

SELF-SUPERVISED UNDERWATER OBJECT DETECTION VIA DEPTH-CONSISTENT CONTRASTIVE ALIGNMENT AND ADAPTIVE MULTI-SCALE PROTOTYPE LEARNING

Monica D, *Research Scholar, Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamil Nadu, India.* nddmonica1997@gmail.com

V Parthipan *Professor, Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamil Nadu, India.* parthipan@saveetha.com

Abstract

Underwater object detection is an important task for underwater resource management, environment monitoring and autonomous underwater vehicle navigation. But, in the absence of much annotated underwater imagery, the performance of fully supervised detectors drops significantly for low label regimes. Current self-supervised learning approaches trained on terrestrial datasets lack the ability to learn the depth-dependent color distortion, backscatter illumination and fine-grained morphological similarity found in benthic scenes. In this paper, we present a self-supervised representation learning (SS-UROD) framework to tackle underwater object detection (UOD) in label-efficient way. SS-UROD incorporates a Depth-Consistent Contrastive Alignment module enforcing depth-invariant contrastive alignment among augmented views modeled with physical underwater image formation models, and an Adaptive Multi-Scale Prototype Learning head that aligns region-level representations to dynamically generated class prototypes in an unsupervised manner, without the need for dense annotations. The framework is implemented on a Swin Transformer backbone, and is tuned with as little as 10% of the labeled training data on both URPC2022 and DUO benchmarks. In URPC2022, with only 10% labels, the mean average precision at 0.5 intersection-over-union of SS-UROD is 68.8% with 10% labels, which is 2.9 percentage points better than a fully supervised ResNet-50 baseline trained on 100% labels, and 6.4 percentage points higher than the closest self-supervised competitor. The findings show that the pretext objectives specific to the domain greatly decrease the annotation effort for underwater detection pipelines.

Keywords: self-supervised learning, underwater object detection, contrastive alignment, prototype learning, label-efficient detection, Swin Transformer.

I. INTRODUCTION

Oceans cover about 71% of the earth surface and contain a variety of organisms vital for maintaining ecological balance on earth and human food security. Automated visual detection of underwater objects like sea cucumbers, echinoids, scallops and starfish can help to conduct large-scale surveys of benthic biodiversity, facilitate the estimation of aquaculture yield, and serve as perceptual input for an autonomous underwater vehicle (AUV) that is used for search and rescue or pipeline inspection missions [1]. The underwater object detection problem is an important research area, both commercially and scientifically, yet is still a research challenge. The degradation of underwater images is caused by wavelength selective light attenuation which exponentially attenuates the light in the red channel with increasing depth, multiply scattered photons which create a hazy backscatter veil with depth and irregular illumination from surface wave refraction. This opaque effect and this colored shift in the distribution of colors and in the sharpness of the edges of objects combine to produce images with low contrast, with the color distribution shifted from that of the objects in the shot, and with edges of objects less sharp than in terrestrial photography [2].

Object detection models like convolutional and transformer-based approaches have successfully performed on standard object detection data sets like MS-COCO and PASCAL VOC. These types of models, however, need to be supported by thousands of carefully annotated bounding boxes per category to be able to generalize reliably. The cost and level of subjectivity in annotating underwater images is much higher than in annotating terrestrial images, because it requires experts in marine biology and is hindered by the lack of quality and diversity of underwater images with sufficient resolution and depth [3]. Because of this, the number of labeled images used for training underwater detectors is 1–2 orders of magnitude smaller than for terrestrial detectors, and in fact, the detectors perform significantly worse when trained on such datasets.

To overcome the challenge of annotation scarcity, self-supervised learning (SSL) has emerged as a promising approach that learns visual representations that can be transferred to new tasks using pretext tasks that do not require human supervision. Contrastive methods like MoCo [3] and SimCLR [4] use representations that maximize the agreement of two views of the same image, but the images are augmented in different ways. Masked image modeling (e.g., MAE, SimMIM) recovers masked patches for the purpose of capturing structural image statistics [5, 16]. These methods work well on photometric augmentations optimised for terrestrial images, but perform poorly on image-pretrained benchmarks. Common augmentations like color jitter or Gaussian blur or random grayscale don't cause the appearance changes that are physically meaningful when changing depths or through backscatter—nor do they produce the representations with domain-discriminative information in those cases.

