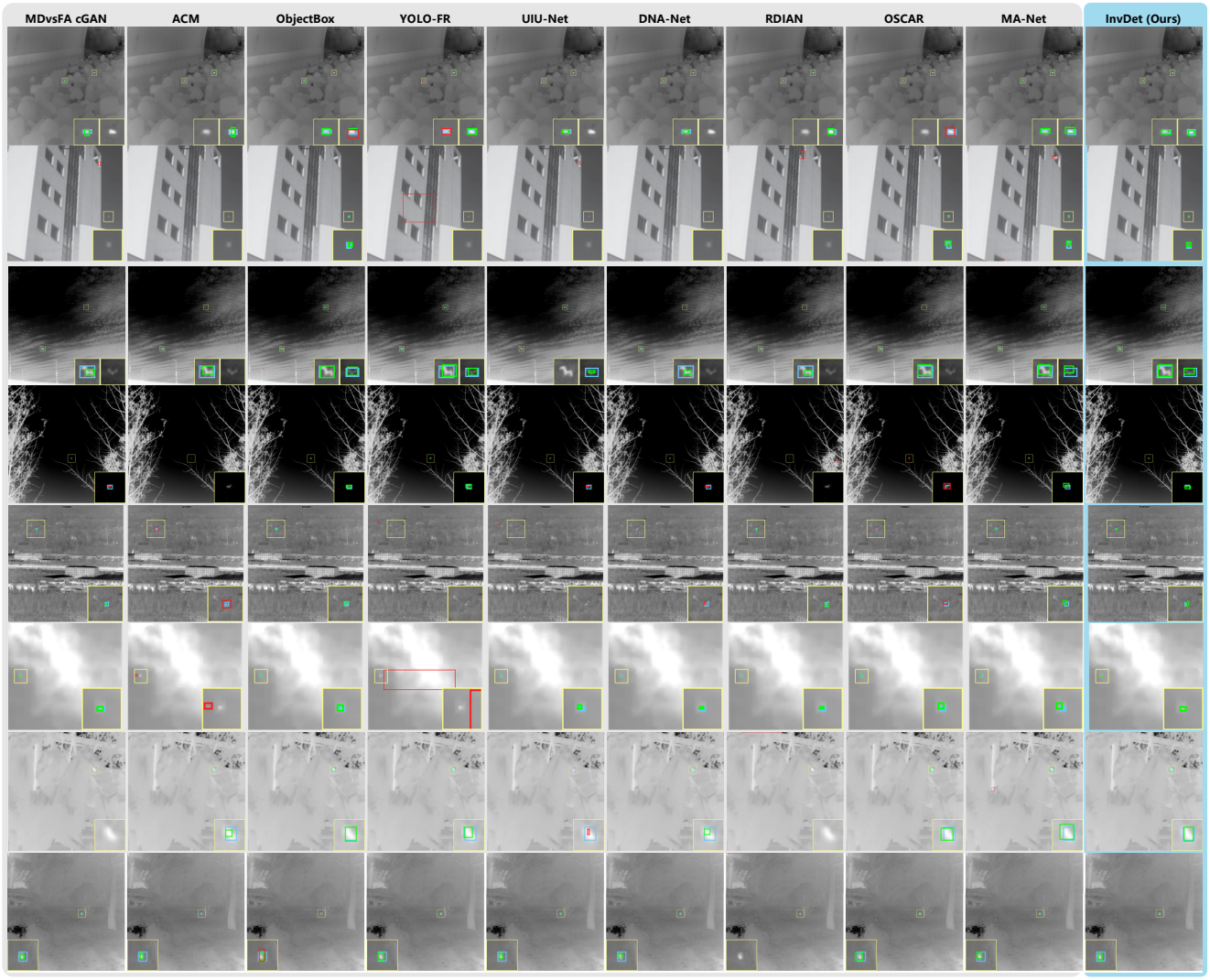

Figure 9. Qualitative results on challenging scenarios. Eight difficult cases showcase robustness under complex backgrounds, varying illumination, and strong interference. Correct detections (green boxes), false alarms (red boxes), and ground truths (blue boxes) are highlighted. Despite adverse conditions including structured clutter, low contrast, and competing distractors, the proposed method maintains reliable detection with minimal false positives.

