SSDENODET: SELF-SUPERVISED CROSS-DOMAIN FEATURE DENOISING FOR SAR TARGET DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Labeled synthetic aperture radar (SAR) detection data are costly, while speckle and small targets challenge conventional detectors. We present SSDenoDet, which transfers a corruption-robust masked-autoencoding representation learned from hyperspectral imagery (HSI) to SAR. A frozen MAESSI encoder and centered exponential-moving-average targets train Yield-Aware MAESSI-Aligned Learner (YAMAL), a grayscale SAR image encoder; an FPN then localizes targets from its spatial features. HSI and the teacher are absent at inference. SSDenoDet-E reaches 57.55 COCO mAP on untouched SARDet-100K validation versus 55.60 for our matched DenoDet V2 reproduction, and improves common-evaluator AP_{50} over its single-model counterpart on four independently trained SAR benchmarks.

1 INTRODUCTION

Synthetic aperture radar (SAR) supports observation across illumination and cloud conditions, but costly annotations, speckle, and small targets make robust detection difficult. Self-supervised pretraining reduces dependence on labels, yet natural-image features face an optical-to-SAR domain gap, while SAR-only supervision is limited to within-domain observations. We investigate hyperspectral imagery (HSI) as an auxiliary source of corruption-robust representations: its densely sampled, correlated bands provide context for recovering incomplete or corrupted measurements.

SSDenoDet transfers this recovery-trained prior from a Masked AutoEncoder with Spatial–Spectral Integration for Hyperspectral Imagery (MAESSI) to YAMAL (*Yield-Aware MAESSI-Aligned Learner*), a grayscale image encoder. Cross-Modal Dual-Stream (CMDS) distillation aligns the student's local and global HSI-projection features with a frozen spectral teacher, while a centered EMA teacher supplies consistency targets on unlabeled SAR views. The retained spatial hierarchy then initializes an FPN-based detector. Neither HSI nor the teachers are required at inference.

Our contributions are threefold. First, we formulate HSI-to-SAR adaptation as *robustness transfer*, using a recovery-trained prior without requiring paired HSI–SAR scenes. Second, we introduce an FPN-compatible student with co-located local/global alignment, SAR self-distillation, and anti-collapse constraints. Third, we evaluate the scale-aware detector on four independently trained SAR benchmarks, with matched controls for representation transfer, input resolution, target-scale preservation, and inference aggregation. SSDenoDet-E reaches 57.55 COCO mAP on official SARDet-100K validation; the controls distinguish its modest representation benefit from the larger effect of target-scale preservation.

2 RELATED WORK

SAR detection and cross-domain transfer. SARDet-100K provides a large-scale benchmark; DenoDet and DenoDet V2 address frequency-domain and phase–amplitude feature denoising Li et al. (2024); Dai et al. (2025); Ni et al. (2026). Natural-image transfer faces a substantial SAR domain gap Huang et al. (2020). MSFA introduces an intermediate optical domain, whereas SAR-specific masked pretraining learns directly from unlabeled SAR Li et al. (2024); Pu and Xu (2025). Paired HSI–SAR fusion instead uses complementary sensing modalities jointly Li et al. (2023); Wang et al. (2023a). SSDenoDet uses HSI only as training-time supervision and operates solely on SAR intensity at inference.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

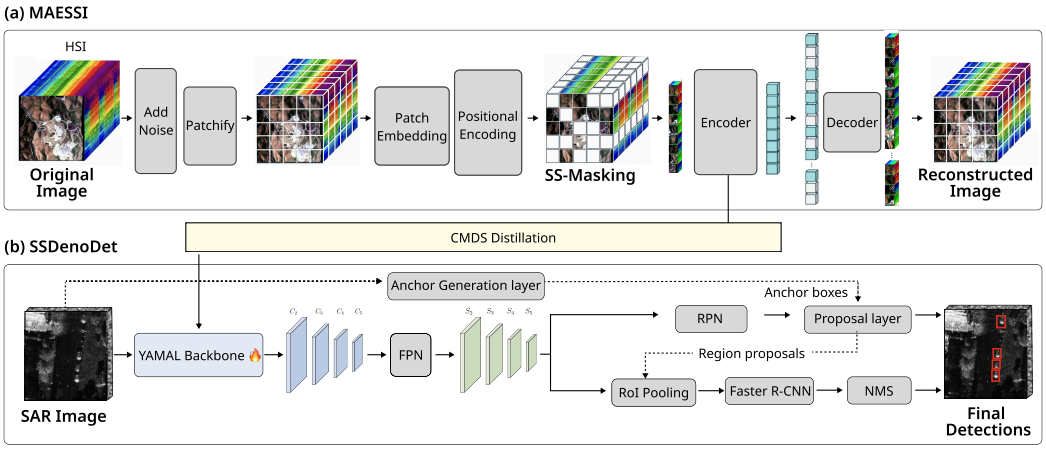


Figure 1: Overview of the proposed HSI-to-SAR representation learning and object detection framework. (a) MAESSI is pretrained to reconstruct clean HSI from noise-corrupted inputs selectively masked along the spatial and spectral dimensions, thereby learning robust spatial–spectral representations. (b) During CMDS, the frozen MAESSI encoder provides HSI-derived guidance to the YAMAL backbone, while view-consistent self-distillation on unlabeled SAR data adapts the backbone to SAR-specific variations. The distilled backbone is then integrated with an FPN-based Faster R-CNN detector to produce the final SAR object detections.

Masked autoencoding and self-distillation. MAE learns representations by reconstructing hidden inputs He et al. (2022); remote-sensing variants incorporate spectral, temporal, scale, and spatial structure Cong et al. (2022); Reed et al. (2023); Lin et al. (2023). MAESSI supplies our corruption-pretrained HSI encoder. EMA teachers stabilize targets Tarvainen and Valpola (2017), DINO sharpens centered teacher distributions Caron et al. (2021); Oquab et al. (2024), and VICReg regularizes representation variance and covariance Bardes et al. (2022). We distinguish the frozen HSI encoder from the within-SAR EMA teacher.

Small-object detection. Tiny targets require sufficient input resolution and spatial feature support Hu and Ramanan (2017). FPN combines lateral and top-down features Lin et al. (2017a); Deformable DETR and ViTDet offer transformer-based alternatives Zhu et al. (2021); Li et al. (2022). SSDenoDet preserves a high-resolution FPN hierarchy and selects scale and suppression settings using validation data.

3 METHOD

Figure 1 gives the complete training-to-inference path. We first learn an image-space representation with two training-only teachers, then fine-tune its spatial hierarchy for dataset-specific SAR detection.

MAESSI ENCODER

Masked AutoEncoder with Spatial–Spectral Integration for Hyperspectral Imagery (MAESSI) denotes the complete wavelength-aware autoencoder pretrained with structured HSI corruption. We denote only its encoder by f_M. SSDenoDet freezes and transfers f_M; the MAESSI reconstruction decoder is used during HSI pretraining but is never part of SAR distillation or inference.

WAVELENGTH-AWARE SPECTRAL ENCODING.

The MAESSI encoder f_M embeds each observed HSI band together with its measured wavelength and processes the sequence using an eight-block, 128-wide Transformer. Pretraining draws approximately 800K spectra from sensors with 157–426 channels; each step samples a contiguous 80-band window, preserving local spectral continuity while varying wavelength range and sensor sampling.

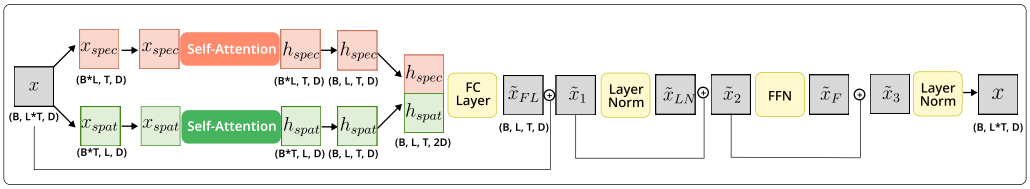


Figure 2: MAESSI encoder block. Spectral and spatial self-attention process complementary token arrangements before feature fusion and a residual feed-forward update.

Spectral–Spatial Masking (SS-Masking).

SS-Masking is the masking operator used when MAESSI receives an HSI cube. Tokens index both a non-overlapping spatial patch and a spectral group. A spatial mask hides every spectral token at selected locations, whereas an independently sampled spectral mask hides selected band groups across the remaining locations; their union defines the missing tokens.

Dual-Stream Spatial–Spectral Attention.

The MAESSI encoder and reconstruction decoder use the same Transformer-block design. Each block rearranges its tokens for parallel spectral and spatial self-attention, modeling inter-band correlations and cross-location structure, respectively, as in spatial–spectral masked pretraining Lin et al. (2023). The branch outputs are fused before a residual feed-forward update (Fig. 2). The decoder reconstructs masked tokens with intensity, spectral-shape, and smoothness losses. The checkpoint supplying f_M was selected under Case-4 dead-line corruption and also exposed to stripe-plus-dead-line corruption; full pretraining details are supplementary.

YAMAL: Spatial MAESSI-to-SAR Distillation

YAMAL (*Yield-Aware MAESSI-Aligned Learner*) is an FPN-compatible ResNet-50 student g_ϕ He et al. (2016) that consumes grayscale intensity images directly. "Yield-aware" means that view construction preserves target-bearing spatial support instead of forcing agreement between local crops that may contain different objects. Copying intensity to three planes only reuses the pretrained stem; no synthetic spectral profile is constructed.