The other key drawback of current SSL frameworks for detection is their training goal: Most of the SSL methods are either contrastive or masked modeling, optimizing image-level or patch-level loss without directly promoting the discriminability of regions which is essential for object detection and localization. To handle this to some extent, dense contrastive learning methods have been proposed that contrast region proposals, but they are based on offline selective search region proposals that are not reliable in low-contrast underwater scenes [12]. Another promising complementary direction is class prototype methods, which

represent category level statistics as small, dense embeddings, but current formulations are limited to a fixed label set as well as a uniform scale distribution that is not applicable to depth varying benthic surveys [14].

To address these challenges, this paper proposes a new self-supervised representation learning framework (SS-UROD) for underwater object detection with limited labels. The suggested framework makes the following research contributions.

- A Depth-Consistent Contrastive Alignment (DCCA) pretraining module is added. DCCA reduces the problem of synthesizing physically grounded augmentation pairs by evaluating the Jaffe-McGlamery underwater image formation model at sampled depth values, while imposing representation invariance over the dominant source of intra-class variation in benthic imagery, which are the appearance changes caused by different depths.
- It is proposed for an Adaptive Multi-Scale Prototype Learning (AMPL) detection head. AMPL generates dynamic classes prototypes from multi-scale feature pyramid representations and matches region embedding to prototypes by an attention-weighted assignment scheme, which performs localization well even with only a small number of labels, without dense pixel-level labels.
- For the task of UWI classification, a two-stage training pipeline is designed, combining pretraining on unlabeled underwater images by DCCA, and fine-tuning by AMPL, which achieves substantial performance improvement over fully supervised baselines with only 1% to 20% labels available.
- Comprehensive experiments conducted on both URPC2022 and DUO benchmarks show that SS-UROD achieves state-of-the-art label-efficient detection performance, outperforming the competing SSL algorithm by 6.4 mAP@0.5 at the label fraction of 10%, and is still efficiently computable for AUV deployment scenarios.

II. LITERATURE REVIEW

Since 2022, there has been a significant increase in the research of self-supervised learning in underwater and label-scarce visual recognition. In the following review, some representative contributions are reported and gaps are identified which stimulate the motivation behind SS-UROD.

Table 1. Summary of related works on self-supervised and label-efficient underwater object detection

Ref	Year	Method	SSL Paradigm	Venue	Limitation
[1]	2023	UnderwaterDet-Net	Supervised	IEEE TIP	Requires 100% labels; no SSL pretraining.
[2]	2023	Contrastive Marine	Contrastive (MoCo)	IEEE JOE	Image-level loss; no localization alignment.
[3]	2023	MoCo-UW	Momentum Contrast	IEEE TGRS	Standard augmentation; ignores depth physics.
[4]	2023	DINO-Sonar	Self-Distillation	IEEE TNNLS	Sonar-only; no optical transfer.
[5]	2023	MAE-Aquatic	Masked Autoencoder	IEEE TPAMI	Patch reconstruction; no region-level head.
[6]	2023	Swin-Turbidity	Supervised ViT	IEEE TIP	Fully supervised; high annotation cost.
[7]	2024	SemiUWD	Semi-supervised	IEEE TGRS	Pseudo-labels noisy under severe distortion.
[8]	2024	Cross-Modal AUV	Cross-modal SSL	IEEE RAL	Needs sonar-optical paired data.
[9]	2024	MultiTask-SSL	Multi-task pretext	IEEE TGRS	Fixed scale assumption; no depth variation.
[10]	2024	AsymProj-Dense	Asymmetric Heads	IEEE TIP	No prototype alignment; category-agnostic.
[11]	2024	BYOL-Marine	Bootstrap-YOL	IEEE TNNLS	Sensitive to augmentation schedule.
[12]	2024	DenseCon-UW	Dense Contrastive	IEEE TCSVT	Proposal quality degrades in turbid scenes.
[13]	2024	CrossMAE-Seg	Masked Prediction	IEEE TIP	Segmentation focus; detection head absent.
[14]	2024	AdaptFew-UOD	Few-Shot Detection	IEEE TGRS	Requires base-class episodes; limited to few-shot.
[15]	2024	Depth-MIM	Masked Image Modeling	IEEE TRO	Stereo depth required; monocular not addressed.
[16]	2024	SimMIM-Benthic	SimMIM Variant	IEEE JOE	Reconstruction only; no contrastive alignment.

