PROCREDET: PROGRESSIVE CREDIBLE SAMPLE SELECTION FOR SPARSELY ANNOTATED UNDERWATER OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparsely Annotated Object Detection (SAOD) is designed to address the high cost of dataset annotation. In underwater scenarios, its inherent challenges are particularly pronounced: variable underwater illumination conditions lead to significant discrepancies in instance features across different images, the mixing of unlabeled positive samples into negative sets disrupts the training logic, and the model tends to generate numerous misclassifications and struggles to effectively preserve the inter-class correlation information of targets. These factors collectively impede the improvement of model performance. To tackle these challenges, we propose ProCreDet, a teacher-student framework based on progressive training with credible samples. This framework employs a complementary feature learning mechanism to handle diverse underwater illumination conditions for positive sample mining, integrates a risky sample filtering module to resolve the issue of unlabeled positive samples being mixed into negative samples, and introduces a conflict mitigation module to suppress false predictions while preserving inter-class relationship information, thereby enhancing the generalization ability of the model.

1 INTRODUCTION

Underwater object detection is crucial for marine resource exploration Fu et al. (2023). However, fully annotating densely grouped underwater organisms is labor-intensive and costly. Sparsely Annotated Object Detection (SAOD) reduces annotation cost by training on partially annotated images.

Existing SAOD methods reduce the impact of erroneous supervision signals or mine pseudo-labels through geometric sampling Niitani et al. (2019), background recalibration Zhang et al. (2020), teacher-student distillation Suri et al. (2023); Hinton et al. (2015), collaborative detectors Wang et al. (2021); Bromley et al. (1993); Wu et al. (2024), and entropy filtering Lin et al. (2026). Nevertheless, they largely ignore three key issues in underwater scenarios: (1) intra-class color variations caused by light attenuation and scattering; (2) unlabeled positives mixed into the negative set; and (3) underutilized inter-class correlations among category scores. As illustrated in Fig. 1(a), existing methods easily misclassify ambiguous unlabeled positives as negatives, resulting in unstable training curves.

To address these issues, we propose ProCreDet, a **Pro**gressive **Cre**dible sample selection network for sparsely annotated underwater object **Det**ection. Built on a teacher-student framework, ProCreDet progressively explores credible samples. Specifically, we design a Risky Sample Filtering (RSF) module in the teacher branch to dynamically partition positive/negative samples using labeled sample confidence distribution, mitigating false negative impacts. For varying underwater color conditions, the student model adopts a BiMask strategy to extract complementary features, enhancing positive identification and reducing misclassification penalties. Additionally, we develop a Conflict Mitigator (CM) to suppress teacher model false detections while preserving inter-class information for student guidance. In contrast, our method leaves undecided samples in a risky buffer and only accepts gradually credible positives, thereby stabilizing the learning process (Fig. 1(b)). Experiments verify that our method effectively suppresses the adverse effects of unlabeled data during training. As quantified in Table 1, representative SAOD methods Zhang et al. (2020); Wu et al. (2024); Chen et al. (2025); Wang et al. (2021); Suri et al. (2023); Lin et al. (2026) treat numerous unlabeled positives as negatives during training, leading to severe performance fluctuations—their performance

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.

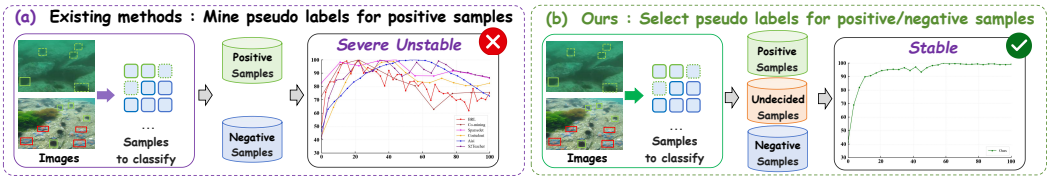

Figure 1: Motivation of progressive credible sample selection for underwater SAOD. (a) Existing methods force every sample into an immediate positive/negative decision, resulting in unstable training curves. (b) Our method places undecided samples into a risky buffer and only accepts gradually credible positives, yielding stable evaluation curves.

Table 1: Post-peak performance drop on UODD with 70% annotations removed (lower is better).

Method	RetinaNet	BRL	Co-mining	SparseDet	Co-Student	Aisi	S2Teacher	Ours
Drop (↓)	13.6	14.2	13.3	5.5	6.5	11.3	5.0	**0.6**

drops by 5.0–14.2 mAP after the training peak (i.e., the maximum gap between the peak mAP and any subsequent evaluation). By comparison, our method stabilizes after a certain number of epochs, with a drop of only 0.6 mAP. Our contributions are summarized as follows:

- For underwater SAOD, we propose a progressive credible sample selection teacher-student network with a BiMask strategy for complementary feature extraction, boosting positive identification across varying instances.
- We design a Risky Sample Filtering (RSF) module to dynamically partition positive/negative intervals, mitigating the adverse impact of misclassifying potential positives as negatives on training.
- A Conflict Mitigator module is designed to resolve prediction-ground truth inconsistency and fully exploit inter-class confidence correlations.
- Tailored for underwater scenarios, ProCreDet outperforms state-of-the-art methods. Under 70% sparsity, ProCreDet improves mAP by 7.6% on UODD and 4.9% on DUO over the previous state of the art, achieving an average gain of 6.3% across the two datasets.

2 Related Work

2.1 Sparsely Annotated Object Detection

Existing SAOD methods were mostly designed for generic scenarios such as COCO Lin et al. (2014) with only a few sparse targets per image, and addressed the problem by either suppressing erroneous supervision or generating pseudo-labels Niitani et al. (2019); Zhang et al. (2020); Wang et al. (2021); Suri et al. (2023); Wu et al. (2024); Lin et al. (2026). Early works focused on reducing training bias from unlabeled instances Niitani et al. (2019); Zhang et al. (2020), while recent approaches mine pseudo-labels via collaborative mining Wang et al. (2021), teacher-student distillation Suri et al. (2023), co-student detectors Wu et al. (2024), or entropy filtering Lin et al. (2026). Nevertheless, these methods overlook the harm of misclassifying potential positives as negatives and ignore inter-class correlations among prediction scores.

Semi-supervised object detection (SSOD) differs from SAOD in the granularity of missing annotations: SSOD drops entire image labels, whereas SAOD faces instance-level missing annotations within labeled images. Consequently, unlabeled positives in SAOD are mixed into the negative set and corrupt the training signal, a problem SSOD methods do not address, so we exclude SSOD baselines.