Cross-Modal Dual-Stream Distillation (CMDS)

YAMAL is optimized from the two synchronized streams in Fig. 3. In the cross-domain stream, an HSI patch $X \in \mathbb{R}^{H\times W\times B}$ is reduced to a grayscale projection $\Pi(X)$ for the image student, while the frozen MAESSI encoder f_M receives the original spectrum x_{ij} at every location of a $K \times K$ grid. A 1×1 bridge maps the student's stride-8 feature z_{ij} into the 128-dimensional teacher space. We align both co-located features and their spatial means:

$$\begin{aligned} \mathcal{L}_H^l &= 1 - \frac{1}{K^2}\sum_{ij}\cos(z_{ij}, f_M(x_{ij})), \\ \mathcal{L}_H^g &= 1 - \cos(\bar{z}, \bar{f}_M). \end{aligned} \tag{1}$$

This correspondence avoids matching a whole HSI patch to one randomly chosen spectrum. On unlabeled SAR, two *global* speckle-aware views I_1, I_2 are passed through the student and a stop-gradient EMA view teacher $g_{\bar{\phi}}$. We avoid aggressive local crops because they can contain different tiny targets. Let u_ϕ and $u_{\bar{\phi}}$ denote their 4096-dimensional projection logits. Student and teacher distributions are

$$\begin{aligned} p_\phi(I_a) &= \mathrm{softmax}(u_\phi(I_a)/\tau_s), \\ q_{\bar{\phi}}(I_b) &= \mathrm{softmax}((u_{\bar{\phi}}(I_b) - c)/\tau_t), \\ \mathcal{L}_D &= -\tfrac{1}{2}\sum_{(a,b)\in\{(1,2),(2,1)\}} q_{\bar{\phi}}(I_b)^\top \log p_\phi(I_a), \end{aligned} \tag{2}$$

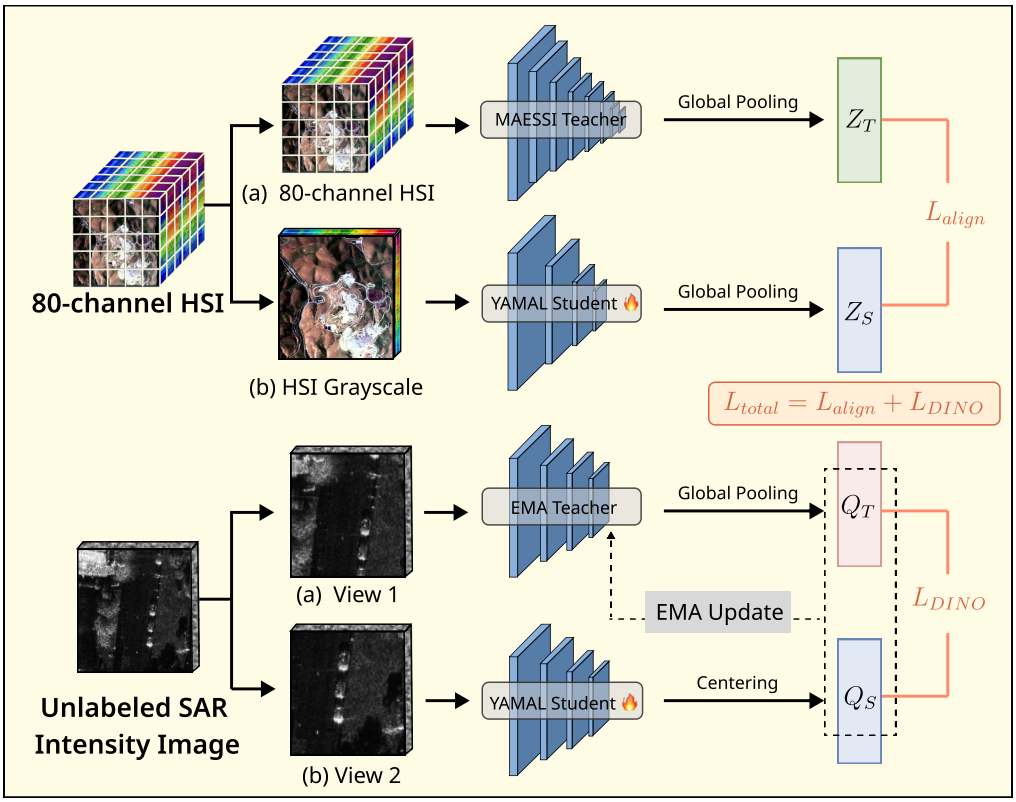


Figure 3: Overview of the proposed CMDS framework. Without requiring direct HSI–SAR sample correspondence, CMDS integrates a frozen spectral teacher and an online SAR teacher within a shared student backbone, combining HSI-derived structural knowledge with representations robust to SAR-specific variations.

where $\tau_s = 0.10$ and $\tau_t = 0.07$. The center c is a non-gradient running mean of teacher logits, updated from each teacher-view batch as

$$c \leftarrow \mu_c c + (1 - \mu_c) \frac{1}{N} \sum_{n=1}^{N} u_{\bar{\phi}}(I_b^{(n)}), \qquad \mu_c = 0.9. \tag{3}$$

Subtracting c before the sharp teacher softmax prevents a few projection dimensions from dominating every target; it does not normalize SAR pixels or use labels. The view teacher itself is updated after each optimizer step by $\bar{\phi} \leftarrow m_t \bar{\phi} + (1 - m_t)\phi$, where m_t follows a cosine schedule from 0.996 to 1.0. We use *feature denoising* to mean invariance to independently speckled views in representation space; SSDenoDet does not smooth the input image or erase bright scatterers before localization. To further prevent collapse, we add VICReg variance/covariance regularization $\mathcal{L}_V$ and a frozen-initialization anchor $\mathcal{L}_A$ that preserves generic edge and texture responses:

$$\mathcal{L}_Y = \mathcal{L}_D + 0.2\mathcal{L}_V + 0.2\mathcal{L}_H^g + 0.35\mathcal{L}_H^l + 0.5\mathcal{L}_A. \tag{4}$$

Each step draws one unlabeled SAR view pair and one HSI patch batch, jointly constraining SAR view stability and the spatial geometry of f_M. The frozen encoder, EMA projection head, and HSI bridge are discarded after distillation, so cross-domain supervision adds no detector inference cost. Without the anchor, the objective decreases while effective rank and held-out kNN macro-F1 deteriorate; we therefore treat rank as a collapse diagnostic rather than selecting distillation by training loss alone. The terms in $\mathcal{L}_Y$ address distinct failure modes. Global HSI alignment alone can match an image mean while discarding the local contrast needed to place a box; $\mathcal{L}_H^l$ instead constrains every

sampled grid location. Local alignment without $\mathcal{L}_H^g$ can preserve neighborhoods but drift in overall scale across sensors. The centered SAR objective supplies intensity-view invariance where no HSI counterpart exists, whereas the initialization anchor limits catastrophic replacement of generic image edges by a spectral-only solution. This division is the core cross-domain design: f_M contributes corruption-tolerant geometry only where a co-located spectrum is available, and the EMA teacher adapts that geometry to SAR without ever receiving optical tokens. Unlike direct feature imitation between unrelated SAR and HSI scenes, no semantic pairing across sensors is assumed. For detector initialization, YAMAL-C applies one fixed anchor calibration, $\theta_C = 0.25\theta_{\mathrm{EMA}} + 0.75\theta_{\mathrm{init}}$, to floating-point backbone weights. This is not an ensemble and adds no inference parameters. Pure and calibrated variants share the detector protocol and are selected only by validation mAP.

The teachers never exchange sensor semantics. Encoder f_M is queried only with co-located HSI spectra; the EMA teacher sees only augmented SAR training images. Its center uses training-batch logits only—never validation/test images, labels, or boxes. The two global views retain at least 80% of image support and combine percentile normalization with flips, contrast, blur, and multiplicative Gamma speckle. Centering, sharpening, stop-gradient, and EMA therefore stabilize target-bearing evidence without imposing agreement between unrelated local crops.

Gradient routing and domain isolation. The HSI and SAR batches need not depict the same scene: pairing exists only within an HSI patch, where $\Pi(X)_{ij}$ and x_{ij} share a location. The student is the sole interaction point between streams; f_M, the EMA teacher, and the initialization anchor receive no gradients.

