AuxDet: Auxiliary Metadata Matters for Omni-Domain Infrared Small Target Detection

Yangting Shi, Renjie He, Le Hui, Xiang Li, Jian Yang, Ming-Ming Cheng, Yimian Dai

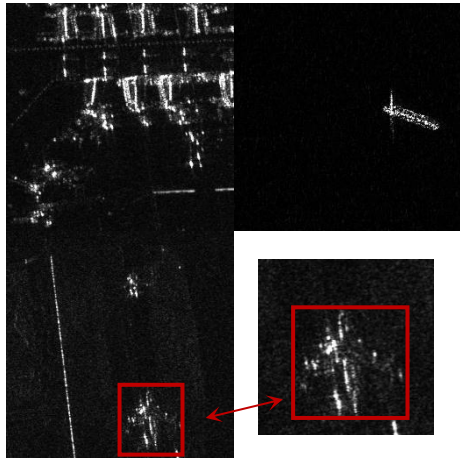
arXiv: https://arxiv.org/abs/2505.15184

code: https://github.com/GrokCV/AuxDet

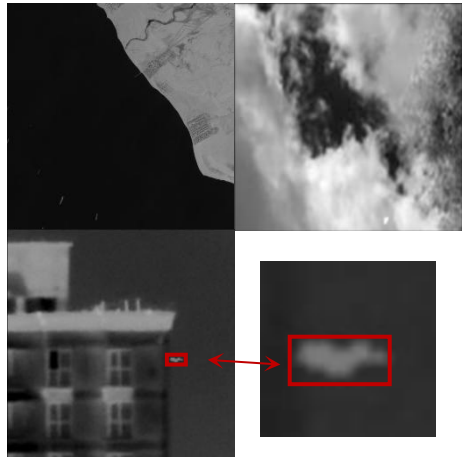
Background

Remote Target Detection

Visible



SAR



Infrared

☑*Why Infrared Target Detection?*

Background

Remote Target Detection

Short-wave infrared defogging effect. Image source: https://zhuanlan.zhihu.com/p/680624604

☑*Why Infrared Target Detection?*

◇*See through darkness and fog*

◇*Spot small targets with thermal contrast*

📌*Typical Applications:*

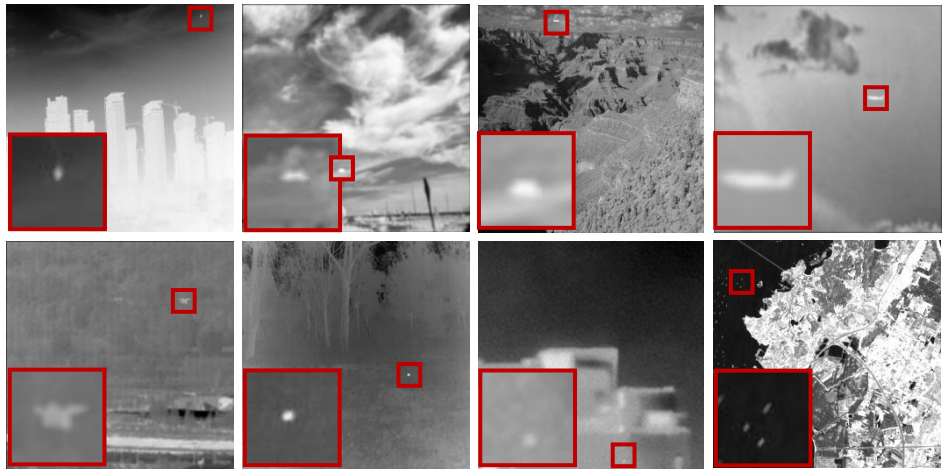
Night surveillance, Search and rescue, Maritime target detection, and more

Background

🎯Infrared Small Target Detection (IRSTD)

Challenges of IRSTD:

- ◇*Target Smallness & Lack of Texture*
  - ⇀ Lacking size and features for reliable detection
- ◇*Complex & Dynamic Background*
  - ⇀ High false alarm rate due to structured noise



Current Solutions:

- ◇*Robust Feature Extraction*
  - ⇀ Multi-scale fusion, local enhancement
- ◇*Refined Attention Mechanisms*
  - ⇀ Context modeling, background suppression

What's missed?

- ◇Wide-area sensing demand
- ◇Imaging condition diversity

Is feature-engineering-only detection still enough?