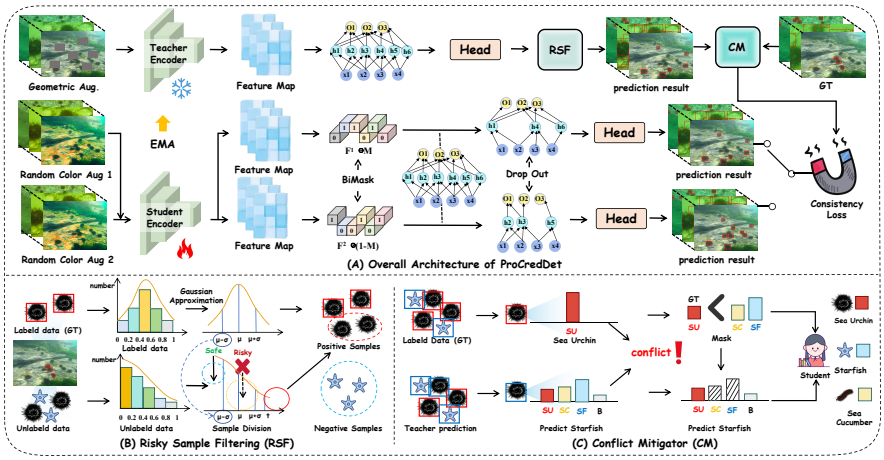

Figure 2: Overview of the proposed ProCreDet method: (a) the overall architecture of ProCreDet, (b) the Risky Sample Filtering (RSF) module, and (c) the Conflict Mitigator (CM) module.

2.2 Underwater Object Detection

Most existing underwater detection methods rely on fully supervised data and improve performance via image quality enhancement Xu et al. (2023); Jian et al. (2024), either through generative restoration Yu et al. (2018); Ge et al. (2022) or feature-level fusion Jiang et al. (2021); Zhang et al. (2025); Chen et al. (2023); Wen et al. (2025). Recently, sparsely annotated underwater detection has emerged, leveraging context and foreground distributions to generate pseudo-labels Chen et al. (2025). Unlike these methods, we focus on underwater SAOD and enhance adaptability to color variations through complementary feature learning rather than image restoration.

3 Methods

3.1 Overview

The architecture of the proposed ProCreDet, which is based on a teacher-student framework, is illustrated in Fig. 2. Let the original image be I_o and the annotated data be A_l. To increase the discrepancy between the two networks, we apply geometric augmentation to the teacher network and color augmentation to the student network, respectively. The geometrically augmented image I_g is fed into the teacher model. First, a Risky Sample Filtering (RSF) module dynamically filters out unlabeled positive samples mixed in negative ones. Subsequently, a Conflict Mitigator (CM) module is designed to prevent erroneous predictions from the teacher model from interfering with student training, yielding the final logits prediction L_t from the teacher branch.

For the student branch, to address insufficient mining of unlabeled positive samples caused by underwater color variations, we apply two random strong color augmentations with different intensities to I_o, generating I_{s1} and I_{s2}. Further, we design complementary features to represent targets under different underwater conditions. Their corresponding logits predictions are denoted as L_{s1} and L_{s2}. In each training iteration, consistency loss is computed between L_t and L_{s1}, and between L_t and L_{s2}, respectively. Finally, the teacher network is updated via Exponential Moving Average (EMA), while the student network is updated by backpropagation.

3.2 Risky Sample Filtering (RSF)

We filter potential positive samples from unlabeled data using the Risky Sample Filtering (RSF) module. As illustrated in Fig. 2(b), we first obtain the teacher model's prediction scores P_g on the augmented image I_g, and split the predictions into a labeled data prediction set P_p and an unlabeled data prediction set P_n based on the annotated data A_l. Specifically, the scores are the pre-NMS anchor-level classification confidences, and each labeled instance in P_p is matched to its ground

truth with an IoU of at least 0.5, taking the confidence of the corresponding ground-truth category. At every iteration, we re-estimate the mean μ and variance σ^2 of the Gaussian distribution over the confidence score set $S = \{s_i\}$ of the labeled instances in the current batch, without any temporal accumulation, where s_i is the confidence score of each detection result. The Gaussian is fitted over all categories jointly rather than per category. This global threshold is a practical simplification; its empirical validity is verified by the interval contamination analysis in Sec. 4.5.

The Gaussian fit provides a simple, data-driven heuristic to locate a dynamic boundary for the current training state. The confidence score is a direct and efficient way to explore potential positive samples in unlabeled data. Misclassifying potential positives as negatives severely degrades training, so we use $\mu - \sigma$ as a boundary, with samples to the right considered potential positives. As shown in Sec. 4.5, over 80% of batches pass a K-S test at the 0.05 level; more importantly, the resulting intervals exhibit the desired contamination/purity properties (Table 6), which is the empirical basis for the boundary choice.

To obtain more reliable positives, we set a high pseudo-label threshold τ to select high-confidence unlabeled samples as pseudo-labels (red circles in Fig. 2(b)). Samples with confidence between $\mu - \sigma$ and τ are ambiguous (mixed positives and negatives), and misclassification of samples in this range harms performance. We thus discard these "risky" unlabeled samples using 0-1 mask M_{Gsn}. Let $S_{\text{neg}} = \{s_j\}$ denote confidence scores of predictions in P_n. $M_{\text{Gsn}}[j]$ is defined as:

$$M_{\text{Gsn}}[j] = \begin{cases} 1, & \text{if } s_j \leq T_{\text{neg}}, \\ 0, & \text{if } s_j > T_{\text{neg}}, \end{cases} \tag{1}$$

where $T_{\text{neg}} = \mu - \sigma$.

The remaining unlabeled samples are relatively "safe" to serve as negatives (blue dashed circles in Fig. 2(b)). Since negatives contribute minimally to training, we randomly remove some via a Bernoulli distribution and generate a 0-1 mask M_{Bsn} to further filter out residual positives from the remaining unlabeled data:

$$M_{Bsn}[j] \sim \text{Bernoulli}(p), \tag{2}$$

where each pending sample is retained with probability p.

3.3 Complementary Mask (BiMask)

To address large feature discrepancies among labeled and unlabeled same-category underwater targets caused by color variations, we design a complementary mask (BiMask) in the student model, ensuring the student's predictions under different color conditions align with the teacher's.

We apply distinct augmentations to generate I_{s1} and I_{s2}, simulating varying illumination perturbations in underwater environments. To enhance feature discrepancy between the two student branches, we construct a BiMask M_{Bi} with the same size as the feature map: approximately half of its elements are set to 0 (the rest to 1) via a binomial distribution. Subsequently, the two feature maps (F_{s1} and F_{s2}) are decoded through Dropout layers masked by M_{Bi} and $1 - M_{Bi}$.

Since both feature maps are extracted by a shared encoder (with consistent channel characteristics), the complementary masks M_{Bi} and $1 - M_{Bi}$ yield two mutually exclusive feature sets with distinct semantic information:

$$F_{Bi1} = F_{s1} \odot M_{Bi}, \tag{3}$$

$$F_{Bi2} = F_{s2} \odot (1 - M_{Bi}). \tag{4}$$

The student model then generates predictions L_{s1} and L_{s2}, and consistency loss is computed between these predictions and the teacher's filtered predictions to enhance robustness.