SMALL-TARGET FPN DETECTION

After distillation, the EMA ResNet stages initialize a shared feature pyramid Lin et al. (2017a). On SARDet-100K we compare a two-stage Faster R-CNN head Ren et al. (2015) and an anchor-free FCOS head Tian et al. (2019) on the same split. Both retain COCO-pretrained FPN/localization modules, replace dataset-specific class logits, and transplant the identical YAMAL backbone. The smaller dataset-specific dense-anchor runs are initialized without COCO head weights. Their Faster R-CNN heads use three aspect ratios and scales spanning 8–512 pixels, including 8/16/32-pixel anchors on the highest-resolution level. Its supervised objective is

$$\begin{aligned}\mathcal{L}_{2s} = {} & \mathrm{BCE}(o, o^*) + \lambda_r \,\mathrm{SmoothL1}(b_r, b_r^*) \\ & + \mathrm{CE}(p, y^*) + \lambda_b \,\mathrm{SmoothL1}(b, b^*),\end{aligned} \tag{5}$$

where o, b_r, p, b are RPN objectness, proposal boxes, RoI class probabilities, and refined boxes. FCOS replaces these terms with focal classification Lin et al. (2017b), generalized-IoU regression, and centerness:

$$\begin{aligned}\mathcal{L}_{focal} &= -\alpha_t(1-p_t)^\gamma \log p_t, \\ \mathcal{L}_{FCOS} &= \mathcal{L}_{focal} + \lambda_g(1-\mathrm{GIoU}) + \lambda_c \,\mathrm{BCE}(q, q^*),\end{aligned} \tag{6}$$

with $\alpha = 0.25$ and $\gamma = 2$. The selected two-stage path protects tiny objects through high-resolution assignment and small anchors rather than an unreported test-specific loss weight; the FCOS control uses focal modulation to reduce abundant easy background cells. The head and input scale are selected using validation COCO mAP.

The detector preserves C2–C5 maps at strides 4, 8, 16, and 32. Lateral 1×1 projections and top-down fusion expose semantically stronger features at the stride-4/8 levels, where an 8–16 pixel scatterer still occupies multiple cells. Assignment follows box scale rather than category, so ships and aircraft of similar size share the same pyramid policy. Crucially, this interface consumes spatial YAMAL maps, not pooled embeddings: distillation changes initialization, but class discrimination and multi-box coordinates remain learned from each benchmark's own annotations.

AIR-SARShip scenes are 3000×3000 pixels; whole-scene resizing makes many ships smaller than the RPN stride. We therefore use 512×512 windows with 200-pixel overlap and remap every visible box. Scale and overlap are fixed on scene-disjoint validation before the public train and validation scenes are combined for the final test fit. Other benchmarks retain their native image units.

SSDENODET-E: VALIDATION-GATED CONSENSUS

SSDenoDet-E is the enhanced inference configuration of the same YAMAL–FPN detector, not an additional backbone. On SARDet-100K and SARBench, it groups same-class boxes from $M = 3$

Algorithm 1 YAMAL Distillation and Scale-Aware Detection

Require: frozen MAESSI encoder f_M; student g_ϕ; EMA teacher $g_{\bar{\phi}}$; center c
for each distillation step **do**
sample HSI patch X and train-split SAR image I
$t_{ij} \leftarrow \mathrm{sg}[f_M(x_{ij})]$; $z_{ij} \leftarrow q(g_\phi(\Pi(X)))_{ij}$
build support-preserving SAR views (I_1, I_2)
compute centered cross-view targets and $\mathcal{L}_Y$
update (ϕ, q); update c and $\bar{\phi}$ by EMA
end for
discard f_M and projection heads; retain EMA C2–C5
for each SAR benchmark d independently **do**
build P2–P6; assign 8–512 px anchors by box scale
optimize $\mathcal{L}_{2s}$ (or matched $\mathcal{L}_{FCOS}$ control)
select scale, NMS, and checkpoint using validation AP only
end for
freeze checkpoint-consensus or flip-view policy on validation
cluster same-class boxes; fuse coordinates and support-aware scores
return SSDenoDet (single pass) or SSDenoDet-E (frozen consensus)

Family	Model	SARDet-100K COCO mAP	AIR-SARShip mAP(07)	SAR-AIRcraft mAP(07)	SARBench COCO AP_{50}	Params
Two-stage	Faster R-CNN	39.22	67.48	64.71	78.92	41.37M
Two-stage	ConvNeXt	53.15	67.52	67.41	79.78	45.07M
End-to-end	Deformable DETR	52.00	55.34	62.43	77.63	40.10M
End-to-end	Conditional DETR	44.04	9.09	62.25	78.03	43.45M
One-stage	FCOS	52.52	63.66	62.63	77.27	32.13M
One-stage	YOLOF	42.83	48.32	66.25	79.89	42.46M
One-stage	ATSS	54.95	64.21	66.01	79.91	32.13M
One-stage	TOOD	54.65	63.92	62.66	78.26	30.03M
One-stage	DenoDet V1	55.90	**70.42**	**68.60**	81.02	70.33M
One-stage	DenoDet V2	55.60	68.92	65.30	83.01	37.15M
Cross-domain	SSDenoDet	**56.77**	67.07	49.19	**88.85**	43.28M
Cross-domain	**SSDenoDet-E (ours)**	**57.55**	**71.98**	**69.88**	**91.12**	129.85M

Table 1: Benchmark-native comparison across four SAR datasets. SARDet-100K uses COCO mAP, AIR-SARShip and SAR-AIRcraft use VOC07 mAP, and SARBench uses COCO AP_{50}; values are comparable only within a column. Sources and reproduction protocols are stated in the text. The SSDenoDet-E parameter entry follows its three-checkpoint SARDet-100K policy; AIR variants reuse one resident checkpoint across three views. **Best** and **2nd-best** results are highlighted within each benchmark column.

independently saved saturation checkpoints when their IoU exceeds a validation-selected threshold, allowing at most one box from each checkpoint per cluster. For cluster $\mathcal{C}$, confidence-weighted box fusion Solovyev et al. (2021) and support-aware scoring give

$$\hat{b} = \frac{\sum_{k\in\mathcal{C}} s_k^\beta b_k}{\sum_{k\in\mathcal{C}} s_k^\beta}, \qquad \hat{s} = \frac{1}{M}\sum_{k\in\mathcal{C}} s_k. \tag{7}$$

Agreement therefore reduces small-box localization jitter, while a detection missing from other checkpoints is penalized by the zero contribution in the M-way score. For the smaller AIR benchmarks, where checkpoint diversity is limited, E instead merges identity, horizontal-flip, and vertical-flip predictions from one selected checkpoint with class-aware NMS. Input scale, checkpoint/view set, fusion IoU, β, and NMS are selected on validation and frozen before one test evaluation. All detector passes are included in the reported computation; HSI and MAESSI remain absent at inference.

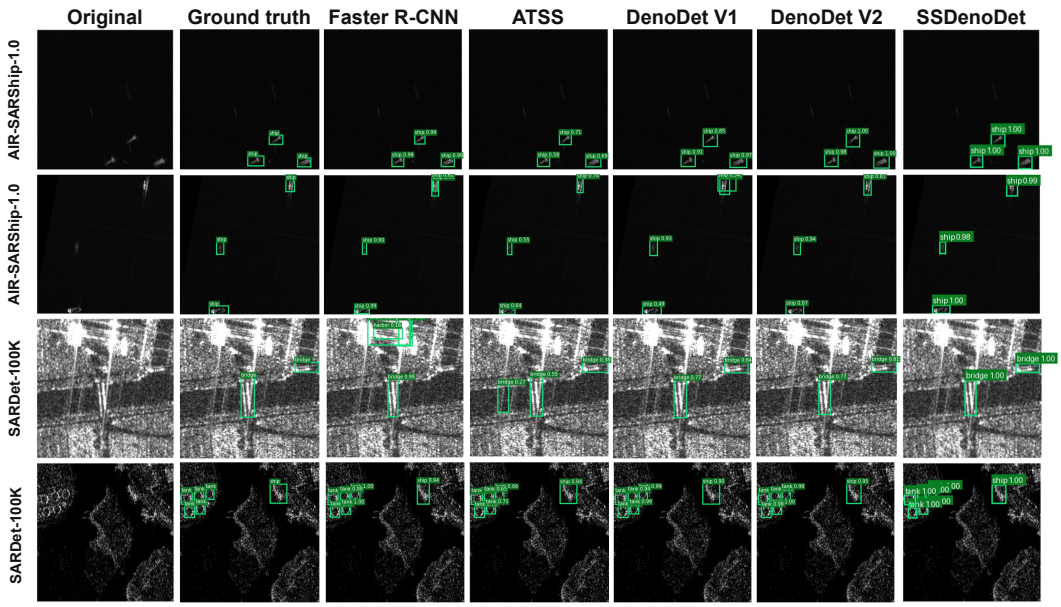


Figure 4: Qualitative detection comparison on AIR-SARShip-1.0 and SARDet-100K. Faster R-CNN and ATSS represent generic detectors, DenoDet V1/V2 are our matched reproductions, and SSDenoDet is our cross-domain model.

Dataset	Train	Valid	Test	Classes	Boxes
SARDet-100K	94,493	10,492	11,613	6	245,568
AIR-SARShip-1.0†	460	324	810	1	1,181
SAR-AIRcraft-1.0	3,047	442	879	7	16,463
SARBench	2,914	728	1,962	1	16,969

Table 2: Independent detection splits. †AIR-SARShip counts are virtual tiles produced after scene-disjoint splitting; Boxes counts visible instances after boundary clipping and tiling. Split construction and geometry audits are detailed in the Supplementary Material.

4 Experiments

Except for DenoDet V1/V2 and our models, Table 1 uses values reported by DenoDet V2 where available. Because public DenoDet V1/V2 checkpoints were unavailable, we retrained both methods on each benchmark's provided training split and evaluated the prescribed validation or test split. SARBench is absent from the DenoDet V2 study, so its entire column comprises our matched runs.