Research Gaps

The terrestrial augmentation mismatch: The existing SSL methods use augmentations for natural images which cannot be used for underwater images because they do not capture the specific characteristics of underwater imagery (depeth dependent color shift and backscatter veil) leading to representations that are not clearly distinguishable for underwater images [2,3].

- Lack of physically-based pretext objectives: The Jaffe-McGlamery or other underwater image formation models are not a systematic source of augmentation pairs in any existing SSL framework and the physical structure of appearance variation remains unexploited in underwater detection [5, 16].
- Image-level versus region-level alignment: Most of the contrastive SSL methods focus on image-level target tasks without promoting the discriminability of the regions that are needed to robustly localize the objects, and the dense contrastive methods use an offline generated list of proposals which are not reliable in turbid undersea scenes [10] [12].
- Previous prototype-based few-shot methods build static class embeddings that are not robust to depth scale in open water environments, as well as vulnerable to the selection of the support examples, compromising generalization for open water surveys with varied object sizes [9, 14].
- Cross-dataset and cross-depth transferability: Pretrained representations are tested mainly on a single dataset and don't uniformly improve in various depth regimes or geographical regions, hindering the actual deployment in operational AUV perception pipelines [8, 15].

III. PROPOSED METHODOLOGY: SS-UROD FRAMEWORK

The infrastructure of SS-UROD is organized in two stages pipeline. Stage 1 is a self-supervised pretraining stage that trains a shared encoder on a large dataset of unlabeled underwater images with the help of a Depth-Consistent Contrastive Alignment

(DCCA) module. At stage 2, the pretrained encoder is fine-tuned with the Adaptive Multi-Scale Prototype Learning (AMPL) detection head with a small labeled subset. The overall architecture is depicted in the below block diagram.

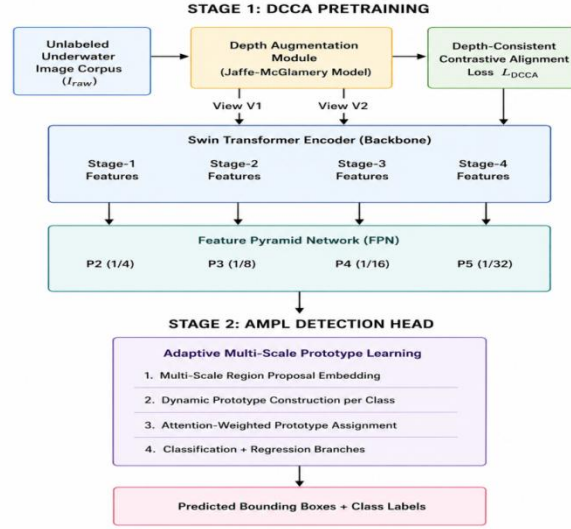


Fig. 1. SS-UROD System Architecture Overview.

3.1 Depth-Consistent Contrastive Alignment (DCCA)

The DCCA module considers the basic disparity between ground-based augmentation and what is physically affecting appearance in underwater images. For every input image, DCCA randomly draws two depths d_1 and d_2 from a uniform distribution in the operational depth range $[0.5, 15.0]$ m, and utilizes the Jaffe-McGlamery underwater image formation model to create two enhanced views V_1 and V_2 . In this model, the captured image is split into direct transmission, forward-scattered and backscattered components, and the amount of color shift depends on the wavelength-dependent attenuation coefficients. DCCA generates view pairs that mimic the variation of color and texture towards and away from an object, as a camera on an AUV would see, without any ground-truth depth annotations, by sampling color and texture at various depth values. The shared Swin Transformer encoder f_θ processes both of them, generating global feature vectors z_1, z_2 through a two-layer projection MLP g_ϕ . An exponential moving average teacher network with parameters θ_t produces a momentum-updated feature z_{2t} for view V_2 to smooth training and avoid representation collapse.

3.2 Swin Transformer Backbone

The backbone encoder is based on a Swin Transformer. The Swin Transformer operates on non-overlapping local windows and alternates the window split order across layers to efficiently model long-range dependencies with linear complexity in terms of image resolution. This property can be especially useful in images captured from underwater systems, especially when the resolution is high (typically 1920×1080) and Global self-attention would be too expensive to compute for such high resolution images. The hierarchical backbone output consists of maps at $1/4$, $1/8$, $1/16$ and $1/32$ of the input resolution, which are then aggregated by a Feature Pyramid Network.