Additionally, the student model generates prediction P_w from a conventionally weakly augmented image I_w. Using pseudo-labels A_p derived from the teacher's prediction P_g together with the annotated data A_l, the student calculates classification and regression losses.

3.4 Conflict Mitigator (CM)

As shown in Fig. 2(c), we develop the Conflict Mitigator module to prevent the student model from being misled by erroneous predictions generated by the teacher model, while preserving inter-class

information to enhance the model's generalization ability. We obtain two logits predictions L_{s1} and L_{s2} from the student model's heads, as well as the logits prediction L_t output by the teacher model. Let L_s denote the prediction output of the student model on the augmented image; we first load the annotation data A_l and generate a mask M_{an} with length equal to the number of anchor boxes (setting mask values to 1 for anchor boxes corresponding to positive samples, 0 otherwise). After applying this mask to the student and teacher predictions, only the logits predictions L_{ps} on positive samples are retained:

$$L_{ps} = L \odot M_{an}, \tag{5}$$

where L denotes either the student's prediction L_s or the teacher's prediction L_t. As shown in Fig. 2(c), we focus on the labeled data.

For the teacher's logits predictions on positive samples (L^t_{ps}), we generate a mask M_{ac} with length equal to the number of categories (setting mask values to 0 for categories with logits greater than the true category's logit, 1 otherwise). Mathematically, M_{ac} is defined as:

$$M_{ac}[k] = \begin{cases} 0, & \text{if } L^t_{ps}[k] > L^t_{ps}[k_{\text{true}}], \\ 1, & \text{otherwise}, \end{cases} \tag{6}$$

where k is the index of the category, and k_{true} is the index of the true category. The student model also uses the mask M_{ac} to ignore certain categories. As shown in Fig. 2(c), the teacher's predicted values for "Starfish" and "Sea Cucumber" are higher than those for the true category "Sea Urchin"; we thus mask these categories and force the student model to align with the remaining categories after masking, transferring valuable inter-class information from the teacher to the student. Taking one student prediction L_{s1} as an example, the consistency loss between the student and teacher predictions is calculated as follows:

$$\mathcal{L}oss_{\text{u}} = \mathcal{L}_{\text{kl}} \left(L_t \odot (1 - M_{\text{an}}), L_{s1} \odot (1 - M_{\text{an}}) \right), \tag{7}$$

For labeled samples, we unify the anchor mask and category mask to produce the final mask for generating high-quality and reliable teacher logits:

$$M_l = M_{\text{an}} \odot M_{\text{ac}}, \tag{8}$$

$$\mathcal{L}oss_{\text{l}} = \mathcal{L}_{\text{kl}} \left(L_t \odot M_l, L_{s1} \odot M_l \right), \tag{9}$$

The complete consistency loss for the first student branch then combines the labeled term with the RSF-filtered final unlabeled term:

$$\mathcal{L}oss_{\text{c1}} = \mathcal{L}oss_{\text{l}} + \mathcal{L}oss_{\text{u}} \odot M_{Gsn} \odot M_{Bsn}, \tag{10}$$

where $\mathcal{L}_{\text{kl}}$ computes the KL loss between two logit predictions (computed directly on logits without temperature scaling), M_l is the final mask for logit predictions of labeled data after unifying the M_{an} and M_{ac} masks, $\mathcal{L}oss_{\text{u}}$ is the consistency loss on the unlabeled data, $\mathcal{L}oss_{\text{l}}$ is the consistency loss on the labeled data, $\mathcal{L}oss_{\text{c1}}$ is the final consistency loss between L_t and L_{s1}, the same operations are applied to the other augmented version L_{s2}.

The full training objective combines the supervised detection loss on the weakly augmented student view with the two consistency regularization terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \lambda \cdot \frac{\mathcal{L}oss_{\text{c1}} + \mathcal{L}oss_{\text{c2}}}{2}, \tag{11}$$

where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{reg}}$ are the classification and regression losses computed from the annotated and pseudo-labeled data, $\mathcal{L}oss_{\text{c1}}$ and $\mathcal{L}oss_{\text{c2}}$ are the consistency losses of the two strong augmentation branches, and λ weights their average; we set $\lambda = 0.5$ by default.

Masking categories whose logits exceed the true category's prevents the student from aligning with incorrect categories under SAOD settings while preserving inter-class information.

4 Experiments

4.1 Datasets and Evaluation Metrics

Experiments are conducted on UODD Jiang et al. (2021) (2,560 train / 506 test images) and DUO Liu et al. (2021) (6,671 train / 1,111 test images). We remove $r\%$ of instance annotations per image to create sparse labels and report $\text{mAP}_{0.5:0.95}$, $\text{mAP}_{0.5}$, and $\text{mAP}_{0.75}$.

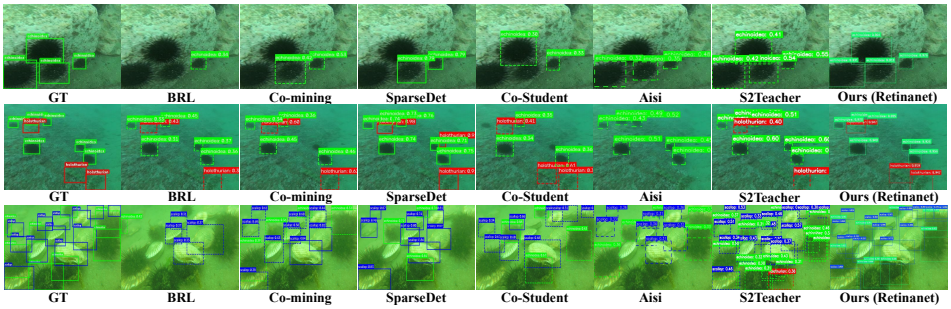

Figure 3: The visualization results on UODD with 70% annotations removed. The green, red, and blue bounding boxes in the figure represent the detected categories of *Echinoidea*, *Holothurian*, and *Scallop*, respectively. Detected instances with confidence above 0.7 are enclosed in solid-line boxes and those with confidence between 0.3 and 0.7 in dashed-line boxes.

Table 2: Performance comparison on UODD dataset under different sparsity ratios. The best and second-best results are shown in bold and underline, respectively.