Datasets and Protocol

We evaluate SARDet-100K Li et al. (2024), AIR-SARShip-1.0 Sun et al. (2019), SAR-AIRcraft-1.0 Wang et al. (2023b), and our SARBench split of HRSID Wei et al. (2020). The released YAMAL checkpoint uses only unlabeled *Train* images from these four corpora; test images are never exposed as unlabeled views. Each detector is trained and evaluated only on its named benchmark, with no mixed-dataset box supervision. SARDet-100K and SAR-AIRcraft follow the provided splits. For SARBench, 20% of the official training set is held out for validation. AIR-SARShip is first split by scene; after architecture and inference settings are fixed on the 17/4-scene development split, all 21 public training scenes are used for the final test fit. We report class-aware COCO AP Lin et al. (2014) with at most 100 detections per chip or image, and additionally merge AIR chip predictions into original-scene coordinates for the DenoDet-native VOC comparison. Different metrics are labeled separately and never averaged.

Encoder	Accuracy	Macro-F1	Eff. rank$_{256}$	Test n
ImageNet C5	0.884	0.350	170.76	4000
YAMAL C5	0.893	0.382	191.61	4000

Table 3: Held-out SAR representation probe. Both rows use the same grayscale images and 4,000/4,000 samples; only the encoder differs.

Data	Model	mAP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$
SD	S	54.51	88.03	58.25	48.91	64.53
	E	**55.41**	**88.75**	**59.49**	**49.70**	**65.59**
AS	S	29.73	69.95	15.66	1.19	32.47
	E	**30.31**	**77.30**	15.35	**4.26**	**33.39**
SA	S	30.60	48.98	30.47	22.27	30.32
	E	**44.04**	**70.84**	**47.85**	**49.63**	**43.38**
SB	S	59.27	88.85	68.00	49.11	71.01
	E	**62.80**	**91.12**	**71.59**	**52.27**	**74.55**

Table 4: SSDenoDet test results under the dataset-isolated common evaluator. SD, AS, SA, and SB denote SARDet-100K, AIR-SARShip-1.0, SAR-AIRcraft-1.0, and SARBench. S is a single checkpoint and view; E is the validation-gated SSDenoDet-E consensus. SARDet-100K uses official test annotations. These common-COCO diagnostics use different splits or metrics from the benchmark-native comparison in Table 1.

Latent-Space Probe Before Detection

Before detector fine-tuning, a distance-weighted 5-NN probe compares the same 4,000 train and 4,000 held-out images. Both encoders are represented by globally pooled C5 features—the exact backbone stage transplanted into the detector—so no randomly initialized projection head favors either model. Relative to the ImageNet initialization, YAMAL improves accuracy from 0.884 to 0.893 and macro-F1 from 0.350 to 0.382. The top-256 effective rank also increases from 170.76 to 191.61, ruling out a gain caused by representation collapse.

Convergence and matched latent-space plots are provided in the Supplementary Material. Model selection uses the held-out probe in Table 3 rather than the self-supervised loss alone.

Detection Results

Under the common COCO evaluator in Table 4, validation-gated consensus improves AP$_{50}$ over the corresponding single model on all four datasets. The largest gain occurs for fine-grained aircraft after small-object chip initialization and final fitting at the benchmark's native resize. The Supplementary Material provides additional representative and quantile-sampled comparisons, saturation curves, and per-class results.

Table 1 labels each benchmark's native metric. Three-checkpoint fusion reaches 57.55 COCO mAP on official SARDet-100K validation versus 55.60 for DenoDet V2. On official test, however, SSDenoDet-E reaches 55.41 versus the reported 56.39, so we claim validation rather than test superiority. On the two AIR benchmarks, the native VOC and common-COCO results are reported separately and are never compared across evaluators. The DenoDet V1/V2 entries are matched reproductions rather than transferred checkpoint outputs. Figure 4 provides a visual comparison on two representative test sets.

The isolated runs expose complementary regimes: tightly spaced multiclass targets stress recall and class-aware suppression on SARDet-100K; the large AP$_{50}$–mAP gap on the 31-scene AIR-SARShip split reveals imprecise box extent; and fine-grained aircraft confusions limit macro AP. SARBench has more consistent ship geometry. This variation motivates dataset-specific fitting rather than a mixed-corpus headline score.

Variant	Components	mAP	AP_S
ImageNet initialization	–	57.91	48.60
SAR self-distillation	E	57.76	48.57
+ MAESSI global	E+G	58.70	49.53
+ MAESSI spatial	E+G+L	58.23	49.29
Full SSDenoDet	E+G+L+R	**58.71**	**49.67**

Table 5: Staged YAMAL component ablation under one matched SARBench validation protocol. E is SAR EMA self-distillation, G/L are global/local f_M alignment, and R is VICReg plus the initialization anchor. This validation mAP is distinct from the benchmark-native AP_{50} in Table 1.

Matched Controls and Optimization

Table 5 isolates representation transfer under one SARBench protocol. Full YAMAL improves on ImageNet initialization by 0.80 mAP and 1.07 AP_S. The sequence is non-monotonic, and one-factor controls in the Supplementary Material do not support an independent gain claim for any single loss. We therefore interpret HSI supervision as a modest, collapse-controlled initialization effect rather than the sole source of the detector improvement.

A matched backbone/resolution control in the Supplementary Material shows that 800-pixel adaptation closes the aggregate initialization gap and that target-scale preservation is the dominant small-object factor.

YAMAL distillation runs for eight epochs of 300 steps with SAR/HSI batches of 128/16, AdamW, peak learning rate 2×10^{-4}, and weight decay 0.05. Detector fine-tuning uses cosine decay, a $4{\times}10^{-4}$ head rate and $0.2\times$ backbone rate at batch 16. Checkpoints, resolution, NMS, and consensus parameters are selected on validation only and frozen before test; datasets are trained separately on one H200 each. The released records contain complete schedules, selection grids, split counts, and COCO metrics.

5 Analysis

Transfer and robustness. HSI redundancy supports the teacher prior: reducing Indian Pines from 220 interpolated bands to 55 changes Random Forest accuracy by only 0.1 points; across five datasets, MAESSI improves mean kNN macro-F1 by 3.7 points and reconstructs held-out spectra at 0.972–0.977 SSIM (Appendix). The HSI bridge transfers corruption-tolerant feature geometry, while SAR semantics and localization remain image- and box-driven. The C5 probe and matched detector controls distinguish this adaptation from zero-shot transfer. Lower self-supervised loss alone is insufficient: the unanchored run reduces effective rank and macro-F1, motivating centered targets, initialization anchoring, and held-out representation checks.

Limitations. Optical reflectance and microwave backscatter remain physically different, and localization still requires box supervision. Rare aircraft classes show validation–test shift; AIR-SARShip has few independent scenes. The 800^2 two-stage detector is more expensive than DenoDet V2, and native VOC scores must remain separate from common-COCO diagnostics. Future work will examine a compressed one-stage decoder, multi-sensor SAR pretraining, rare-class uncertainty, and official-server evaluation.

6 Conclusion

SSDenoDet transfers an HSI recovery-trained prior to a grayscale SAR encoder and combines it with a small-target FPN detector. Four benchmark-specific experiments indicate complementary roles for representation transfer and spatial detection, with target-scale preservation dominating tiny-object gains. HSI and the teachers are used only during training.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, concision, and LaTeX formatting of this manuscript.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, volume 35, pages 197–211, 2022.

Yimian Dai, Minrui Zou, Yuxuan Li, Xiang Li, Kang Ni, and Jian Yang. Denodet: Attention as deformable multi-subspace feature denoising for target detection in sar images. *IEEE Transactions on Aerospace and Electronic Systems*, 61(2):4729–4743, 2025. doi: 10.1109/TAES.2024.3507786.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–959, 2017.

Zixian Huang, Zhirui Pan, and Bin Lei. What, where, and how to transfer in SAR target recognition based on deep CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2324–2336, 2020. doi: 10.1109/TGRS.2019.2947634.

Wei Li, Yunhao Gao, Mengmeng Zhang, Ran Tao, and Qian Du. Asymmetric feature fusion network for hyperspectral and sar image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):8057–8070, 2023. doi: 10.1109/TNNLS.2022.3149394.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, 2022.

Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. SARDet-100K: Towards open-source benchmark and toolkit for large-scale SAR object detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4079.

Junyan Lin, Feng Gao, Xiaocheng Shi, Junyu Dong, and Qian Du. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. doi: 10.1109/TGRS.2023.3331717.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017a.

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

Kang Ni, Minrui Zou, Yuxuan Li, Xiang Li, Kehua Guo, Ming-Ming Cheng, and Yimian Dai. Denodet v2: Phase-amplitude cross denoising for sar object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8142–8150, 2026. doi: 10.1609/aaai.v40i10.37761.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Xinyang Pu and Feng Xu. Enhancing SAR object detection with self-supervised pre-training on masked auto-encoders. In *IEEE International Geoscience and Remote Sensing Symposium*, 2025.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. doi: 10.1016/j.imavis.2021.104117.

Xian Sun, Zhirui Wang, Yuanrui Sun, Wenhui Diao, Yue Zhang, and Kun Fu. AIR-SARShip-1.0: High-resolution SAR ship detection dataset. *Journal of Radars*, 8(6):852–862, 2019. doi: 10.12000/JR19097.

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.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019.