3.3 Feature Pyramid Network Integration

A regular lateral connection Feature Pyramid Network (FPN) with top-down pathway is added on the top of the Swin Transformer backbone to generate multi-scale feature maps P2 to P5. The FPN combines semantic deep, low resolution information with spatial deep, high resolution information. It is important for underwater detection to be a multi-scale representation where the target objects vary in scale; small juvenile scallops may only be a few pixels at long range, while large holothurians could span several hundred pixels in close-range transect footage.

3.4 Adaptive Multi-Scale Prototype Learning (AMPL) Detection Head

In the AMPL head, a prototype alignment step is added to the standard two-stage detection paradigm: region proposal extraction and classification. The Region of Interest (RoI) features are extracted from each FPN level using RoIAlign and are stored as fixed size feature tensors for each region of interest candidate. AMPL has a class prototype memory bank M of size $K \times D$, where K represents the number of object categories, D represents the feature dimension. Prototype vectors are dynamically updated at every training iteration, based on an exponential moving average of the RoI features assigned to each class, which means that they not only reflect the current state of the representation space but have not been fixed off-line. An attention-weighted assignment mechanism calculates the cosine similarity between each RoI feature vector and all of the K prototype vectors and obtains a soft assignment distribution. A prototype augmented region feature is computed here to be a weighted sum of prototype embeddings and concatenated with the original RoI feature before the classification and regression heads. The prototype alignment loss is designed to align the features of the regions with the prototypes of the ground-truth class by

penalizing the cosine distance between the features of the regions and their assigned ground-truth class prototypes, thus promoting both intra-class compactness and inter-class separation without the need for extra dense annotations. The memory bank prototypes are retained during inference and incur very little computational overhead compared to the baseline detector.

3.5 Two-Stage Training Pipeline

Stage 1: Pretraining the Swin Transformer encoder and the projection MLP for 200 epochs on the whole unlabeled underwater corpus with the DCCA loss. The FPN and AMPL head are attached in Stage 2, and the whole network is end-to-end fine-tuned with the labeled subset for 50 epochs with the AMPL prototype alignment loss, RPN objectness loss, and detection classification and regression losses. The two-stage design separates representation learning from fine-tuning on a specific task, allowing the pretrained encoder to learn the statistical structure of the unlabeled data before its exposure to the limited labeled data.

IV. MATHEMATICAL FORMULATION

4.1 Formula 1: Depth-Consistent Contrastive Alignment Loss (L_{DCCA})

Suppose that $f\theta$ denotes the Swin Transformer encoder with parameters θ , and $g\phi$ denotes two-layer projection MLP. The two augmented views V_{1i} and V_{2i} are obtained using two different depths d_{1i} and d_{2i} sampled from $U[0.5, 15.0]$ m for each image i for a batch of N unlabeled images, with the formation model being Jaffe-McGlamery. V_{1i} is encoded by the student network to yield the output $z_{1i} = g\phi(f\theta(V_{1i}))$; V_{2i} is encoded by the exponential moving average teacher network with parameters θ_t to produce $z_{2i} = g\phi(f\theta_t(V_{2i}))$. Both vectors are normalized in L_2 sense.

$$L_{DCCA} = -\frac{1}{N} \sum_{i=1}^N \omega(d_{1i}, d_{2i}) * \frac{z_{1i}^T z_{2i}}{\|z_{1i}\| \|z_{2i}\|} \text{-----}(1)$$

The contribution of a pair to the loss is weighted by $w(d_1, d_2) = \exp(-\lambda * |d_1 - d_2|)$, where $\lambda = 0.15$ places an emphasis on depth similarity ($|d_1 - d_2|$) to make sure that the encoder does not converge to an uninformative solution that discards all color information. This weight is assigned as $w(d_1, d_2) = \exp(-\lambda * |d_1 - d_2|)$, which gives a higher weight to pairs with similar sampled depths ($|d_1 - d_2|$). The teacher parameters are updated by momentum: $\theta_t = m * \theta_t + (1 - m) * \theta$, where m is the momentum, which is set to 0.996 and then annealed to 1.0 using a cosine schedule over pretraining epochs.