Method	**Venue**	**30%**				**50%**				**70%**			
		mAP	mAP_{50}	mAP_{75}	RPR (%)	mAP	mAP_{50}	mAP_{75}	RPR (%)	mAP	mAP_{50}	mAP_{75}	RPR (%)
RetinaNet	ICCV 2017	39.6	71.9	40.3	80.1	35.0	69.2	32.0	71.6	31.4	62.8	26.7	64.2
FCOS	ICCV 2019	42.3	81.5	36.8	89.1	37.4	69.2	36.6	78.7	31.6	64.1	25.6	66.5
BRL (RetinaNet)	ICASSP 2020	42.6	82.4	42.8	87.1	40.3	77.9	36.7	84.8	36.2	72.4	29.9	74.0
Co-mining (RetinaNet)	AAAI 2021	40.4	75.6	38.2	82.6	35.9	67.3	32.5	73.4	31.3	66.4	21.4	64.0
SparseDet (Faster R-CNN)	ICCV 2023	41.9	74.4	43.4	86.6	34.7	62.8	34.9	71.7	30.2	74.3	15.1	62.4
Co-Student (FCOS)	ECCV 2024	44.1	80.1	43.4	92.8	40.6	75.3	37.7	85.5	33.1	66.7	28.3	69.7
Aisi (YOLO)	IEEE TGRS 2025	*48.1*	*88.3*	46.5	92.7	*46.0*	*85.6*	*44.9*	88.6	39.6	*79.0*	33.0	76.3
S2Teacher (FCOS)	AAAI 2026	44.7	83.5	*48.2*	94.1	42.5	77.2	42.3	89.5	37.7	77.0	32.1	79.4
Ours (RetinaNet)	-	46.7	83.5	48.1	*95.5*	44.2	80.8	44.2	*90.4*	*43.2*	78.9	*44.2*	*88.3*
Ours (YOLO)	-	**51.3**	**89.4**	**51.8**	**97.3**	**49.5**	**87.7**	**49.8**	**94.0**	**47.2**	**82.8**	**49.5**	**89.6**

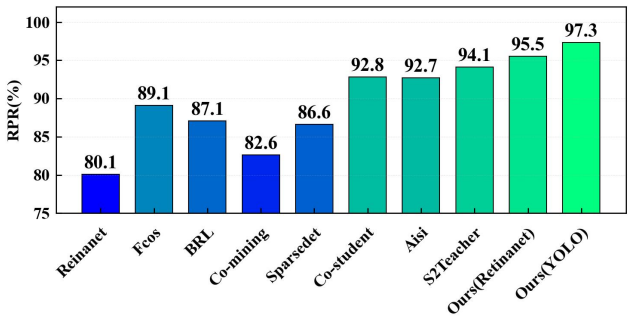

Figure 4: RPR performance for different methods on UODD with 30% annotations removed.

4.2 Experimental Settings

We implement ProCreDet on RetinaNet and YOLOv8, training for 30 epochs (batch size 2, Adam, 1×10^{-5}). All hyperparameters are shared across the UODD and DUO datasets. The teacher uses geometric augmentation, each student branch uses strong color distortion, and the teacher is updated via EMA ($\alpha = 0.99$). At inference, only the teacher is used, so no extra overhead is added.

4.3 Quantitative Experiments

Tables 2 and 3 evaluate ProCreDet on UODD (30% / 50% / 70% sparsity) and DUO (50% / 70%/ 90% sparsity), showing gains up to 7.6% on UODD and 4.9% on DUO.

Our method runs on RetinaNet Lin et al. (2017), Faster R-CNN Ren et al. (2016), FCOS Tian et al. (2019), YOLOv8, and YOLOX Ge et al. (2021). Because detector capacities differ (Table 4), raw

Table 3: Performance comparison on DUO dataset under different sparsity ratios. The best and second-best results are shown in bold and underline, respectively.

Method	Venue	50%				70%				90%			
		mAP	mAP_{50}	mAP_{75}	RPR (%)	mAP	mAP_{50}	mAP_{75}	RPR (%)	mAP	mAP_{50}	mAP_{75}	RPR (%)
RetinaNet	ICCV 2017	49.5	72.0	56.8	84.5	45.4	67.8	53.3	77.5	34.0	52.1	38.2	58.0
FCOS	ICCV 2019	52.5	72.8	60.1	83.6	47.9	69.3	54.7	76.3	34.3	55.4	38.8	54.6
BRL (RetinaNet)	ICASSP 2020	49.3	71.4	55.8	84.1	44.8	68.7	50.4	76.5	30.6	49.1	33.6	52.5
Co-mining (RetinaNet)	AAAI 2021	44.0	61.5	49.7	75.0	40.2	63.3	44.8	68.6	33.3	54.3	35.7	56.8
SparseDet (Faster R-CNN)	ICCV 2023	45.5	60.8	51.6	73.0	41.6	68.2	45.1	66.7	31.7	50.8	35.2	50.8
Co-Student (FCOS)	ECCV 2024	52.6	71.0	58.5	83.8	41.0	60.9	44.7	65.3	31.3	48.6	35.4	49.8
Aisi (YOLO)	IEEE TGRS 2025	*55.0*	*79.0*	*64.7*	88.7	49.1	*72.7*	*57.8*	79.2	*39.3*	*62.6*	*44.0*	63.4
S2Teacher (FCOS)	AAAI 2026	51.5	75.9	58.8	82.0	45.5	69.3	51.4	72.5	35.9	59.3	38.6	57.2
Ours (RetinaNet)	-	54.5	74.3	61.9	**93.0**	*50.0*	71.2	57.0	**86.2**	38.5	59.6	42.8	*65.7*
Ours (YOLO)	-	**57.4**	**79.4**	**65.6**	*90.5*	**54.0**	**75.6**	**61.7**	*85.2*	**42.5**	**64.0**	**48.7**	**67.0**

Table 4: Fully supervised oracle performance on UODD and DUO datasets. The best and second-best results are shown in bold and underline, respectively.

Method	UODD (oracle)			DUO (oracle)		
	mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
RetinaNet	48.9	86.4	50.5	58.6	79.2	67.0
FCOS	47.5	84.8	46.0	*62.8*	80.8	69.5
Faster R-CNN	48.4	85.7	48.6	62.4	**83.8**	*70.8*
YOLOv8	**52.7**	**89.8**	**56.0**	**63.4**	*83.7*	**71.2**
YOLOX	*51.9*	*88.9*	*54.5*	62.0	82.5	70.1



Figure 5: Qualitative visualization of our trained model applied to out-of-domain ImageNet samples. The green, red, and purple boxes represent the detection results for the categories *Echinoidea*, *Holothurian*, and *Starfish*, respectively.

mAP gains can be dominated by detector choice; we report Recovered Performance Ratio

$$\mathrm{RPR}(\alpha) = \frac{\mathrm{mAP}_{\mathrm{sparse}}(\alpha)}{\mathrm{mAP}_{\mathrm{full}}} \times 100\%, \tag{12}$$

where α is the annotation sparsity level. As shown in Fig. 4, our method reaches 97.3% RPR at 30% sparsity, meaning it recovers 97.3% of the fully supervised performance, and retains 67% RPR even at 90% sparsity, significantly outperforming all compared methods.

4.4 Qualitative Experiments

As shown in Fig. 3, on UODD with 70% annotations removed and using a confidence threshold of 0.7, comparison methods often confuse positive and negative samples, producing missed and false detections. We additionally visualize detections with confidence scores between 0.3 and 0.7 as dashed-line boxes; this reveals that baselines assign low confidence even to simple real instances. Thanks to the Risky Sample Filtering (RSF) and Conflict Mitigator (CM) modules, our method achieves more complete detection, with the average confidence of detected ground-truth instances approximately 40% higher than that of other methods. Figure 5 further shows that the BiMask and CM modules allow the trained model to produce reasonable qualitative predictions on out-of-domain ImageNet samples, suggesting a certain degree of cross-domain generalization.