Motivation

Omni-Domain IRSTD

- *multi-platform observations*

e.g., land-based, aerial-based, and space-based

- *multi-target types*

point-like, spot-like, and extended targets

- *multi-spectral bands*

e.g., LWIR, NIR, SWIR

- *multi-resolution imaging*

from 256×256 to 6000×6000

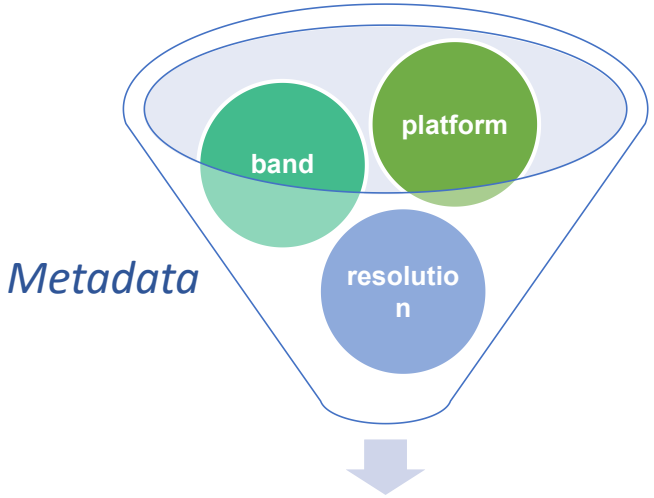
Motivation

Omni-Domain IRSTD

- ◇ Challenge: Existing models struggle to *generalize across platforms, sensors, and resolutions*

  (air-based, land-based, space-based); (LWIR, SWIR, NIR); (256×256 - 6000×6000+)

- ◇ Observation: Metadata is readily available but underused
- ◇ Insight: Metadata encodes global context that complements visual features
- ◇ Motivation: *Leverage **metadata** for efficient, adaptive detection across omni-domain scenarios*