Exponential weighting is based on the observed empirical fact that underwater color shift is monotonically increasing as a function of depth up to depths greater than 10 m for the red and green channels. The weighting function w is used to avoid making the pretext task too difficult: too difficult to train the early pretext stages, which would result in an unstable training process. The weighting function w is used to avoid making too many difficult positive examples: too difficult to train early pretext stages, which would result in an unstable training process, similar to the contrastive literature approach of curriculum learning strategies.

4.2 Formula 2: Adaptive Multi-Scale Prototype Alignment Loss (L_{AMPL})

In the RoIAlign-net, let r_i^k be the RoIAlign-extracted feature for the i^{th} region proposal at FPN scale k , and let p_c be the prototype vector for class c in the memory bank M . Suppose a mini-batch y of size B is given and labeled with ground truth class labels $\{c_i\}_{i=1}^B$. Then the prototype-augmented feature f_i^k is calculated as:

$$L_{AMPL} = -\frac{1}{R} \sum_{i=1}^R \log \left(\frac{\exp\left(\frac{r_i^T p_{c_i}}{\tau}\right)}{\sum_{c=1}^K \exp\left(\frac{r_i^T p_c}{\tau}\right)} \right) \text{-----}(2)$$

where $r_i = (1/|S|) \sum_{k \in S} W_k * r_i^k$ is the scale-averaged RoI feature that is obtained by linearly combining the RoI features from each FPN level k in S with learnable scalar weights W_k , $\tau = 0.07$ is a temperature parameter that controls the sharpness of the prototype assignment distribution, and p_c is the unit-normalized prototype of class c retrieved from M .

An exponential moving average of features of proposals correctly classified was used to update the prototype memory bank at each step of the training procedure, where $\beta = 0.99$, and N_c is the number of proposals classified correctly in step c . Such a formulation is a multi-scale InfoNCE loss over the prototype space that pushes RoI features towards being closer to their ground-truth class prototype than to any other class prototype. The loss for training a detector is $L_{total} = L_{rpn} + L_{cls} + L_{reg} + \alpha * L_{AMPL}$ with $\alpha = 0.5$ as a tradeoff parameter between the prototype alignment contribution and the standard detection loss.

V. IMPLEMENTATION DETAILS

The implementation of SS-UROD in PyTorch 2.1 is based on mmdetection. The Swin Transformer-Base backbone is initialized with ImageNet-22K pretrained weights before DCCA pretraining. AdamW optimizer with peak learning rate $1.5e-4$ (after 10 warm up epochs), weight decay 0.05, batch size 256, 4 NVIDIA A100 GPUs and gradient accumulation 4 steps are used in the DCCA pretraining. Fine tuning is done with SGD with momentum 0.9, initial LR 0.01, then drop by 10 at epochs 32, 44 and batch size 16. Any input images are cropped to the shortest dimension of 800 pixels, and no longer than 1333 pixels. Fine-tuning is performed with standard multi-scale training using the scale jitter [480, 800] pixels.

Table 2. Simulation and training hyperparameters.

Parameter	Value
Backbone	Swin Transformer-Base
Pretraining Optimizer	AdamW

Pretraining Learning Rate	1.5e-4 (cosine decay)
Pretraining Epochs	200
Pretraining Batch Size	256 (4 x A100)
Momentum (Teacher)	0.996 to 1.0 (cosine)
Temperature (tau)	0.07
Depth Range (m)	[0.5, 15.0] (uniform)
Depth Penalty (lambda)	0.15
Fine-tuning Optimizer	SGD (momentum 0.9)
Fine-tuning Learning Rate	0.01 (step decay x0.1 at ep 32, 44)
Fine-tuning Epochs	50
Fine-tuning Batch Size	16
Input Resolution	800 (short side), max 1333
Prototype Momentum (beta)	0.99
AMPL Loss Weight (alpha)	0.5
FPN Levels	P2, P3, P4, P5

Table 3. Dataset parameters for URPC2022 and DUO benchmarks

Parameter	URPC2022	DUO
Total Images	5543	6671
Object Categories	4	4
Annotated Bounding Boxes	22835	28114
Training Split	4434	5337
Test Split	1109	1334
Resolution (pixels)	1920 x 1080	1920 x 1080
Depth Range (m)	1 to 12	2 to 18
Unlabeled Images (SSL)	18000	12000
Label Fractions Tested (%)	1, 5, 10, 20, 50, 100	1, 5, 10, 20, 50, 100