Junjie Wang, Wei Li, Yunhao Gao, Mengmeng Zhang, Ran Tao, and Qian Du. Hyperspectral and sar image classification via multiscale interactive fusion network. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10823–10837, 2023a. doi: 10.1109/TNNLS.2022.3171572.

Zhirui Wang, Yuzhuo Kang, Xuan Zeng, Yuelei Wang, Ting Zhang, and Xian Sun. SAR-AIRcraft-1.0: High-resolution SAR aircraft detection and recognition dataset. *Journal of Radars*, 12(4):906–922, 2023b. doi: 10.12000/JR23043.

Shunjun Wei, Xiangfeng Zeng, Qizhe Qu, Mou Wang, Hao Su, and Jun Shi. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *IEEE Access*, 8:120234–120254, 2020. doi: 10.1109/ACCESS.2020.3005861.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

SUPPLEMENTARY MATERIAL

A SUPPLEMENTARY OVERVIEW

This supplementary document provides additional technical details and experimental evidence for SSDenoDet, including MAESSI pretraining, HSI band-reduction evidence, YAMAL validation, dataset-isolated detection analyses, and the MAE reconstruction protocol. The main paper remains self-contained, and none of its central claims depends on this supplementary material.

B MAESSI PRETRAINING AND HSI ANALYSIS

B.1 PRETRAINING SETUP

MAESSI denotes the complete Masked AutoEncoder with Spatial–Spectral Integration for Hyperspectral Imagery, including its reconstruction decoder. We denote the encoder extracted after HSI pretraining by f_M. SSDenoDet freezes f_M as a training-only teacher and never transfers the MAESSI decoder. The full autoencoder is pretrained on a multi-sensor hyperspectral corpus of approximately 800K spectra. Spectra are sampled from local 10×10 regions with sensor-stratified sampling to reduce source imbalance; sources span 157–426 bands and 380–2510 nm. Table S1 summarizes this corpus.

Sensor	Platform	Bands	Range (nm)	Spectra
AVIRIS	Airborne	194	400–2500	267K
NEON AOP	Airborne	426	380–2510	133K
EO-1 Hyperion	Satellite	157	427–2355	133K
SpectralEarth	Satellite	202	418–2445	267K

Table S1: Hyperspectral corpus used for MAESSI pretraining. Dataset-specific details are kept in the supplementary material because SAR detection is the main task.

During pretraining, an adjacent sequence of $B_t = 80$ channels is randomly selected from each spectrum. The selected physical wavelengths are passed to the positional encoding module, so identical token indices can correspond to different spectral locations. The encoder uses eight Transformer blocks, latent width $d = 128$, 16 attention heads, and roughly 1.5M parameters. Optimization uses AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, initial learning rate 10^{-4}, weight decay 0.05, batch size 256, ten warm-up epochs, cosine decay, FP16 mixed precision, and early stopping after 30 stagnant validation epochs. Training converges at epoch 501 in the recorded run. An NT-Xent term with weight 0.1 and temperature 0.5 is added to encourage nearby latent representations for spectra with similar patterns.

B.2 STRUCTURED NOISE CASES FOR ROBUST PRETRAINING

The frozen encoder f_M used by SSDenoDet is selected from a corruption-aware MAESSI pretraining/evaluation regime. This matters for SAR transfer because speckle, sparse bright scatterers, and sensor artifacts require a teacher robust to more than clean spectral reconstruction. We evaluate four progressively harder HSI noise cases.

Case 1 adds i.i.d. zero-mean Gaussian noise with variance 70 to all bands. **Case 2** combines non-i.i.d. Gaussian noise with band-dependent variance sampled between 10 and 70 and impulse noise sampled from a zero-mean Laplacian distribution with density in $[0.1, 0.7]$. **Case 3** adds stripe corruption to Case 2: 30% of bands are randomly affected, each selected band contains 10–15 stripes, and stripe values are sampled between 0.6 and 0.8. **Case 4** replaces stripes with dead lines: 30% of bands are affected, each selected band has 10–15 dead lines, and their width is 1–3 pixels. We detach the SAR teacher from the checkpoint selected under Case 4. During corruption-aware pretraining, a harder auxiliary setting also composes Case 3 stripes and Case 4 dead lines; this composition broadens exposure but is not mislabeled as a fifth row in Table S2.

Case	Metric	AE	Deep AE	DAE	CAE	VAE	VQ-VAE	C2AE	MAESSI
Case 1	PSNR↑	28.821	29.503	27.319	27.092	28.391	29.131	31.071	31.149
	SSIM↑	0.951	0.957	0.932	0.928	0.945	0.946	0.968	0.972
	SAM↓	0.102	0.098	0.147	0.109	0.118	0.099	0.081	0.079
Case 2	PSNR↑	22.241	24.831	25.912	26.842	27.513	28.476	28.659	28.903
	SSIM↑	0.809	0.872	0.901	0.921	0.938	0.943	0.952	0.959
	SAM↓	0.226	0.247	0.195	0.111	0.127	0.124	0.107	0.101
Case 3	PSNR↑	22.098	23.731	25.492	26.645	27.053	28.071	28.148	28.629
	SSIM↑	0.823	0.868	0.905	0.927	0.935	0.938	0.949	0.957
	SAM↓	0.216	0.251	0.194	0.114	0.135	0.129	0.108	0.102
Case 4	PSNR↑	22.011	24.281	25.501	26.493	27.267	27.968	28.651	28.729
	SSIM↑	0.851	0.856	0.890	0.922	0.929	0.939	0.951	0.956
	SAM↓	0.222	0.261	0.201	0.113	0.131	0.127	0.104	0.096

Table S2: HSI denoising robustness under four corruption cases. PSNR/SSIM are higher-is-better and SAM is lower-is-better. MAESSI remains strongest under the severe dead-line case used to select the teacher encoder for SAR transfer.

Table S2 shows that the complete MAESSI autoencoder is consistently strongest or tied for strongest under these corruptions, including Case 4. We extract f_M from this checkpoint so YAMAL receives a robustness-oriented target rather than a clean-only embedding.

B.3 Wavelength-Aware HSI Tokenization

An observed HSI band provides a scalar intensity but its token index alone does not identify the same physical quantity across sensors. MAESSI therefore applies a shared linear value embedding and adds a positional code derived from the measured wavelength rather than the ordinal band index. Two sensors can consequently assign different positional codes at the same ordinal index when their wavelengths differ. The resulting token is the sum of the intensity embedding and wavelength code, allowing one encoder to process varying band counts and sampling grids without conflating spectrally distinct measurements.

Beyond this value embedding, MAESSI adds a single positional code derived from the measured wavelength, computed via sinusoidal encoding as described above. No spatial position or learnable positional parameter is used, since each token corresponds to one spectral measurement rather than a spatial image patch.

B.4 MAESSI Encoder Structure

Main-paper Figure 2 details the block retained in f_M. Given tokens arranged over spatial locations and spectral groups, MAESSI forms two views of the same tensor. Spectral attention compares band groups within a location, while spatial attention compares locations within a band group. Their fused residual update therefore exposes the teacher embedding to both inter-band continuity and local structure without requiring a sensor-specific channel count.

B.5 HSI Band-Reduction Evidence

Indian Pines experiments indicate that a compact selected-band representation retains nearly the same Random Forest performance as the full 220-band representation. Reducing the input from 220 to 55 bands yields a $4\times$ channel reduction while changing robust-RF accuracy by only -0.0010 and macro-F1 by -0.0067. This supports the 80-band stochastic window used by MAESSI: the model does not need every raw channel at once if the selected channels preserve local spectral continuity and are paired with wavelength-aware positional encoding. The class distribution is also included (Table S4) because the benchmark is strongly imbalanced and therefore macro-F1 is a more informative companion to accuracy. See Table S3 for full experiment details.

Input / model	Bands	Reduction	Accuracy	Macro-F1
Full spectrum + robust RF	220	1.00×	0.8699	0.8566
Selected bands + robust RF	55	4.00×	0.8689	0.8499
Selected bands + LR	55	4.00×	0.8104	0.8244
Accuracy drop, RF	–	–	-0.0010	-0.0067

Table S3: Indian Pines band reduction summary from the auxiliary analysis. RF denotes Random Forest, and LR denotes logistic regression.

Class	ID	Samples
Alfalfa	1	46
Corn-notill	2	1428
Corn-mintill	3	830
Corn	4	237
Grass-pasture	5	483
Grass-trees	6	730
Grass-pasture-mowed	7	28
Hay-windrowed	8	478
Oats	9	20
Soybean-notill	10	972
Soybean-mintill	11	2455
Soybean-clean	12	593
Wheat	13	205
Woods	14	1265
Buildings-Grass-Trees-Drives	15	386
Stone-Steel-Towers	16	93

Table S4: Indian Pines class distribution used in the auxiliary band-selection analysis.

B.6 HSI Embedding Separability

We also inspect whether MAESSI embeddings improve simple nonparametric class separation before transferring the encoder to SAR. A 5-fold kNN probe is run on raw spectra and MAESSI embeddings. Table S5 reports macro-F1 before and after incorporating MAESSI embeddings, improving average macro-F1 from 0.8884 to 0.9255 across five HSI datasets, with the largest gain on Indian Pines.