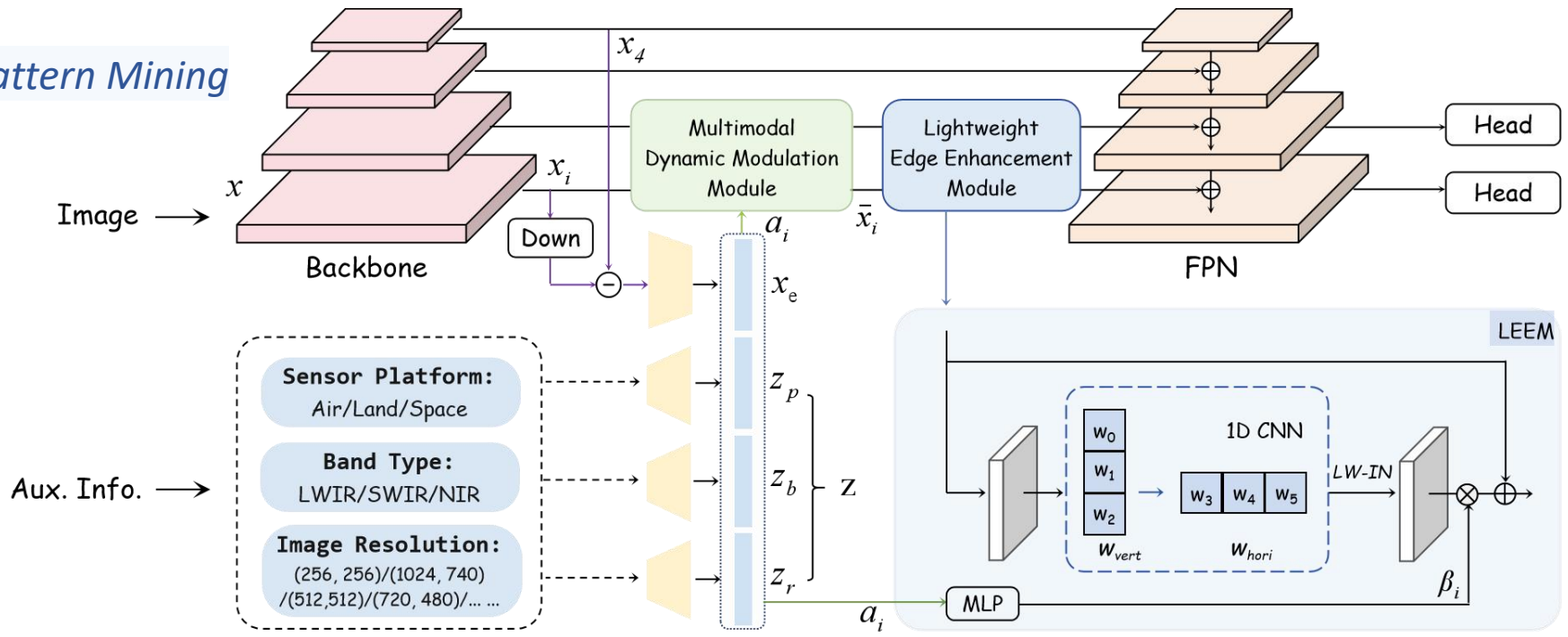
Conventional visual-only paradigm

↓

Leveraging auxiliary metadata for multi-modal scene-aware optimization

Method

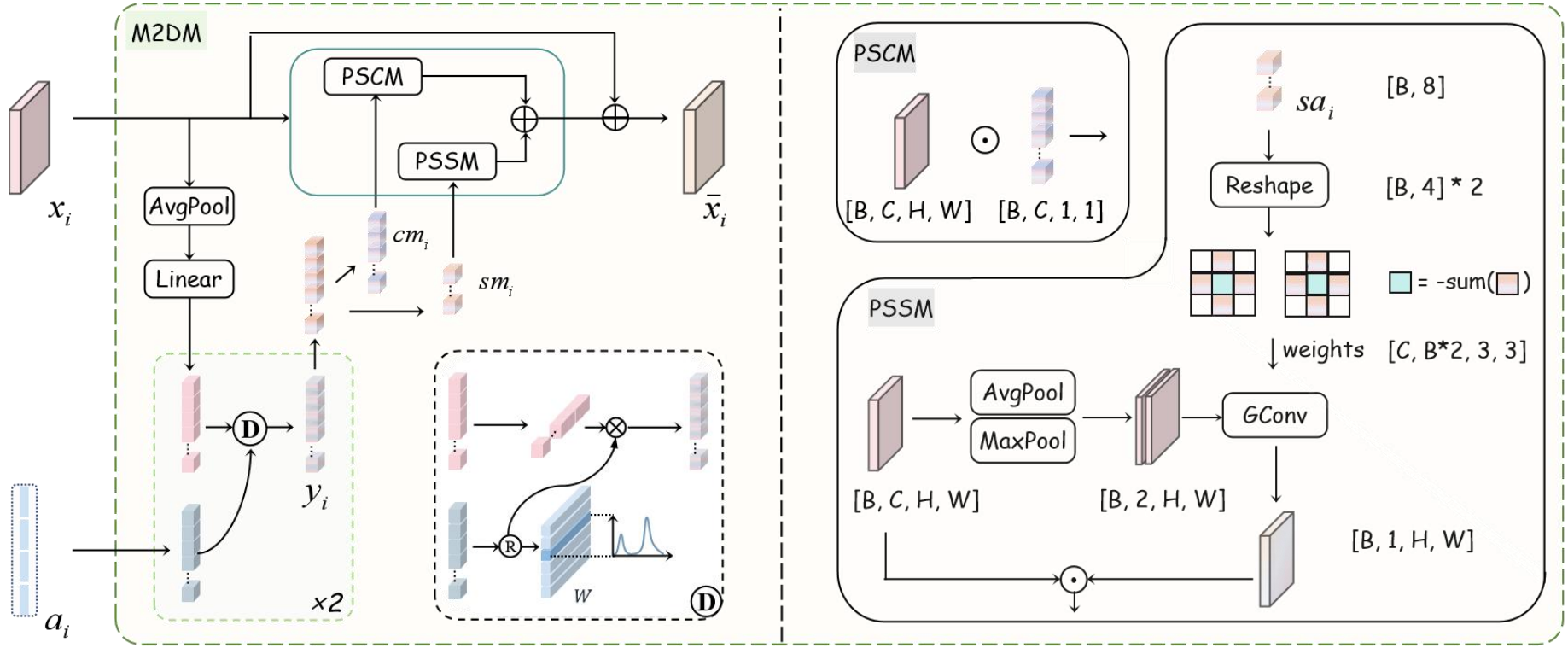
AuxDet (*Visual-Metadata Dual-Pathway*)