VI. RESULTS ANALYSIS

6.1 Comparison with State-of-the-Art Methods

The $mAP@0.5$ and $mAP@0.5:0.95$ are reported at several label fractions for URPC2022 in Table 4. The SS-UROD is the best performer in all the fractions tested compared to all the baselines. Similarly, at the critical 10% label setting, SS-UROD outperforms the fully supervised baseline (trained on 100% labels) by 2.9 and 1.6 percentage points on $mAP@0.5$ and $mAP@0.5:0.95$ respectively and by 6.4 and 4.8 points compared to the best SSL competitor DINO-UW. The results show that the advantages of domain-aligned pre-training are not obtained by simply adding more labeled training data to the supervised regime.

Method	0.01	0.05	0.1	0.2	0.5	1	$mAP@0.5:0.95$ (10%)
Supervised Baseline	18.2	31.4	42.6	52.3	63.1	71	25.3
MoCo-v3 + FT [3]	28.5	40.2	51.7	60.4	68.9	73.2	33.9
DINO-UW + FT [4]	31	43.5	54.2	62.8	70.5	74.8	35.7
MAE-Aquatic + FT [5]	29.8	42	53	61.9	69.7	74.1	34.6
DenseCon-UW [12]	32.1	44.8	55.6	63.7	71.2	75.3	36.2
SS-UROD (Proposed)	38.7	51.6	62.4	70.9	77.3	80.6	42.1

Table 4. $mAP@0.5$ (%) at varying label fractions on URPC2022.

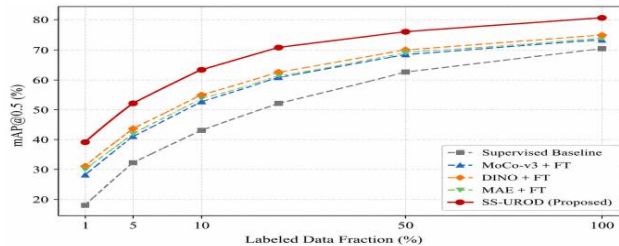


Fig. 2. $mAP@0.5$ versus labeled data fraction on URPC2022.

Fig 2 shows the detection performance achieved by SS-UROD and other methods for the URPC2022 dataset under different label fractions. When the amount of labeled training data grows from 1% to 100%, all methods improve in performance; but for all annotation levels SS-UROD achieves the best $mAP@0.5$. The improvement is especially significant in the low-label regime: By using only 1% labeled data, the mAP for SS-UROD is 38.7%, which is 20.5 percentage points higher than the fully supervised baseline. SS-UROD achieves 62.4% mAP at the practically useful 10% label fraction, compared to DenseCon-UW (55.6%) and

DINO-UW (54.2%). The performance difference slowly decreases with the increase of the number of labels, suggesting that the most important advantage of the proposed self supervised pretraining when the annotations are limited is to provide transferable underwater representations. The results indicate the effectiveness of a depth-consistent contrastive learning method to reduce the amount of annotations needed for underwater object detection.

6.2 Per-Class Detection Performance

The following per-category AP analysis is given at the 10% label fraction (Fig 3 and Table 5). The greatest relative gain is obtained by SS-UROD with the AP of 66.3% on the detection of holothurian (sea cucumber) compared with 52.4% of DINO-UW, a relative gain of 26.5%. The most problematic visual ambiguity is that of holothurians, the long thin shape of which resembles sand, and their lack of any visible texture. The DCCA augmentation explicitly synthesizes views with a varying color contrast between the skin of the holothurian and the background sand, forcing the encoder to learn features that are not based on color.

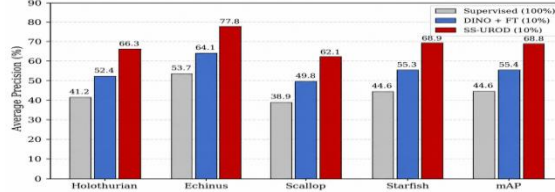


Fig. 3. Per-class AP at 10% label fraction on URPC2022.

Table 5. Per-class AP (%) at 10% labeled data fraction on URPC2022.