Dataset	Raw kNN F1	MAESSI kNN F1	Gain
Indian Pines	0.7518	0.8904	+0.1386
Botswana	0.9271	0.9446	+0.0175
Pavia University	0.8873	0.9011	+0.0138
Pavia Centre	0.9359	0.9436	+0.0077
HyRank Dioni	0.9400	0.9478	+0.0078
Average	0.8884	0.9255	+0.0371

Table S5: Raw spectra versus MAESSI embedding kNN macro-F1 across five HSI datasets.

Figure S1 complements the kNN scores with matched raw-spectrum and MAESSI projections for all five HSI datasets. Across all five HSI datasets, MAESSI embeddings form more compact, better-separated class neighborhoods than raw spectral vectors under t-SNE visualization.

B.7 MAE Reconstruction Evaluation Protocol

We evaluate the MAESSI decoder checkpoint on held-out HSI spectra, its native pretraining domain. This experiment verifies that the transferred encoder comes from a checkpoint that preserves spectral structure under masked reconstruction; it is not used as a surrogate metric for SAR detection. We report the fixed checkpoint evaluation (Table S6) rather than a second optimization curve because neither the decoder nor its reconstruction loss is optimized or used for checkpoint selection anywhere in SSDenoDet.

Dataset	LPIPS ↓	SSIM ↑	PSNR ↑
Indian Pines	0.025	0.977	39.001
Botswana	0.037	0.977	37.265
Pavia University	0.028	0.976	38.970
Pavia Centre	0.045	0.972	34.751
HyRank Dioni	0.041	0.972	37.108

Table S6: HSI masked reconstruction with the MAESSI decoder checkpoint.

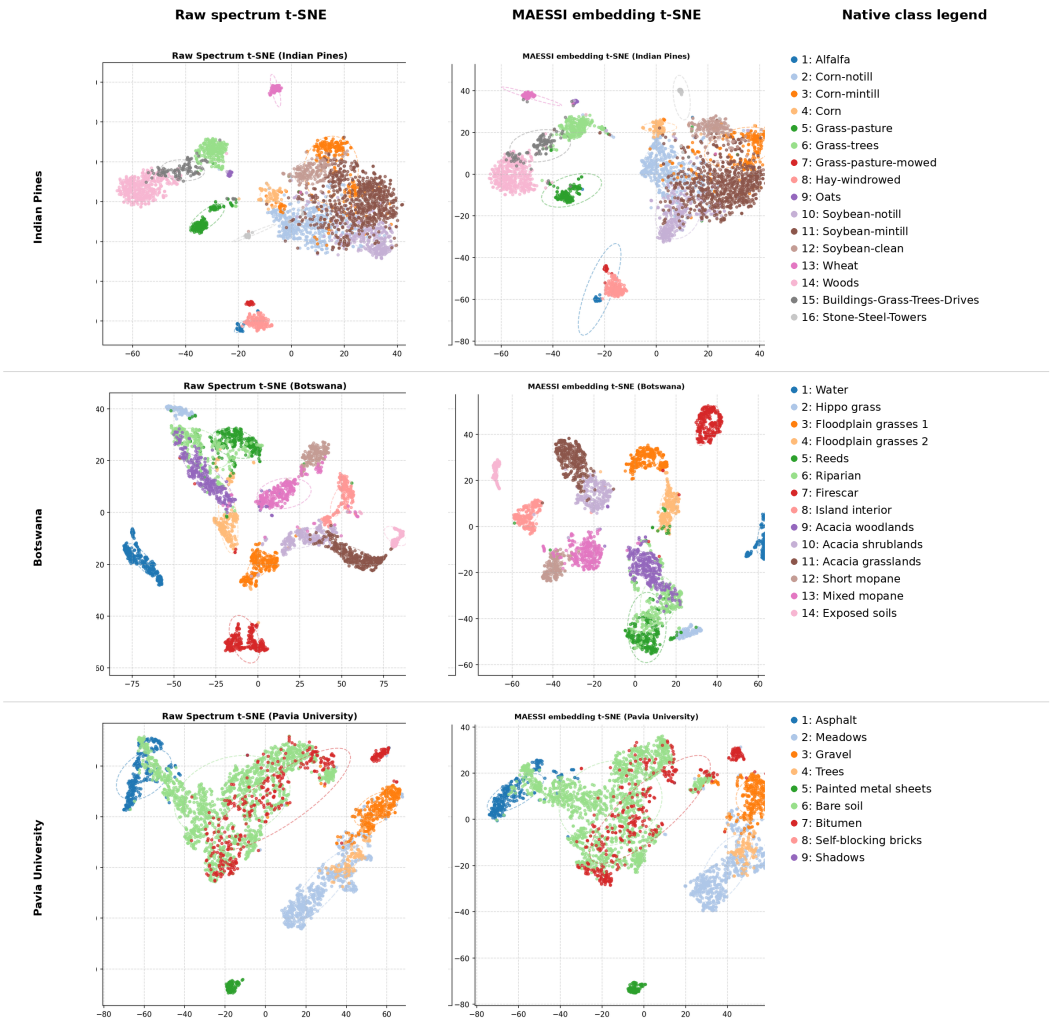


Figure S1: HSI latent-space diagnostics (first panel). Raw-spectrum and MAESSI t-SNE projections use identical samples and native class colors. Indian Pines, Botswana, and Pavia University are shown here; Pavia Centre and HyRank continue on the next page.

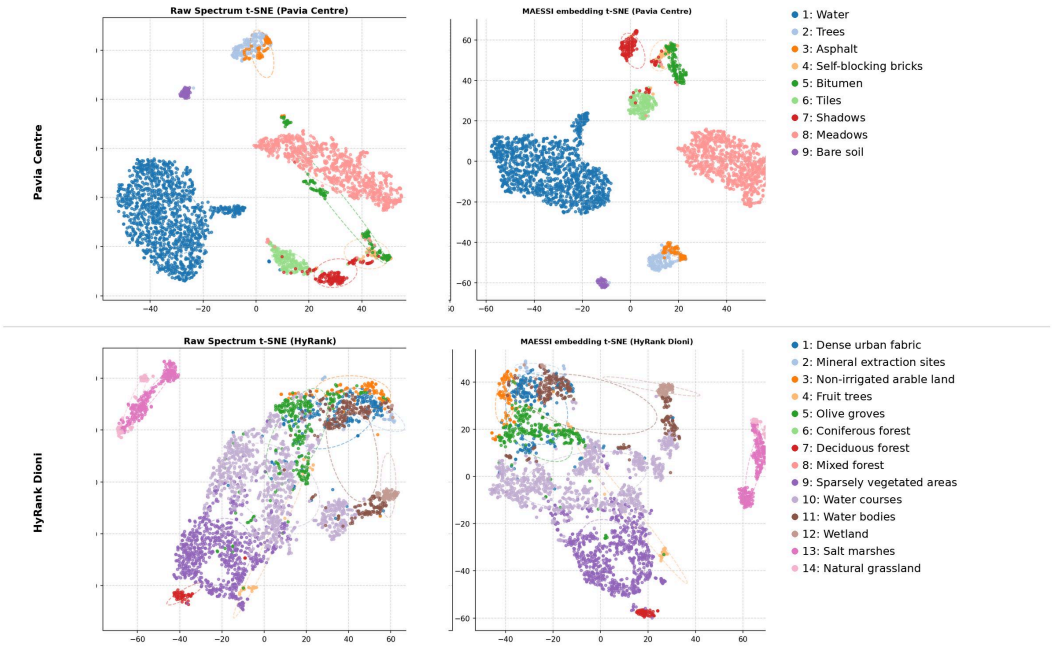


Figure S1: HSI latent-space diagnostics (continued): Pavia Centre and HyRank. The dedicated legend column keeps class labels separate from the projections.

C YAMAL Distillation and Representation Diagnostics

C.1 Distillation Setup and Collapse Diagnostics

YAMAL is a grayscale SAR image encoder. We optimize eight epochs of 300 steps with SAR batch size 128 and HSI batch size 16 at 224×224 resolution. The student uses AdamW with peak learning rate 2×10^{-4}, a 0.15 backbone learning-rate multiplier, weight decay 0.05, and cosine decay. The EMA momentum increases from 0.996 to 1.0. Multiplicative Gamma speckle, contrast, blur, and flips form two global SAR views; local crops are excluded because two crops can omit or contain different tiny targets. Teacher projection logits are centered by a 0.9-momentum running mean before the temperature-0.07 softmax, while the student uses temperature 0.10. Centering and EMA updates are stop-gradient operations and consume neither image labels nor boxes.

The unanchored objective reached a lower training loss but reduced held-out macro-F1 and effective rank. We therefore select the anchored run using a held-out representation probe rather than its self-supervised loss. On the same 4,000 train and 4,000 held-out images, globally pooled C5 features from YAMAL raise accuracy by 0.85 and macro-F1 by 3.15 percentage points over the ImageNet initialization. The top-256 effective rank rises from 170.76 to 191.61. Using the transplanted C5 feature on both sides avoids comparing a trained YAMAL projection head with a random ImageNet head. Image-level target labels are used only to evaluate this frozen kNN probe; no labels update either encoder.

The matched probe is summarized in main-paper Table 3. Figures S2 and S3 provide the corresponding convergence and latent-space diagnostics.