Method

Dynamic Feature Modulation Based on Auxiliary Cues

Sample-wise Channel and Spatial Modulation

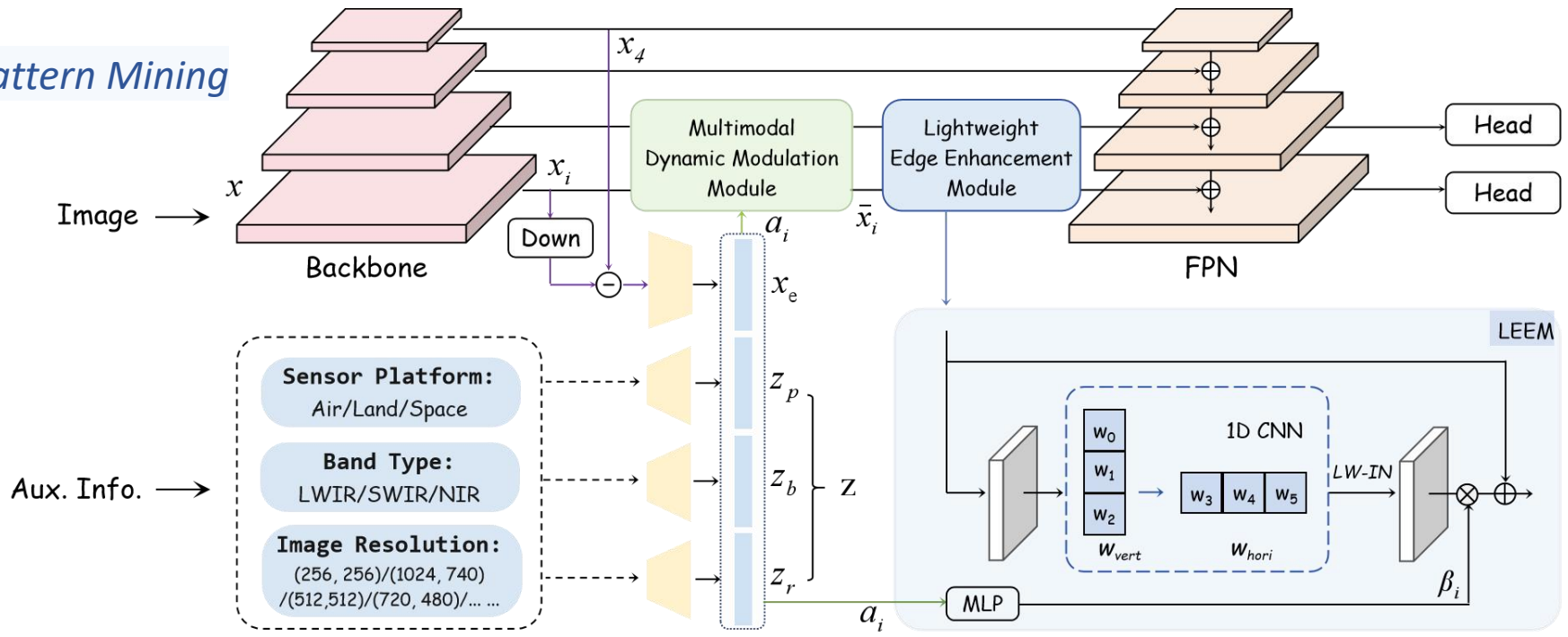


Dynamic MLP-based[1] Multimodal Feature Fusion

1. Yang L, Li X, Song R, et al. Dynamic MLP for fine-grained image classification by leveraging geographical and temporal information[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 10945-10954.

Method

AuxDet (*Visual-Metadata Dual-Pathway*)



Method

Detail Feature Completion

Gaussian-Laplacian-inspired 2D initialization scheme

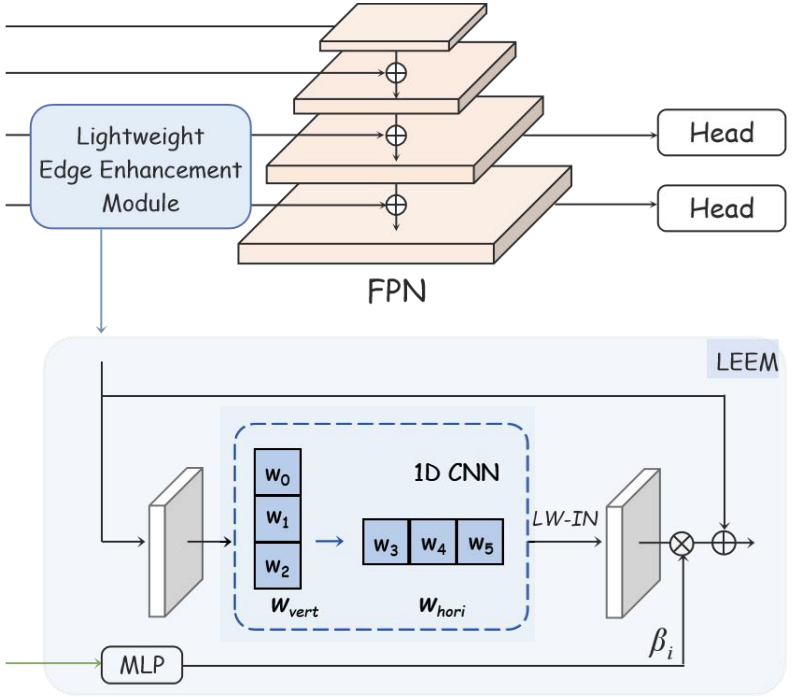
$$\begin{cases} W_{3\times1}^{(0)} = \frac{1}{4}\cdot[1,-2,1]^T \odot I_C \\ W_{1\times3}^{(0)} = \frac{1}{4}\cdot[1,-2,1] \odot I_C \end{cases}, \quad W_{3\times1} * W_{1\times3} = \frac{1}{16}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} \odot I,$$