Table 5: Per-batch K-S test results across training stages on UODD (50% annotations removed).

Stage	Early	Middle	Late	Overall
Passed (%)	79.0	83.1	82.8	81.7
Avg. D	0.142	0.135	0.135	0.137
Avg. p	0.429	0.459	0.455	0.448

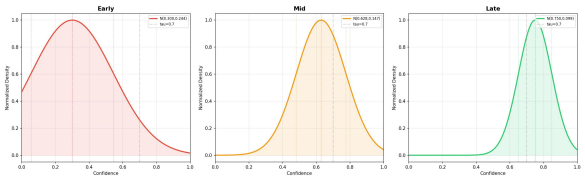

Figure 6: Gaussian fitting results of samples from random batch at the early, middle, and late training stages on UODD with 50% annotations removed.

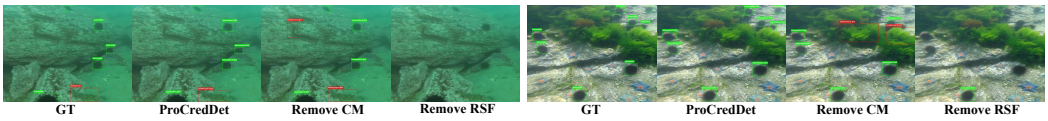

Figure 7: Ablation visualization results. The green, red, and blue bounding boxes in the figure represent the detected categories of *Echinoidea*, *Holothurian*, and *Scallop*, respectively. False detections occur when the CM module is removed from the model, and a large number of missed detections appear after removing the RSF module.

4.5 Ablation Experiments

As shown in Table 7, we conduct module ablations on UODD with 70% annotations removed. Removing CM causes a 1.9% mAP drop (erroneous teacher predictions propagate), removing RSF causes a 5.8% drop (unlabeled positives hidden in the negative set are no longer filtered), and removing both drops performance by 7.6%. As illustrated in Fig. 7, CM removal leads to false detections while RSF removal leads to many missed detections. The teacher-student baseline improves over RetinaNet by 1.6%, and adding BiMask further improves by 2.6%, confirming that this gain comes solely from complementary masking.

The bottom part of Table 7 compares CM strategies. Directly shielding erroneous categories improves mAP by 1.9%, whereas soft distillation with temperatures 0.5, 1.0, and 2.0 changes performance by 0%, -0.6%, and -2.7%, respectively. Thus, hard masking should be applied once an erroneous category is detected, while soft distillation preserves misleading information and fails to address the problem.

To validate the RSF boundary, we examine the conditional purity of the three confidence intervals across training stages. Table 6 shows that pseudo-label samples ($> \tau$) remain over 90% true positives, the risky interval $[\mu - \sigma, \tau]$ is mixed early (30.5% positives) but clears later (4.5%), and the negative interval $[0, \mu - \sigma]$ contains fewer than 0.2% positives. It further shows risky samples shifting to pseudo-label and negative intervals as training converges, justifying the three-way split: $[0, \mu - \sigma]$ is safe for negatives, $[\mu - \sigma, \tau]$ is too risky, and $(\tau, 1]$ is reliable for pseudo-labels. The K-S test in Table 5 and Gaussian fits in Fig. 6 confirm the global Gaussian approximation is reasonable, but the empirical contamination rates are the primary justification.

4.6 Hyperparameter Experiments

Figure 8 summarizes the hyperparameter sensitivity on UODD with 70% annotations removed. The pseudo-label threshold τ peaks at 0.7. A lower threshold ($\tau \leq 0.6$) introduces many low-quality pseudo-labels and produces erroneous training signals, whereas a higher threshold ($\tau \geq 0.8$) filters out valid pseudo-labels and leads to insufficient supervision. The Bernoulli keep ratio p is best at

Table 6: True positives (left) and sample distribution ratios (right) in each confidence interval at early, middle, and late training stages on UODD with 70% annotations removed.

	True positives (%)			**Sample distribution (%)**		
Stage	$> \tau$	$[\mu-\sigma, \tau]$	$< \mu-\sigma$	$> \tau$	$[\mu-\sigma, \tau]$	$< \mu-\sigma$
early	94.3	30.5	0.2	23.7	56.7	19.6
middle	92.7	24.7	0.1	39.2	40.5	20.3
late	93.2	4.5	0.1	45.2	31.4	23.4

Table 7: Top: ablation experiment results on the UODD dataset with 70% annotations removed; TS denotes the teacher-student model. Bottom: correction strategies and distillation temperatures.

Model	M1	M2	M3	M4	M5	M6
TS	-	-	√	√	√	√
BiMask	-	-	-	√	√	√
CM	-	-	-	√	-	√
RSF	-	-	-	-	√	√
mAP	31.4	33.0	35.6	37.4	41.3	**43.2**

Strategies	No CM	CM (shield)	T=0.5	T=1.0	T=2.0
Results (%)	41.3	**43.2**	41.3	40.7	38.6

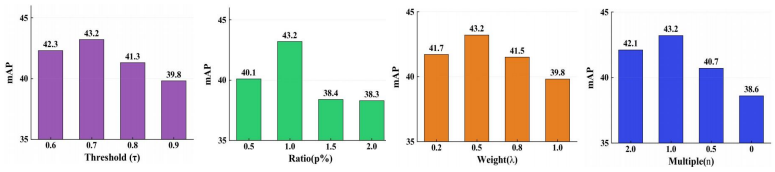

Figure 8: Hyperparameter sensitivity of ProCreDet on UODD with 70% annotations removed (RetinaNet), evaluated by mAP. (a) Pseudo-label threshold τ. (b) Bernoulli keep ratio $p\%$ in RSF. (c) Consistency-loss weight λ for the averaged two-branch consistency losses. (d) RSF boundary coefficient n in $\mu - n\sigma$. Defaults: $\tau = 0.7$, $p = 0.01$, $\lambda = 0.5$, $n = 1$ (i.e., $\mu - \sigma$).

0.01, indicating that most low-quality negatives contribute little. The consistency weight λ peaks at 0.5; smaller values provide insufficient regularization, while larger values overwhelm the supervised detection loss. The RSF boundary coefficient n in $\mu - n\sigma$ is best at 1: a stricter boundary excludes too many potential positives, and a looser one retains risky samples.