9.1. Cross-Dataset Evaluation Protocol

For each training dataset $\mathcal{A}$, the best checkpoint is directly evaluated on the test split of every dataset $\mathcal{B}$ without any fine-tuning, using the target domain's preprocessing (mean/std normalization and input resolution). This yields a 3×3 cross-evaluation matrix (Tab. 7).

We further define the *F1 retention rate* as:

$$\mathrm{Ret}(\mathcal{A}\to\mathcal{B}) = \frac{\mathrm{F1}_{\mathcal{A}\to\mathcal{B}}}{\mathrm{F1}_{\mathcal{B}\to\mathcal{B}}} \times 100\%, \tag{22}$$

which measures the fraction of in-domain performance preserved under domain shift.

Analysis. InvDet achieves a mean cross-domain F1 retention of 84.9%, demonstrating strong generalization of the learned representation across domains. Three observations are notable:

Table 7. Cross-dataset generalization (F1 %). Rows: training set; columns: test set. **Bold**: in-domain. Subscripts: retention rate (%). Mean cross-domain retention: **84.9%**.

Train \ Test	IRSTD-1K	NUAA-SIRST	NUDT-SIRST	Mean
IRSTD-1K	**84.4**	77.8 89.1	74.3 86.1	78.8
NUAA-SIRST	74.3 88.0	**87.4**	75.3 87.4	79.0
NUDT-SIRST	63.7 75.5	72.6 83.1	**86.2**	74.2

Real → real transfer is robust. IRSTD-1K and NUAA-SIRST differ in resolution (512^2 vs. 256^2) and scale (1001 vs. 427 images), yet models trained on either retain 88–90% F1 when evaluated on the other. This indicates the invertible encoder captures resolution- and scene-agnostic information-preservation patterns rather than dataset-specific features.

Synthetic → real gap is bounded. NUDT-SIRST contains synthetic imagery with cleaner backgrounds than real data, producing a larger domain gap when transferred to real datasets (retention 76–83%). The NUDT-SIRST → IRSTD-1K pair is the hardest (75.5%), consistent with the well-known synthetic-to-real distribution shift.

Real → synthetic transfer is easier. Real-trained models transfer to NUDT-SIRST with 86–87% retention, confirming that real-world feature diversity subsumes the simpler synthetic patterns. This asymmetry further supports that InvDet's information-preservation mechanism does not overfit to either domain type.

9.2. Cross-Dataset Feature Visualization

To further examine whether the invertible encoder develops dataset-agnostic feature representations, we select one trained checkpoint (IRSTD-1K) and visualize its internal features on test samples from all three datasets (Fig. 10).

For each sample, we extract the low-pass (LP) and high-pass (HP) components from the deepest reversible stage, the GCTM-derived weight map $\mathbf{W}_s$, the TARM-modulated high-pass HP′, the reconstructed image $\mathbf{X}^{\text{rec}}$, and the reconstruction error $|\mathbf{X}^{\text{rec}} - \mathbf{X}|$.

Across all three datasets, the feature maps exhibit consistent behavior: (i) $\mathbf{W}_s$ concentrates sharply on target re-