$$\mathbf{x}_{\text{vert}} = \sigma(\alpha) \odot \text{ReLU}(W_{3\times1} * \mathbf{x}),$$

$$\mathbf{x}_{\text{hori}} = W_{1\times3} * \mathbf{x}_{\text{vert}},$$

$$\bar{\mathbf{x}} = \mathbf{x} + \gamma \odot \left(\frac{\mathbf{x}_{\text{hori}} - \mu_{\text{space}}}{\sqrt{\sigma_{\text{space}}^2 + \text{ò}}} \right).$$

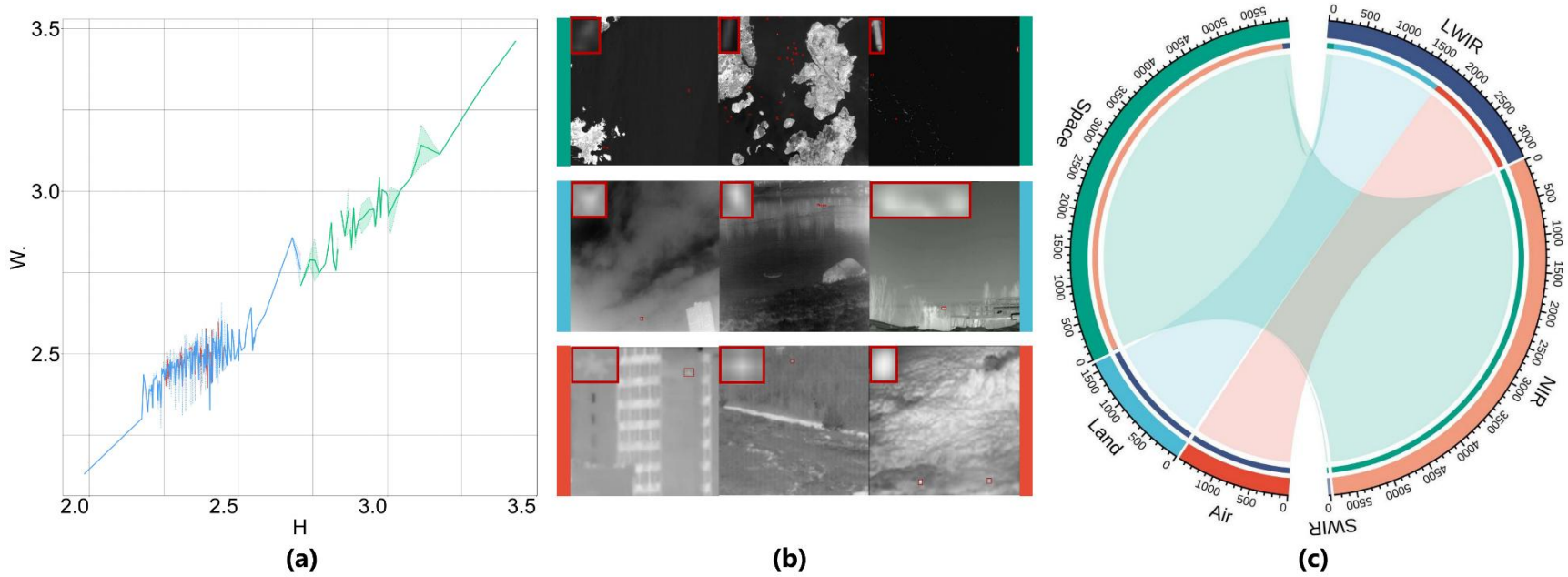
Balanced design between smoothing and edge detection



Experiments

Dataset： WideIRSTD-Full Dataset[2]

This dataset integrates seven prominent public datasets, namely SIRST-V2, IRSTD-1K, IRDST, NUDT-SIRST, NUDT-SIRST-Sea, NUDT-MIRSDT, and Anti-UAV, augmented by an additional dataset from the National University of Defense Technology.