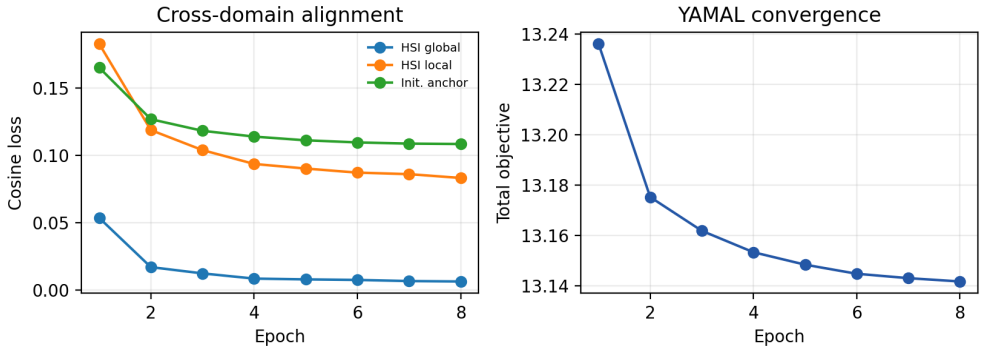


Figure S2: YAMAL distillation convergence. Local/global f_M alignment and centered SAR view consistency are monitored together; held-out macro-F1 and effective rank guard against collapse.

C.2 Distillation-Loss Ablation

The loss ablation changes one mechanism at a time while fixing the sampled SAR/HSI manifests, augmentation budget, optimizer, EMA schedule, and random seeds. Frozen C5 kNN accuracy, macro-F1, and effective rank diagnose whether a lower self-supervised objective hides dimensional collapse. Only checkpoints that complete this representation audit are transplanted into the matched SARBench detector, and all detector decisions use validation rather than test annotations. main-paper Table 5 reports the staged component additions; Table S7 retains the more detailed one-factor loss controls.

D Dataset-Isolated Detection Evaluation

D.1 Dataset Protocols and Validation-Only Selection

Every row in main-paper Table 4 has its own training loader, validation-based configuration selection, and untouched test loader. SARDet-100K uses the complete provided split. AIR-SARShip is split at scene level before 512×512 tiling with 200-pixel overlap, preventing chips from the same

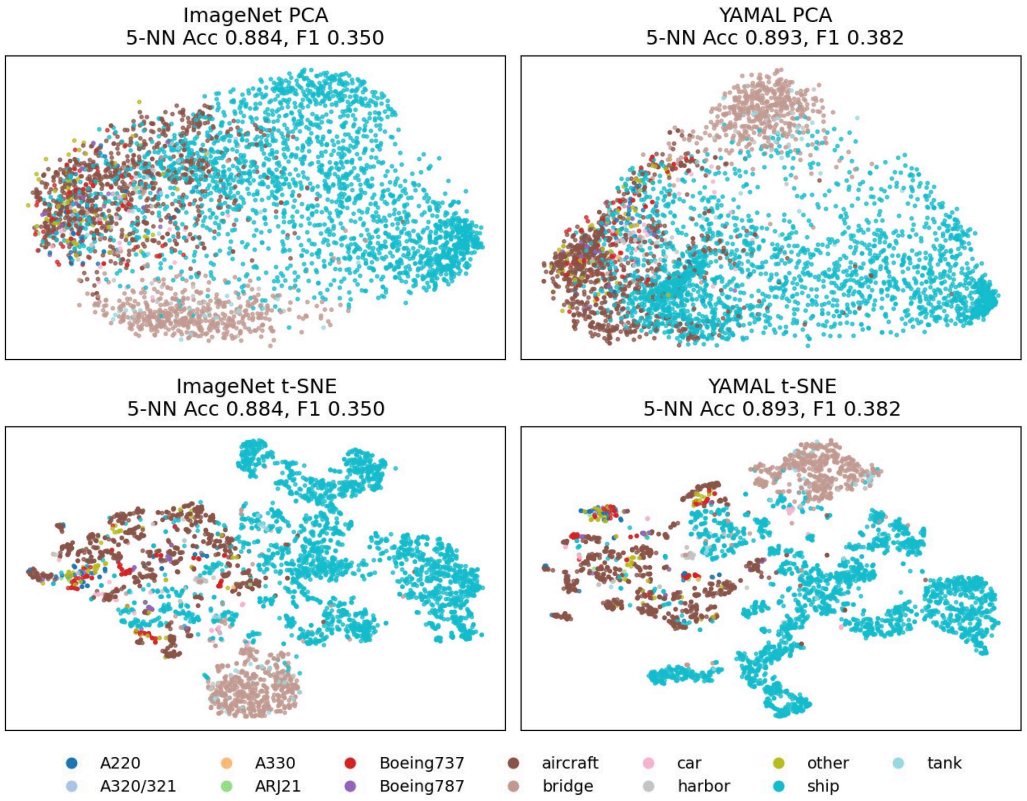


Figure S3: Matched SAR representation diagnostic before box-supervised fine-tuning. ImageNet and YAMAL receive identical grayscale images and sampling. Distillation improves minority-sensitive kNN macro-F1 without making global separation artificially clean.

Distillation variant	kNN Acc.	Macro-F1	Eff. rank	Val. mAP
Full YAMAL	0.932	0.438	945.09	58.71
− teacher centering	0.931	0.426	967.25	58.48
− VICReg	0.928	0.424	953.31	58.78
− initialization anchor	0.917	0.404	1020.00	57.97
− local HSI alignment	0.929	0.436	962.85	57.98
− all HSI alignment	0.930	**0.469**	**1098.22**	58.61

Table S7: One-factor YAMAL loss ablation. The final column uses the same SARBench validation detector as main-paper Table 5; representation probes use the same held-out split.

3000×3000 scene from crossing development splits. Once scale, overlap, and epoch count are fixed, the 17 training and four validation scenes are combined for the final fit; the ten public test scenes remain untouched. SAR-AIRcraft uses small-object chip initialization but returns to the benchmark's native image unit for final fitting and test. SARBench retains its native image unit. Multi-class runs use inverse-square-root image sampling capped at $6\times$ to expose rare classes without repeatedly drawing only the few rarest scenes.

We intersect each annotation with the visible raster before detector training. In SARDet-100K, 7,804 axis-aligned boxes cross an image boundary: 7,719 partially visible boxes are clipped symmetrically and 85 fully invisible boxes are excluded, leaving the 245,568 visible instances reported in main-paper Table 2. Direct comparison with DenoDet instead uses the unmodified official validation JSON (22,703 annotations), so the geometry audit and benchmark-native evaluator are never silently mixed.

SARDet-100K saturates after 12 base epochs and two visible-box refinement epochs. Close targets reveal a proposal-suppression bottleneck: with RoI NMS fixed at 0.7, raising RPN NMS from 0.7 to 0.95 improves official-box validation mAP from 56.48 to 56.77. SSDenoDet-E then fuses three already-saved saturation checkpoints. Consensus scoring penalizes boxes unsupported by the other checkpoints, while score-weighted coordinate averaging reduces localization jitter. Fusion IoU 0.70 is selected on validation and frozen before test. It reaches 57.55 validation mAP and improves official test mAP from 54.51 to 55.41 (main-paper Table 4), but remains below the reported DenoDet V2 test value; the paper therefore claims a validation improvement only. The corresponding inference-selection sweep on the official box is provided in Table S8.

The same selection discipline is applied independently elsewhere rather than copying the SARDet policy across domains. SARBench uses three checkpoints and a validation-frozen fusion IoU of 0.55. AIR-SARShip and SAR-AIRcraft retain one checkpoint and combine identity, horizontal-flip, and vertical-flip outputs with class-aware NMS at IoU 0.50; overlapping ship chips are subsequently merged in original-scene coordinates at the benchmark-native threshold. Consequently, "E" denotes a declared validation-selected consensus family, not one test-tuned ensemble shared by all datasets.

Model	Validation-only policy	mAP	AP_S	AP_M
SSDenoDet	RoI NMS 0.7 / RPN NMS 0.7	56.48	49.60	66.11
SSDenoDet	RoI NMS 0.7 / RPN NMS 0.8	56.66	49.82	66.19
SSDenoDet	RoI NMS 0.7 / RPN NMS 0.9	56.71	49.90	66.28
SSDenoDet	RoI NMS 0.7 / RPN NMS 0.95	56.77	49.89	66.37
SSDenoDet-E	RPN 0.9 / Fusion IoU 0.65	57.51	50.49	67.26
SSDenoDet-E	RPN 0.9 / Fusion IoU 0.70	57.53	50.50	67.27
SSDenoDet-E	**RPN 0.95 / Fusion IoU 0.70**	**57.55**	**50.52**	**67.27**

Table S8: Official-box SARDet-100K inference selection. Every row is computed on validation; bold marks the policy frozen before the single test evaluation. SSDenoDet-E fuses three saturation checkpoints and is reported with its full three-model cost.

D.2 Protocol-Matched Comparison with DenoDet

Table S10 separates native VOC evaluation from the common COCO diagnostic. Predictions from overlapping AIR chips are first translated to original-scene coordinates and merged class-wise with the GrokSAR NMS threshold of 0.1; AP then uses IoU 0.5 and MMDetection's legacy-coordinate convention. SAR-AIRcraft is evaluated on its original test images with the same VOC convention. SSDenoDet-E exceeds DenoDet V2 on SAR-AIRcraft mAP(12) by 0.32 points, is within 0.05 on mAP(07), and remains below DenoDet V2 on AIR-SARShip. These native results, rather than chip-level COCO AP_{50}, determine the direct DenoDet comparison.