5 Conclusion

This paper proposes ProCreDet, a progressive credible sample selection network for sparsely annotated underwater object detection. To address defects of existing sparsely annotated detection methods and tackle limited annotation resources in underwater vision tasks, ProCreDet designs tailored modules. To handle severe color variations in underwater images caused by light attenuation and scattering, it employs a BiMask strategy to extract complementary features and enhance positive target mining. The Risky Sample Filtering (RSF) module places ambiguous unlabeled samples in a risky buffer and gradually accepts credible positives, suppressing false negatives and stabilizing model training. The Conflict Mitigator (CM) module avoids misleading supervision from erroneous teacher predictions while preserving valuable inter-class correlation information. Experiments on UODD and DUO datasets verify that ProCreDet surpasses state-of-the-art methods, delivering significant accuracy improvements under high annotation sparsity and stable training performance. Future work will explore SAOD under ultra-high annotation sparsity to advance deployment of sparse supervised detection algorithms in marine environments.

REFERENCES

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Gangqi Chen, Zhaoyong Mao, Junge Shen, and Zhiyong Cheng. Pseudo label guided object detection in sparsely annotated underwater optical images. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Liang Chen, Yuyi Yang, Zhenheng Wang, Jian Zhang, Shaowu Zhou, and Lianghong Wu. Underwater target detection lightweight algorithm based on multi-scale feature fusion. *Journal of Marine Science and Engineering*, 11(2):320, 2023.

Chenping Fu, Risheng Liu, Xin Fan, Puyang Chen, Hao Fu, Wanqi Yuan, Ming Zhu, and Zhongxuan Luo. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing*, 517:243–256, 2023.

Huilin Ge, Yuewei Dai, Zhiyu Zhu, and Xu Zang. Single-stage underwater target detection based on feature anchor frame double optimization network. *Sensors*, 22(20):7875, 2022.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO series in 2021, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Muwei Jian, Nan Yang, Chen Tao, Huixiang Zhi, and Hanjiang Luo. Underwater object detection and datasets: a survey. *Intelligent Marine Technology and Systems*, 2(1):9, 2024. doi: 10.1007/s44295-024-00023-6.

Lihao Jiang, Yi Wang, Qi Jia, Shengwei Xu, Yu Liu, Xin Fan, Haojie Li, Risheng Liu, Xinwei Xue, and Ruili Wang. Underwater species detection using channel sharpening attention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4259–4267, 2021.

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. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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*, pp. 2980–2988, 2017.

Yu Lin, Jianghang Lin, Kai Ye, You Shen, Shengchuan Zhang, and Liujuan Cao. S^2teacher: Step-by-step teacher for sparsely annotated oriented object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6997–7005, 2026.

Chongwei Liu, Haojie Li, Shuchang Wang, Ming Zhu, Dong Wang, Xin Fan, and Zhihui Wang. A dataset and benchmark of underwater object detection for robot picking. In *2021 IEEE international conference on multimedia & expo workshops (ICMEW)*, pp. 1–6. IEEE, 2021.

Yusuke Niitani, Takuya Akiba, Tommi Kerola, Toru Ogawa, Shotaro Sano, and Shuji Suzuki. Sampling techniques for large-scale object detection from sparsely annotated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6510–6518, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Saksham Suri, Saketh Rambhatla, Rama Chellappa, and Abhinav Shrivastava. Sparsedet: Improving sparsely annotated object detection with pseudo-positive mining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6770–6781, 2023.

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*, pp. 9627–9636, 2019.

Tiancai Wang, Tong Yang, Jiale Cao, and Xiangyu Zhang. Co-mining: Self-supervised learning for sparsely annotated object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2800–2808, 2021.

Junjie Wen, Guidong Yang, Benyun Zhao, Lei Lei, Zhi Gao, Xi Chen, and Ben M Chen. Joint image enhancement for underwater object detection in various domains. *IEEE Journal of Oceanic Engineering*, 2025.

Lianjun Wu, Jiangxiao Han, Zengqiang Zheng, and Xinggang Wang. Co-student: Collaborating strong and weak students for sparsely annotated object detection. In *European Conference on Computer Vision*, pp. 459–475. Springer, 2024.

Shubo Xu, Minghua Zhang, Wei Song, Haibin Mei, Qi He, and Antonio Liotta. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing*, 527:204–232, 2023.

Xiaoli Yu, Yanyun Qu, and Ming Hong. Underwater-GAN: Underwater image restoration via conditional generative adversarial network. In *International Conference on Pattern Recognition*, pp. 66–75. Springer, 2018.

Dehua Zhang, Changcheng Yu, Zhen Li, Chunbin Qin, and Ruixue Xia. A lightweight network enhanced by attention-guided cross-scale interaction for underwater object detection. *Applied Soft Computing*, pp. 113811, 2025.

Han Zhang, Fangyi Chen, Zhiqiang Shen, Qiqi Hao, Chenchen Zhu, and Marios Savvides. Solving missing-annotation object detection with background recalibration loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1888–1892. IEEE, 2020.

Appendix

A Implementation Details

A.1 Training Settings

The learning rate is initialized to 1×10^{-5}. No warm-up or weight decay is used. Gradient clipping with a threshold of 0.1 is applied.

A.2 Input Pre-processing

Input images are resized so that the shorter side is 608 pixels and the longer side does not exceed 1024 pixels, with padding to multiples of 32.

A.3 Data Augmentation

Teacher branch: random horizontal flipping (50%) and random cropping with a scale range of 0.8–1.0.

Each student branch is independently augmented with strong color augmentations: brightness, contrast, and saturation are jittered within $[0.3, 1.7]$; hue is shifted within $[-0.3, 0.3]$; and grayscale conversion is applied with a probability of 20%.

A.4 Key Hyperparameters

Pseudo-label threshold: $\tau = 0.7$ (fixed throughout training).

Consistency-loss weight: $\lambda = 0.5$; the consistency losses of the two strong branches are averaged before weighting.

A.5 Inference and Backbone

At inference, only the teacher weights are used, which is equivalent to a single standard detector without any extra branches or overhead.

For other compared models, we maintain their original experimental settings. For models using ResNet as the backbone, we consistently employ the 101-layer network.

B Training Stage Definitions

For the analyses reported below, we divide the 30-epoch training into three stages based on validation performance: the early stage (performance rises sharply), the middle stage (performance rises steadily), and the late stage (performance stabilizes).

C K-S Statistic Evolution over Training

Figure 9 plots the K-S statistic D and the corresponding p-values versus training iterations on UODD with 50% annotations removed. Raw batch-level values are rendered as low-opacity scatter to avoid overplotting across approximately 39,000 points, while the moving average (window = 1000) and ± 1 standard-deviation band highlight the trend. The p-value subplot is shown on a logarithmic scale (10^{-12} to 1.5) to reveal differences across orders of magnitude. The K-S statistic gradually decreases over training, indicating that the labeled-sample confidence scores remain close to a Gaussian distribution throughout training. On a per-batch basis, approximately 79.0%, 83.1%, and 82.8% of batches pass the K-S test in the early, middle, and late stages, respectively.