2. https://github.com/XinyiYing/WideIRSTD-Dataset

Result

Table 1: Comparison with Other State-of-the-Art Methods on WideIRSTD-Full Dataset

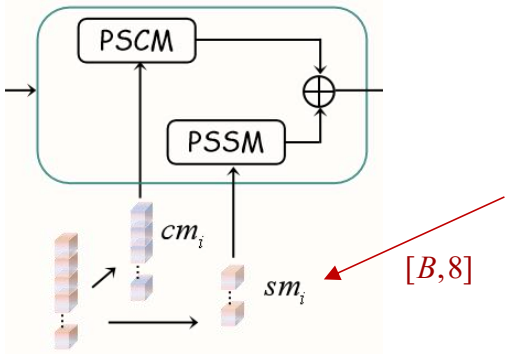
Method			WideIRSTD-Full Dataset			
Category	**Venue**	**Backbone**	FLOPs ↓	Params. ↓	AP_{50} (%) ↑	Recall (%) ↑
General Object Detection	▹ *One-stage* ▹ *Two-stage* ▹ *End-to-End*					
RetinaNet [22]	ICCV[17]	ResNet50	0.197T	36.330M	60.7	74.8
TOOD [10]	ICCV[19]	ResNet50	0.191T	32.018M	53.2	61.9
FCOS [36]	ICCV[19]	ResNet50	0.194T	32.113M	49.2	58.3
GFL [21]	NeurIPS[20]	ResNet50	0.197T	32.258M	48.5	60.2
ATSS [54]	CVPR[20]	ResNet50	0.194T	32.113M	50.1	60.6
AutoAssign [57]	arXiv[20]	ResNet50	0.195T	36.244M	51.2	62.1
Faster R-CNN [29]	NeurIPS[15]	ResNet50	0.200T	41.348M	31.2	33.5
Cascade R-CNN [1]	TPAMI[19]	ResNet50	0.228T	69.152M	35.8	37.5
Grid R-CNN [25]	CVPR[19]	ResNet50	0.313T	64.467M	32.2	34.8
Cascade RPN [37]	NeurIPS[19]	ResNet50	0.190T	42.527M	74.4	84.6
Libra R-CNN [28]	CVPR[19]	ResNet50	0.209T	41.611M	30.7	33.8
Dynamic R-CNN [50]	ECCV[20]	ResNet50	0.200T	41.348M	35.0	37.2
SABL [38]	ECCV[20]	ResNet50	0.198T	36.357M	64.1	77.8
Deformable DETR [59]	ICLR[21]	ResNet50	0.189T	40.099M	33.1	49.8
Sparse R-CNN [32]	CVPR[21]	ResNet50	0.146T	0.106G	40.4	62.1
DINO [51]	arXiv[22]	ResNet50	0.268T	47.540M	67.1	82.8
DAB-DETR [23]	ICLR[22]	ResNet50	99.511G	43.702M	17.6	31.9
Small Target Specialized						
ALCNet [8]	TGRS[21]	ResNet20	6.047G	42.700M	36.3	69.7
NWD [42]	ISPRS[22]	DetectoRS (ResNet50)	0.235T	0.123G	48.5	53.6
RFLA [43]	ECCV[22]	DetectoRS (ResNet50)	0.235T	0.123G	71.8	80.6
OSCAR [5]	TGRS[23]	ResNet50	0.327T	30.730M	40.2	65.4
DNANet [18]	TIP[23]	-	0.228T	4.697M	73.9	46.9
RDIAN [30]	TGRS[23]	-	0.059T	0.217M	62.5	51.8
DQ-DETR [16]	ECCV[24]	ResNet50	1.476T	58.680M	70.1	81.3
EFLNet [45]	TGRS[24]	-	0.102T	38.336M	73.7	68.1
PConv [47]	AAAI[25]	-	0.012T	2.922M	71.7	64.5
AuxDet (Ours)	-	ResNet50	0.215T	45.279M	**77.9**	**87.2**

Result

The guiding role of metadata

Table 3: Ablation Study on Module Combinations and Parameter Analysis

Strategy	Module Config.	Performance		Params. (M)
	Architecture	AP_{50} (%) ↑	Recall (%) ↑	
(a)	Baseline	74.7	84.5	42.527
(b)	PSSM Only	75.8	85.7	44.234
(c)	PSCM Only	75.7	85.3	44.230
(d)	PSSM → PSCM	76.3	86.1	44.367
(e)	PSCM → PSSM	76.8	86.3	44.367
(f)	PSSM + PSCM	77.2	86.7	44.367