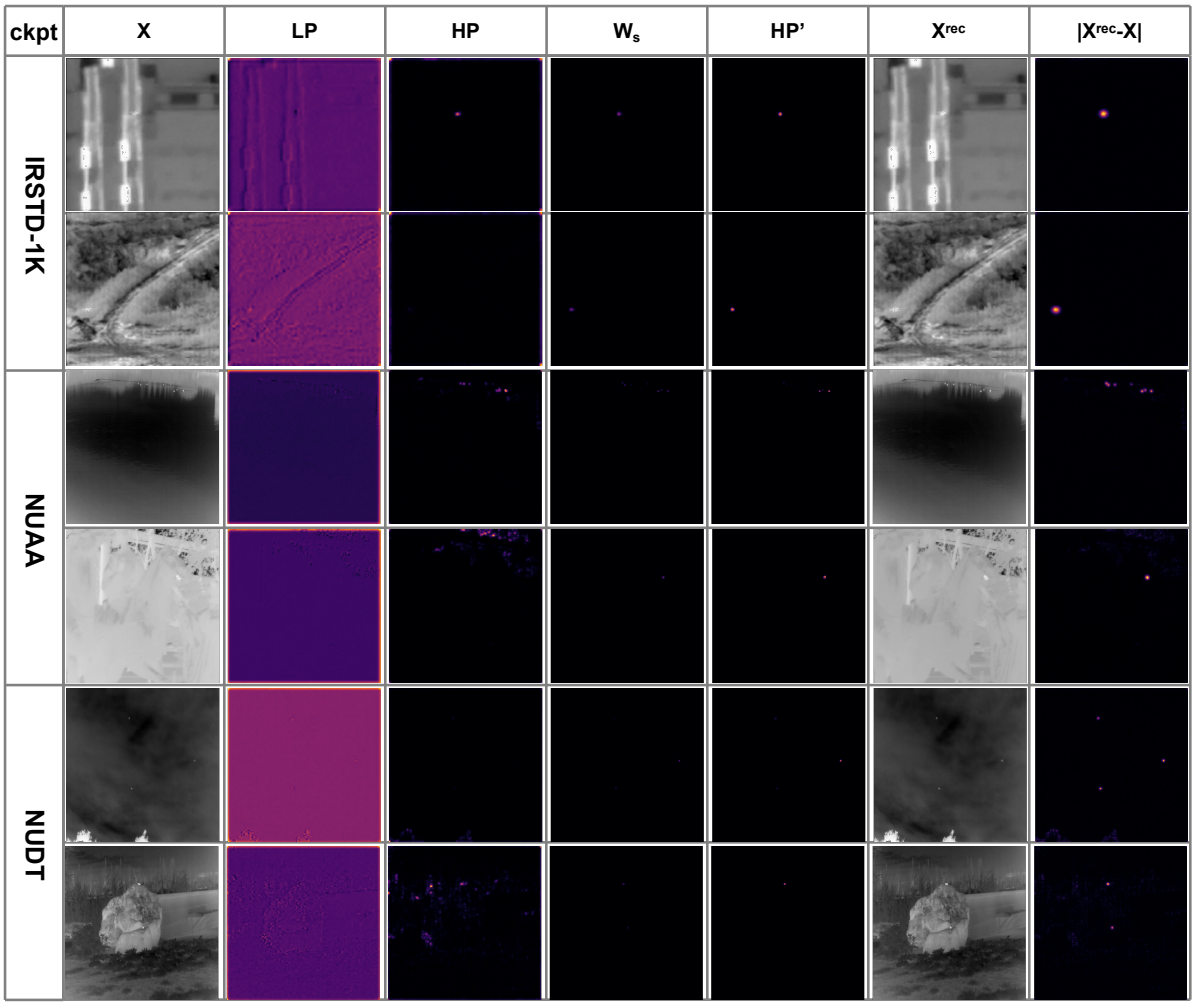

Figure 10. Cross-dataset feature visualization. Every model(.ckpt) trained on source is applied to samples from other cross-domain datasets. Each row shows: input image (X), low-pass component (LP), high-pass energy (HP), GCTM weight map ($\mathbf{W}_s$), TARM-modulated high-pass (HP′), reconstruction ($\mathbf{X}^{\text{rec}}$), and reconstruction error ($|\mathbf{X}^{\text{rec}} - \mathbf{X}|$).

gions regardless of background type; (ii) the LP branch elevates the baseline signal at target locations while remaining suppressed elsewhere; (iii) the HP branch after TARM modulation (HP′) retains target boundary details and attenuates background textures; and (iv) the reconstruction error $|\mathbf{X}^{\text{rec}} - \mathbf{X}|$ is low at target positions and higher in background, confirming that reconstruction fidelity is strategically allocated. These patterns are qualitatively identical across the real IRSTD-1K, real NUAA, and synthetic NUDT data, supporting that the invertible encoder learns a general information-preservation mechanism rather than dataset-specific shortcuts.

10. Comprehensive Ablation Study

To validate the necessity of each design choice, we conduct ablations along four orthogonal axes under the same IRSTD-1K protocol (identical split, augmentation, and hyperparameters). The full model achieves F1 = 86.4%. Results are consolidated in Tab. 8.