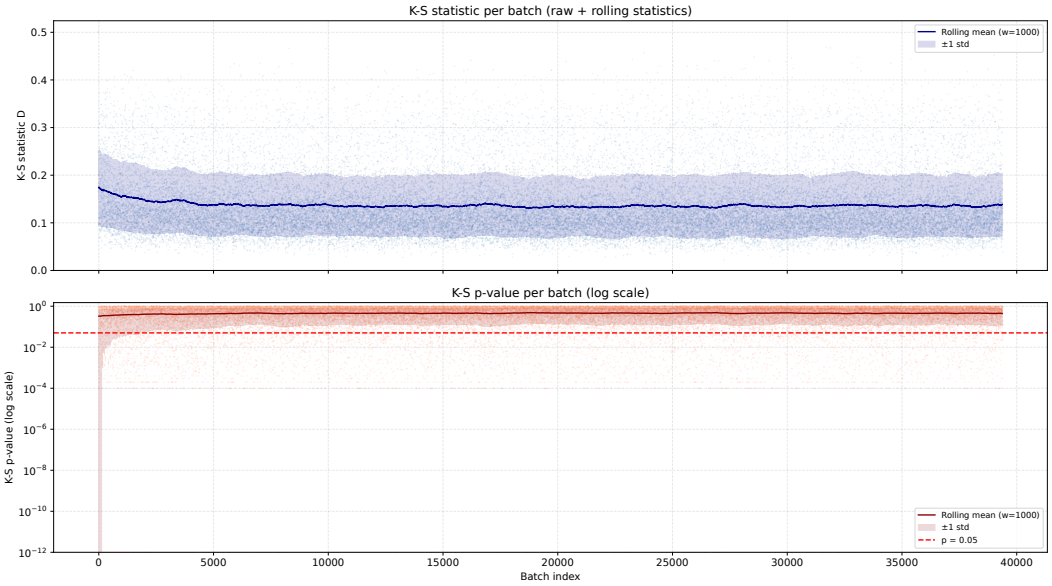

Figure 9: Evolution of the K-S statistic and p-values over training iterations on UODD with 50% annotations removed. Raw batch-level values are shown as low-opacity scatter points; the solid line is the moving average with window 1000, and the shaded band is the moving ± 1 standard deviation. The p-value subplot uses a logarithmic scale from 10^{-12} to 1.5.

D Gaussian Fitting Analysis across Training Stages

We randomly pick one epoch from the early, middle, and late stages, aggregate all labeled samples in that epoch, and visually fit a Gaussian to their confidence scores (see the Gaussian fitting results in the main paper). The fitted mean shifts to the right as training progresses, verifying that the Gaussian can serve as a criterion for evaluating pseudo-label correctness at different stages. Meanwhile, the bell-shaped curve narrows progressively: early on the model cannot reliably distinguish positive and negative samples and produces inconsistent predictions, whereas in the middle and late stages the model converges and makes more consistent judgments.

E Interval Purity and Distribution Shift

Tables in the main paper report the proportion of genuine positives in each confidence interval and the distribution of unlabeled samples across intervals on UODD with 70% annotations removed. For pseudo-labels above τ, the true-positive ratio remains above 90% throughout training. In the risky interval $[\mu - \sigma, \tau]$, positives account for 30.5% of unlabeled samples in the early stage, dropping to 24.7% in the middle stage and 4.5% in the late stage; the total number of samples in this interval also decreases toward zero. In the negative interval $[0, \mu - \sigma]$, positives remain around 0.1% across all stages. Thus, sampling a false negative from $[0, \mu - \sigma]$ is roughly $1/200$–$1/300$ as likely as sampling a false positive from $[\mu - \sigma, \tau]$. As training converges, risky-interval samples progressively shift to the pseudo-label and negative intervals; the residual samples entering the negative interval correspond to hard positives that remain unidentifiable by the model.

F Training Stability Curves

Figure 10 shows the training/evaluation curves of representative SAOD methods on UODD with 70% annotations removed. After the training peak, existing methods drop by 5.0–14.2 mAP, whereas ProCreDet drops by only 0.6 mAP, demonstrating that the risky-buffer design stabilizes training.

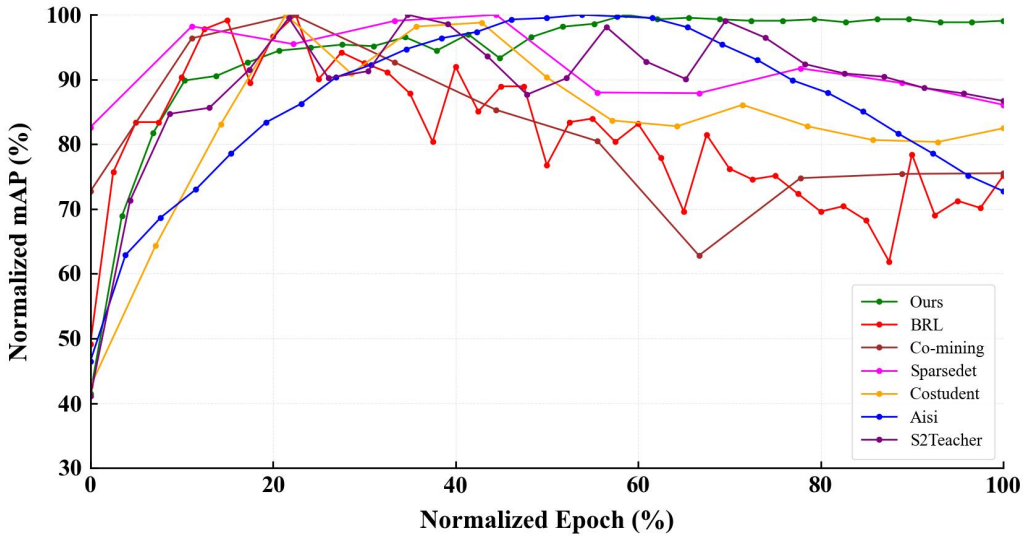

Figure 10: Training/evaluation curves of representative SAOD methods on UODD with 70% annotations removed. After the training peak, existing methods drop by 5.0–14.2 mAP, whereas ProCreDet drops by only 0.6 mAP.

Table 8: Three independent runs (seeds) of ProCreDet on UODD with 70% annotations removed, showing AP and the maximum fluctuation after the best training validation point.

Run	AP	Drop after Peak (↓)
Seed 1	42.7	0.4
Seed 2	43.2	0.6
Seed 3	43.3	0.8
Mean	43.1	0.6

G Multi-Seed Stability and RSF Ablation Curves

G.1 Multi-Seed Stability on UODD

Table 8 reports three independent runs of ProCreDet on UODD with 70% annotations removed. The AP values are 42.7, 43.2, and 43.3, with maximum post-peak fluctuations of only 0.4, 0.6, and 0.8, respectively, confirming that the method is stable across random seeds.