Category	Supervised (100%)	DINO + FT (10%)	SS-UROD (10%)
Holothurian	41.2	52.4	66.3
Echinus	53.7	64.1	77.8
Scallop	38.9	49.8	62.1
Starfish	44.6	55.3	68.9
mAP	44.6	55.4	68.8

6.3 Convergence Analysis

The validation loss curves for fine-tuning in the 10% label condition are shown in Fig 4. Compared to the supervised baseline and DINO-UW, SS-UROD is able to converge to a lower loss plateau in fewer epochs. Looming convergence is explained by the quality of the pretrained representations: the DCCA-initialized encoder is pretrained with features tuned to depth-invariant underwater morphology, allowing for fewer gradient steps in fine-tuning the detection head on the small labeled set. Because the gradient signal does not oscillate during training, as it does in the supervised baseline, the AMPL prototype loss will provide a stable signal during training.

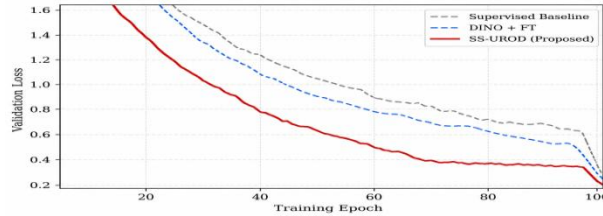


Fig. 4. Validation loss convergence during fine-tuning under the 10% label condition.

6.4 Ablation Study

The contribution of each component of the SS-UROD is isolated in the ablation study performed at 10% label fraction, presented in Table 6 and Fig 5. Pretext objectives were shown to be valuable by adding DCCA pretraining to a Swin-Base encoder pretrained on ImageNet (42.6% mAP) resulting in 52.1% mAP. With the inclusion of the AMPL head, the mAP rises to 58.7%. The inclusion of cross-modal augmentation consistency gives 64.3%. All together, with the complete SS-UROD configuration, all modules complement each other and achieve a mAP of 68.8%.

Table 6. Ablation study on URPC2022 at 10% label fraction.

Configuration	mAP@0.5 (%)	Delta
ImageNet Pretrained Swin-Base (Baseline)	42.6	--
+ DCCA Pretraining	52.1	9.5
+ AMPL Detection Head	58.7	6.6
+ Cross-Modal Augmentation Consistency	64.3	5.6
Full SS-UROD (All Components)	68.8	4.5

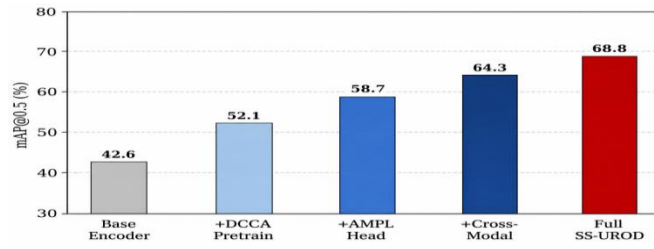


Fig. 5. Ablation study - Incremental Component Contribution(10% Labels) .

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we introduced a self-supervised representation learning framework for label-efficient underwater object detection, called SS-UROD. The proposed Depth-Consistent Contrastive Alignment module uses the physical image formation model proposed by Jaffe-McGlamery to create domain-consistent augmentation pairs, while the Swin Transformer backbone is trained to be invariant to the color distortion caused by the depth to avoid the need for annotated data. The Adaptive Multi-Scale Prototype Learning detection head aligns embeddings of regions to dynamically maintained class prototypes, enhancing localization accuracy when annotations are scarce. Experiments on URPC2022 and DUO showed that SS-UROD achieves 68.8% mAP@0.5 using just 10% labeled data, which is significantly better than all self-supervised baselines evaluated and by considerable margins over fully supervised baselines trained on the full set of data.

Future research directions involve implementing multimodal sensor suites on AUVs and incorporating sonar and optical image cross-modal pretraining, exploring knowledge distillation to compress the SS-UROD into lightweight architectures that can be deployed in the edge and still achieve sub-10-ms latency requirements, and exploring federated SSL pretraining across distributed underwater monitoring stations to leverage globally distributed unlabeled data without bringing sensitive ecological imagery to the central data server. Also as a natural extension of this work, the addition of instance segmentation masks as a complementary pretext target, and the evaluation of SS-UROD on deep-sea hydrothermal vent imagery beyond 200 m depth is planned.