Table 8. Unified ablation study on IRSTD-1K (F1 %). Four panels isolate orthogonal design dimensions: (A) training strategy, (B) core components, (C) TARM internals, and (D) GT noise robustness.

A. Training Strategy			B. Components		
ID	**Config**	**F1**	**ID**	**Config**	**F1**
S1	ResNet50, det-only	76.5	C1	ResNet50 backbone	73.2
S2	Inv+$\mathcal{L}_{\text{rec}}$, single opt	42.3*	C2	Haar + ConvBlock	57.2†
S3	Inv+$\mathcal{L}_{\text{rec}}$, dual opt	46.1*	C3	Haar + InvBlock, no $\mathcal{L}_{\text{rec}}$	79.0
S4	S3 + decoupling	79.7	C4	C3 + $\mathcal{L}_{\text{rec}}$+TARM (no GCTM)	76.8
S5	S4 + TARM+GCTM	**86.4**	C5	C4, no MMFB	74.1
			C6	C3 + $\mathcal{L}_{\text{rec}}$+GCTM (no TARM)	78.5

*S2/S3 non-convergence; †C2 Haar-only (no learnable coupling) insufficient.

C. TARM Internals			D. GT Noise Robustness		
ID	**Ablation**	**F1**	**ID**	**GT Noise**	**F1**
T1	No LP gain	81.8	G1	Loc Center jitter ±5% diag	83.3
T2	No HP gating	83.0	G2	Loc Center jitter ±10% diag	77.0
T3	No High-Boost	82.5	G3	Scale jitter ∈[0.8, 1.2]	84.5
T4	Full TARM (forward)	—‡	G4	Drop 20% GT boxes (W-gen only)	75.7

‡T4: modulation on the detection path causes feature distribution drift.

Panel A — Training strategy. Simply adding reconstruction to the invertible encoder collapses training: S2 (single optimizer) and S3 (dual optimizers, shared gradients) both fail to converge (F1 ≈ 42–46%), because reconstruction gradients bias features toward global fidelity and conflict with detection discrimination. Gradient decoupling (S4) immediately stabilizes training to 79.7%, exceeding even the ResNet-50 baseline (S1: 76.5%), and adding TARM+GCTM (S5) yields a further +6.7%. This establishes that the decoupled training strategy—not merely the invertible architecture—is the key enabler for reconstruction to serve as a useful regularizer.

Panel B — Core components. The invertible backbone alone (C3: 79.0%) outperforms both the ResNet-50 baseline (C1: 73.2%) and the incomplete Haar-only backbone (C2: 57.2%), confirming that learnable InvBlock coupling is essential. Adding reconstruction with only TARM (C4: 76.8%) or only GCTM (C6: 78.5%) individually shows that each addresses a distinct aspect: TARM localizes reconstruction to target regions, while GCTM stabilizes supervision via geometry–content tolerance. Combining both achieves the full 86.4% (+7.4 over C3), demonstrating their complementary nature. Removing MMFB from C4 (C5: 74.1%, −2.7%) confirms the necessity of multi-rate frequency fusion in the detection neck.

Panel C — TARM internals. All three sub-components contribute: removing LP gain (T1: −4.6%), HP gating (T2: −3.4%), or the High-Boost residual (T3: −3.9%) each degrades F1 notably. Critically, T4 shows that applying TARM on the forward detection path (rather than only on the inverse reconstruction path) causes training divergence, validating the core design principle: the forward feature distribution must remain unperturbed for stable detection.

Panel D — GT noise robustness. GT boxes are perturbed only for weight-map generation; detection supervision uses clean GT. Moderate perturbations (G1: ±5% center jitter, −3.1%; G3: scale jitter, −1.9%) cause bounded degradation, thanks to GCTM's radiometry-aware normalization and soft sigmoid gating. Under aggressive perturbations (G2: ±10% center jitter, −9.4%; G4: 20% box dropout, −10.7%), performance drops more substantially but does not collapse—F1 under G4 (75.7%) remains close to the S4 baseline (79.7%) which uses clean GT. These results delineate the robustness boundary: GCTM tolerates moderate annotation noise well, but performance degrades when the noise magnitude approaches or exceeds the target scale itself, as the reconstruction constraint increasingly loses spatial grounding.