LEEM is lightweight and enhances small-target details

Table 4: Ablation study on Lightweight Edge Module vs 3×3 Standard Convolution

Module Structure	Performance		Params. (M)
Kernel Decomposition	AP_{50} (%) ↑	Recall (%) ↑	
Baseline	74.7	84.5	42.527
$W_{3\times3}$ (Standard)	77.3	86.7	43.116
$W_{1\times3} \rightarrow W_{3\times1}$	77.2	86.8	42.723

Result

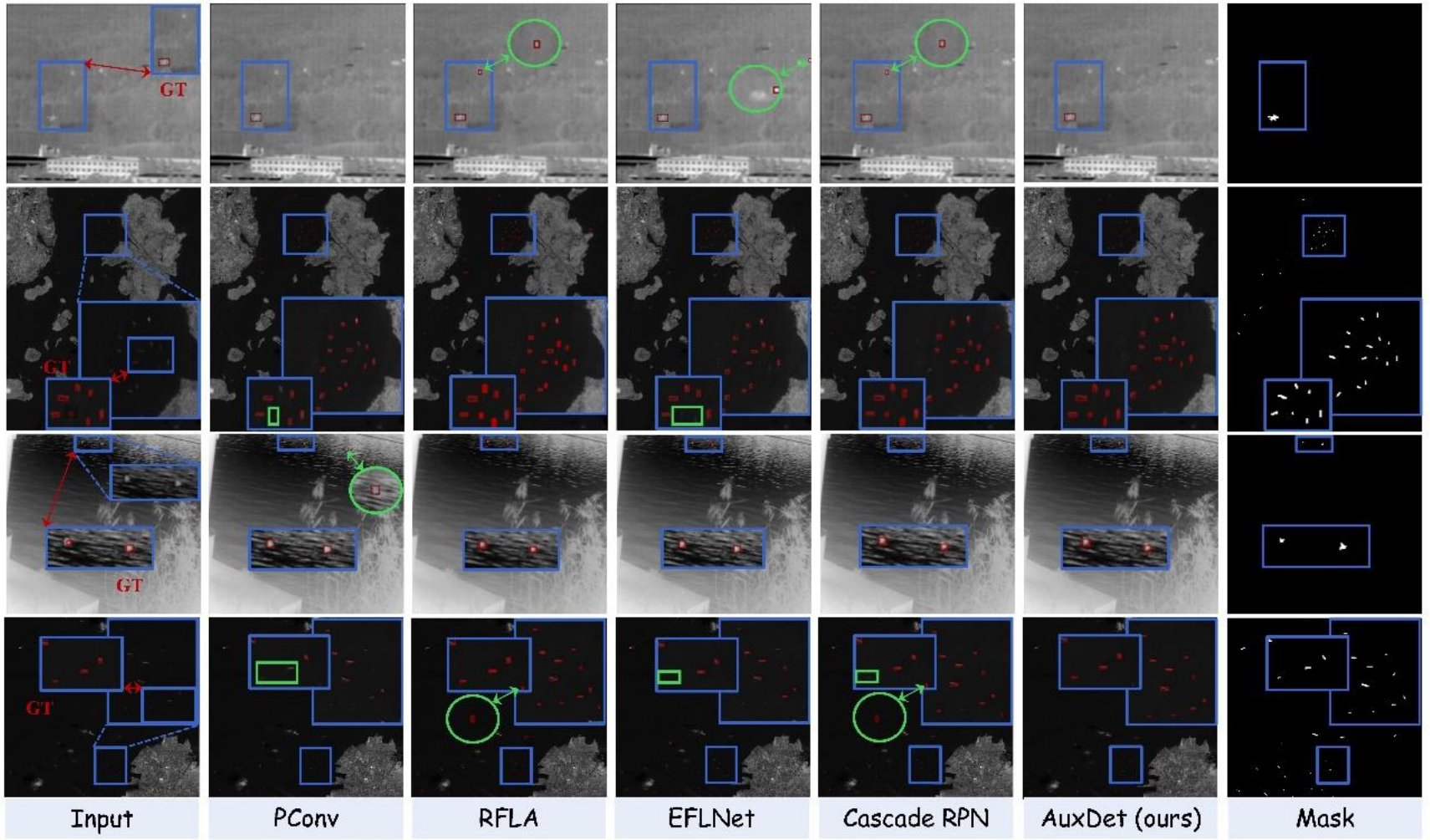


Figure 3: Visualization of detection results. Green circles indicate false alarms, while green rectangles mark missed detections. Compared to other methods, AuxDet has fewer missed detections and false alarms.