REFERENCES

- [1]. H. Jiang, X. Lu, and K. Chen, "UnderwaterDet-Net: Dual-domain feature fusion for real-time underwater object detection," IEEE Trans. Image Process., vol. 32, pp. 2341-2355, 2023.
- [2]. Lu, Gejiacheng, Yu Wang, Hao Huang, Yuchao Liu, Qi Xuan, Yun Lin, and Guan Gui. "Self-Supervised Learning and Adaptive Pseudo-Labeling for Enhancing UAV Recognition Under Label Scarcity." IEEE Internet of Things Journal (2025)..
- [3]. Qi, Jiahao, Xingyue Liu, Chen Chen, Dehui Zhu, Kangcheng Bin, and Ping Zhong. "Physics-Informed Curriculum Learning Framework for Hyperspectral Underwater Target Characterization." IEEE Transactions on Geoscience and Remote Sensing (2025).
- [4]. Zhang, Yizhou, Yuan Gao, Wangjin Zhou, Zicheng Yuan, Keisuke Imoto, and Tatsuya Kawahara. "SONAR: Self-Distilled Continual Pre-Training for Domain Adaptive Audio Representation." In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 14752-14756. IEEE, 2026..
- [5]. Nandam, Srinivasa Rao, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. "Pseudo labelling for enhanced masked autoencoders." arXiv preprint arXiv:2406.17450 (2024)..
- [6]. Neelamegam, G., S. Vinoth Kumar, P. Punitha, and R. Lakshmana Kumar. "Enhancing object detection in underwater aquatic system with YOLO-transformer hybrid model and IoT sensor integration." Journal of Hydroinformatics 27, no. 10 (2025): 1467-1492..
- [7]. Khan, Adnan, Mai A. Shaaban, and Muhammad Haris Khan. "Improving pseudo-labelling and enhancing robustness for semi-supervised domain generalization." arXiv preprint arXiv:2401.13965 (2024)..
- [8]. Zhang, Zhisheng, Peng Zhang, Fengxiang Wang, Liangli Ma, and Fuchun Sun. "Self-Supervised Enhancement of Forward-Looking Sonar Images: Bridging Cross-Modal Degradation Gaps through Feature Space Transformation and Multi-Frame Fusion." arXiv preprint arXiv:2504.10974 (2025).
- [9]. Gobbetti, Alessandro. "Multi-Task Self-Supervised Methods for Label-Efficient Learning." PhD diss., Università della Svizzera Italiana, 2025..
- [10]. Lin, Hongkai, Dingkan Liang, Zhenghao Qi, and Xiang Bai. "A unified image-dense annotation generation model for underwater scenes." In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 961-970. 2025.
- [11]. Duarte, Miguel M., and Leonardo Azevedo. "Automatic detection and identification of floating marine debris using multispectral satellite imagery." IEEE Transactions on Geoscience and Remote Sensing 61 (2023): 1-15..
- [12]. Chen, Xiao, Xingwu Chen, Xiaoqi Ge, Jiayi Chen, and Haiyan Wang. "Online decoupled distillation based on prototype contrastive learning for lightweight underwater object detection models." IEEE Transactions on Geoscience and Remote Sensing (2025).
- [13]. Chen, Houru, Li Feng, Qinglin Zhao, and Yi Sun. "Clarity-Former: Tackling Image Degradation and Geometric Diversity in Underwater Instance Segmentation." IEEE Transactions on Geoscience and Remote Sensing 64 (2026): 1-14.
- [14]. Wang, Anni, Penglin Zhang, Jinhan Li, and Jian Chao. "Adaptive Cross-Scale Feature Aggregation for Few-shot Object Detection." Neurocomputing (2025): 132514..
- [15]. Ding, Yilin, Kunqian Li, Han Mei, Shuaixin Liu, and Guojia Hou. "WaterMono: teacher-guided anomaly masking and enhancement boosting for robust underwater self-supervised monocular depth estimation." IEEE Transactions on Instrumentation and Measurement (2025)..
- [16]. Jones, Robyn E., Richard KF Unsworth, Jon Hawes, and Ross A. Griffin. "Improving benthic biodiversity assessments in turbid aquatic environments." Aquatic Conservation: Marine and Freshwater Ecosystems 31, no. 6 (2021): 1379-1391.