G.2 Training Curves with and without RSF

Figure 11 compares the training/validation curves with and without the Risky Sample Filtering (RSF) module. The full model fluctuates by only 0.4 after the best validation point in this seed, whereas removing RSF leads to a maximum fluctuation of 9.5. Note that this fluctuation metric is distinct from the final mAP drop reported in the main paper's module ablation; across seeds the full model's average post-peak drop is 0.6 (Table 8), matching the main paper. These results demonstrate that RSF is essential for stable training.

H Hyperparameter Sensitivity Details

The sensitivity curves shown in the main paper are obtained by varying one hyperparameter at a time on UODD with 70% annotations removed. The pseudo-label threshold τ peaks at 0.7: lower values

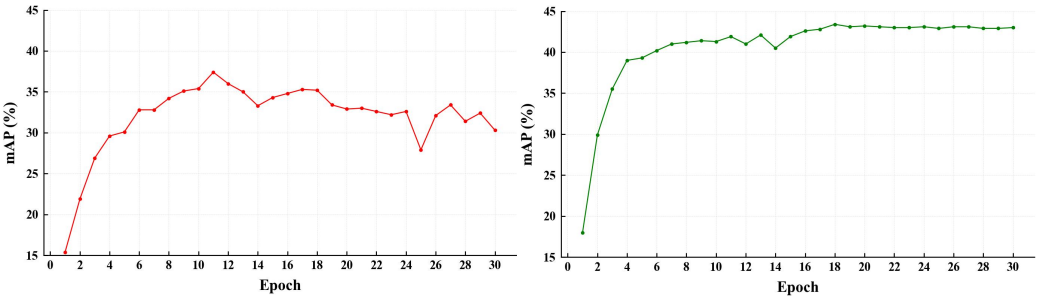

Figure 11: Training/validation curves with and without the RSF module on UODD. Removing RSF increases the maximum post-peak fluctuation from 0.4 (full model) to 9.5.

introduce noisy pseudo-labels, while higher values discard valid ones. The Bernoulli keep ratio p is best at 0.01, confirming that most low-quality negatives contribute little. The consistency weight λ peaks at 0.5, balancing consistency regularization and supervised losses. The RSF boundary coefficient n in $\mu - n\sigma$ is best at 1 (i.e., $\mu - \sigma$), striking the best trade-off between excluding too many potential positives and retaining risky samples.

I Formulas and Derivations

I.1 RSF Gaussian Parameter Estimation

At every training iteration, the mean and variance of the labeled-sample confidence scores in the current batch are estimated as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} s_i, \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(s_i - \mu)^2, \tag{13}$$

where N is the number of labeled instances in the batch. These statistics are re-estimated per batch without temporal accumulation.

I.2 Confidence Interval Definitions

Given μ, σ, and the pseudo-label threshold τ, the unlabeled samples are partitioned into three sets:

$$\begin{aligned} \mathcal{P} &= \{s_j : s_j > \tau\}, \\ \mathcal{R} &= \{s_j : \mu - \sigma \le s_j \le \tau\}, \\ \mathcal{N} &= \{s_j : s_j < \mu - \sigma\}, \end{aligned} \tag{14}$$

corresponding to pseudo-labels, the risky buffer, and negatives, respectively. Samples in $\mathcal{R}$ are discarded; samples in $\mathcal{N}$ are randomly retained with probability p.

I.3 BiMask Generation

The complementary mask is generated as

$$M_{Bi} \sim \text{Binomial}(n, 0.5), \tag{15}$$

where n is the number of elements in the feature map. The two student-branch feature maps are then masked as

$$F_{Bi1} = F_{s1} \odot M_{Bi}, \qquad F_{Bi2} = F_{s2} \odot (1 - M_{Bi}), \tag{16}$$

ensuring that F_{Bi1} and F_{Bi2} are mutually exclusive.

I.4 EMA UPDATE

The teacher network parameters θ_t are updated from the student parameters θ_s by

$$\theta_t \leftarrow \alpha\,\theta_t + (1-\alpha)\,\theta_s, \qquad \alpha = 0.99, \tag{17}$$

after every training iteration.

I.5 STABILITY DROP METRIC

The performance drop after the training peak is defined as

$$\text{Drop} = \max_{t > t_{\text{peak}}} \left(\text{mAP}_{\text{peak}} - \text{mAP}_t\right), \tag{18}$$

where t_{peak} is the epoch at which the peak mAP is attained.

I.6 K-S STATISTIC

The Kolmogorov–Smirnov statistic between the empirical confidence distribution F_{emp} and the fitted Gaussian F_N is

$$D = \sup_x \left|F_{\text{emp}}(x) - F_N(x)\right|; \tag{19}$$

the p-value is computed from the asymptotic Kolmogorov distribution.

I.7 INTERVAL PURITY AND CONTAMINATION

For an interval I, purity and contamination are defined as

$$\text{Purity}(I) = \frac{TP_I}{|I|}, \qquad \text{Contamination}(I) = \frac{FN_I}{|I|}, \tag{20}$$

where TP_I is the number of true positives and FN_I is the number of false negatives (unlabeled positives) assigned to I.

I.8 TOTAL LOSS AND GRADIENT FLOW

The total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \lambda \cdot \frac{\mathcal{L}oss_{\text{c1}} + \mathcal{L}oss_{\text{c2}}}{2}. \tag{21}$$

Gradients of $\mathcal{L}oss_{\text{c1}}$ and $\mathcal{L}oss_{\text{c2}}$ flow only to the student network; the teacher is updated exclusively by EMA and does not receive gradients.

ETHICS STATEMENT

(This section is recommended and does not count toward the page limit.)

This work studies sparsely annotated underwater object detection. All experiments are conducted on two publicly available benchmark datasets, UODD and DUO, which are used in accordance with their original licenses and intended research purposes. The study does not involve human subjects, personally identifiable information, or the collection of new data. The proposed method aims to reduce the annotation cost of underwater perception and is intended to support marine resource exploration and ecological monitoring. We are not aware of any potential harmful insights, discriminatory or fairness concerns, privacy or security issues, or conflicts of interest arising from this work.

REPRODUCIBILITY STATEMENT

(This section is recommended and does not count toward the page limit.)

We have made extensive efforts to ensure the reproducibility of our results. The two datasets used in this work (UODD and DUO) are publicly available. Complete implementation details, including training settings, input pre-processing, data augmentation, key hyperparameters, and inference protocol, are provided in Appendix A; training stage definitions and full metric derivations are given in Appendices B and I. Sensitivity analyses for all key hyperparameters are reported in Section 4.6 of the main paper, and multi-seed stability results are reported in Appendix G. Upon acceptance, we will release the source code to facilitate reproduction.

AI USE STATEMENT

(This section is required and does not count toward the page limit.)

In this work, we used generative AI tools solely for language polishing and grammar correction of the manuscript. We have not used generative AI tools for research ideation, method design, experimental execution, data analysis, or the generation of figures, tables, and results, and these tasks are not applicable to the AI assistance used in this work. We have carefully reviewed all AI-polished text for accuracy and fidelity to the intended technical content. We take full responsibility for the final content of this work, including text produced with the aid of generative AI.