Result

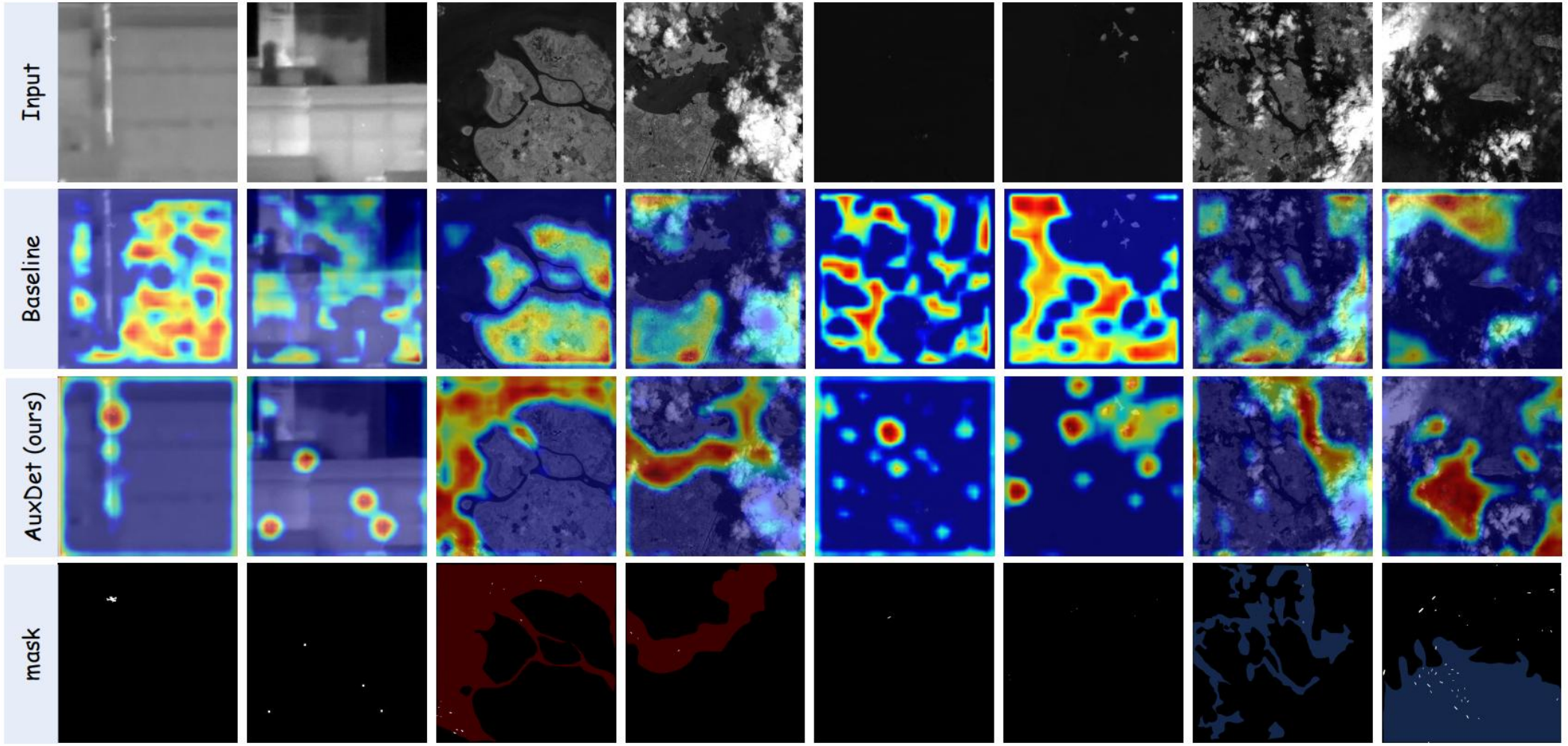


Figure 2: Feature visualization comparison. White masks annotate small targets, where comparative analysis demonstrates our method's enhanced small-target focusing capability over the baseline. In ground truth annotations, we use red to mark distinct contiguous maritime regions and blue to denote barely discernible maritime areas under heavy cloud interference.

Summary

Core Problem: The visual-only paradigm struggles to generalize in Omni-Domain IRSTD.

Motivation: Metadata is naturally available, low-cost, semantically rich, and encodes imaging domain characteristics.

Our Method: AuxDet, a metadata-driven IRSTD paradigm, *deeply leverages auxiliary cues to facilitate the dynamic injection of imaging condition priors and adaptive perceptual tuning for cross-domain images*.

Key Findings: Auxiliary metadata proves essential for enhancing robustness and accuracy, showing broad adaptability in Omni-IRSTD scenarios.

Future Work: Exploring richer metadata sources.

Leveraging vision-language models for metadata understanding and extension.

THANK YOU!

arXiv: https://arxiv.org/abs/2505.15184

code: https://github.com/GrokCV/AuxDet