Fig. S4 verifies validation convergence.

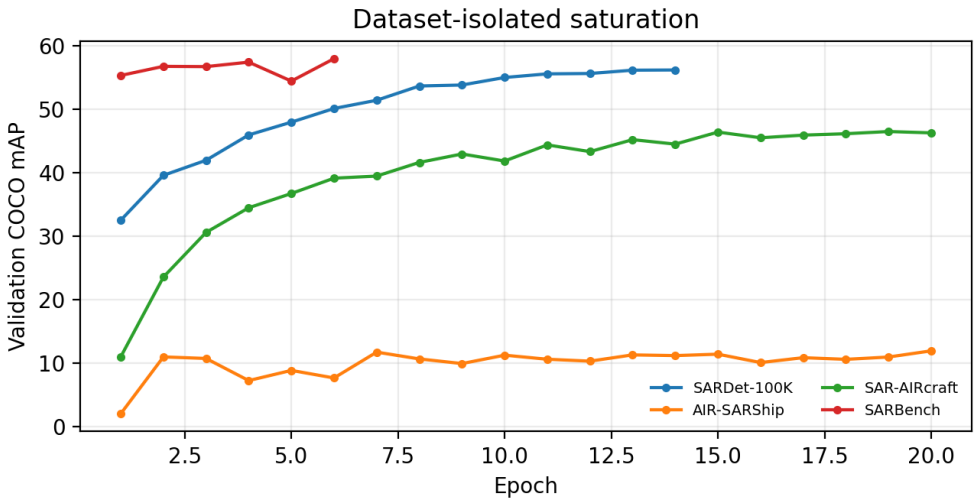


Figure S4: Validation convergence of the four dataset-isolated detectors. Curves are kept separate because epochs, category balance, and image units differ across benchmarks. All checkpoint decisions use validation only.

Dataset	Class	Test boxes	mAP	AP_{50}	AP_{75}
SARDet-100K	aircraft	6778	46.55	81.33	47.22
	bridge	3281	38.68	79.43	32.08
	car	1230	63.56	97.46	75.59
	harbor	399	78.28	97.85	92.11
	ship	10738	53.27	90.78	57.23
	tank	1771	48.94	82.63	48.45
AIR-SARShip-1.0	ship	204	29.73	69.95	15.66
SAR-AIRcraft-1.0	A220	460	23.14	40.17	21.91
	A320/321	52	45.52	68.50	46.29
	A330	31	57.28	75.05	62.80
	ARJ21	362	20.02	37.51	17.56
	Boeing737	550	12.43	24.98	10.89
	Boeing787	454	27.89	43.10	29.65
	other	1041	27.90	53.58	24.21
SARBench	ship	5922	59.27	88.85	68.00

Table S9: Per-class full-test COCO AP for the four independently trained YAMAL/FPN detectors. Instance counts expose rare-class uncertainty that aggregate mAP can obscure.

Table S9 reports per-class test AP for four dataset-isolated detectors. Test-box counts vary across classes even within the same dataset; SARDet-100K has 399 harbor boxes and 10738 ship boxes. Class-level AP is therefore comparatively less stable than the dataset-level aggregate in main-paper Table 4.

Method	Dataset	Metric	Main	AP_S	AP_M	FLOPs
DenoDet V1 Dai et al. (2025)	SARDet-100K val	COCO mAP	55.88	50.63	68.47	–
DenoDet V2 Ni et al. (2026)	SARDet-100K val	COCO mAP	56.71	51.45	68.75	52.47G
DenoDet V2 Ni et al. (2026)	SARDet-100K test	COCO mAP	56.39	–	–	52.47G
DenoDet V2 Ni et al. (2026)	SAR-AIRcraft-1.0	mAP(07)	65.30	–	–	0.21T
DenoDet V2 Ni et al. (2026)	SAR-AIRcraft-1.0	mAP(12)	70.73	–	–	0.21T
DenoDet V2 Ni et al. (2026)	AIR-SARShip-1.0	mAP(07)	68.92	–	–	48.61G
DenoDet V2 Ni et al. (2026)	AIR-SARShip-1.0	mAP(12)	74.86	–	–	48.61G
SSDenoDet-E	SAR-AIRcraft-1.0	mAP(07)	69.88	–	–	–
SSDenoDet-E	SAR-AIRcraft-1.0	mAP(12)	**71.05**	–	–	–
SSDenoDet-E	AIR-SARShip-1.0	mAP(07)	71.98	–	–	–
SSDenoDet-E	AIR-SARShip-1.0	mAP(12)	72.36	–	–	–

Table S10: Protocol-matched comparison with DenoDet. DenoDet values are paper-reported; SSDenoDet-E uses original-image test annotations and legacy-coordinate VOC evaluation. A dash denotes an unavailable scale-specific value, not zero.

D.3 Backbone, Resolution, and Tiling Controls

The matched backbone control in Table S11 attributes the large 512-to-800 gain primarily to scale-aware detection and treats cross-domain distillation as a modest representation effect.

Backbone	Input	mAP	AP_{50}	AP_{75}	AP_S	AP_M
ImageNet	512	58.47	88.56	66.07	46.59	71.76
YAMAL	512	55.48	85.87	62.05	42.73	69.84
ImageNet	800	58.67	88.73	67.52	49.52	69.87
YAMAL	800	58.98	89.05	67.39	48.73	70.64
YAMAL-C	800	59.27	88.85	68.00	49.11	71.01

Table S11: Matched SARBench validation control. The split, FPN head, optimizer, and checkpoint rule are fixed. These common-COCO validation values diagnose initialization and resolution; they are not the benchmark-native headline scores in main-paper Table 1.

At 512 pixels, ImageNet initialization is stronger. At 800 pixels, YAMAL closes the gap and YAMAL-C reaches 59.27 mAP, while ImageNet retains a small AP_S advantage. The larger within-YAMAL improvement from 512 to 800 supports target-scale preservation as the dominant small-object factor.

Table S12 tests whether patch magnification alone explains the small-target gains. Its validation result supports full-image high-resolution inference, which retains context and avoids duplicate boundary proposals.

SARBench validation inference	mAP	AP_{50}	AP_S	AP_M
Full image, 800 px	**57.95**	**87.92**	**48.77**	**69.66**
Full + 512-px overlapping tiles	53.49	84.08	43.75	65.54

Table S12: Validation-only inference ablation. Overlapping tiles enlarge targets but lose context and create duplicate proposals, so the untouched test set is evaluated with full images only.

E Qualitative Detection Evidence

The main paper's qualitative comparison shows matched detector outputs on AIR-SARShip-1.0 and SARDet-100K. Figure S5 shows matched detector outputs on SARBench, AIR-SARShip-1.0 and SARDet-100K complementing the main paper's qualitative comparison.

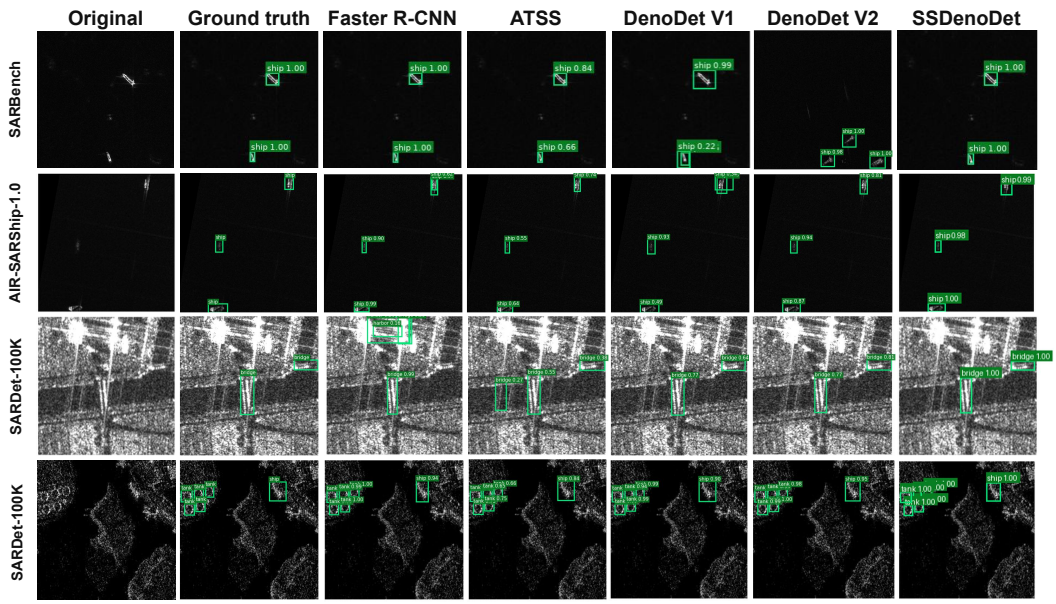


Figure S5: Qualitative detection comparison on SARBench, AIR-SARShip-1.0, and SARDet-100K. Faster R-CNN and ATSS represent generic detectors, DenoDet V1/V2 are our matched reproductions, and SSDenoDet is our cross-domain model.