Boundary-Aware Refinement with Environment-Robust Adapter Tuning for Underwater Instance Segmentation

Pin-Chi Pan r12942103@ntu.edu.tw
Graduate Institute of Communication Engineering, National Taiwan University, Taiwan

Soo-Chang Pei peisc@ntu.edu.tw
Department of Electrical Engineering, National Taiwan University, Taiwan

Editors: Hung-yi Lee and Tongliang Liu

Abstract

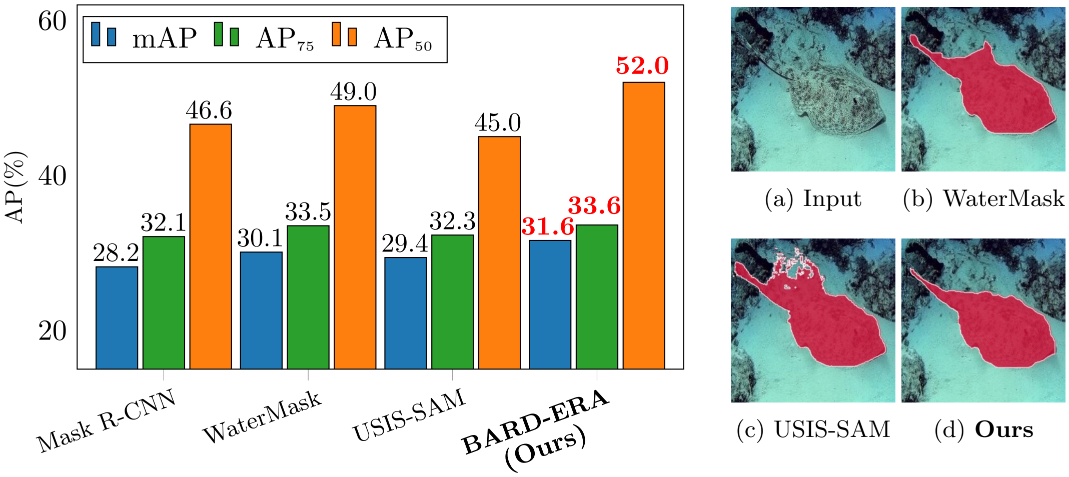
Underwater instance segmentation is a challenging task due to adverse visual conditions such as light attenuation, scattering, and color distortion, which severely degrade image quality and hinder model performance. In this work, we propose **BARD-ERA**, a unified framework that integrates three novel components to address these challenges. First, the **Boundary-Aware Refinement Decoder (BARDecoder)** improves mask quality through progressive feature refinement and lightweight upsampling using a Multi-Stage Gated Refinement Network and Depthwise Separable Upsampling. Second, the **Environment-Robust Adapter (ERA)** enables efficient adaptation to underwater degradations by injecting environment-specific priors with over 90% fewer trainable parameters than full fine-tuning. Third, the **Boundary-Aware Cross-Entropy (BACE) loss** enhances boundary supervision by leveraging range-null space decomposition. Together, these modules achieve state-of-the-art performance on the UIIS dataset, surpassing Mask R-CNN by 3.4 mAP with Swin-B and 3.8 mAP with ConvNeXt V2-B, while maintaining a compact model size. Our results demonstrate that BARD-ERA enables robust, accurate, and efficient segmentation in complex underwater scenes. The source code is available at https://github.com/PANpinchi/BARD-ERA.

Keywords: Underwater Instance Segmentation; Environment-Robust Adapter Tuning; Boundary-Aware Refinement; Boundary-Aware Cross-Entropy

1. Introduction

Instance segmentation is a fundamental task in computer vision, with applications in autonomous robotics, medical imaging, remote sensing, and environmental monitoring Liu et al. (2020). While substantial progress has been made in terrestrial environments, underwater instance segmentation remains highly challenging due to various visual distortions such as light attenuation, scattering, and wavelength-dependent color shifts Akkaynak et al. (2017). These degradations obscure object boundaries and vary dynamically with depth and lighting. Additionally, suspended particles (marine snow) and surface reflections further complicate the scene, leading to misclassifications and the loss of fine-grained details. As a result, segmentation models designed for terrestrial datasets often fail to generalize well to underwater scenes due to differences in texture, lighting, and water clarity.

Existing approaches rely on multi-scale feature fusion Lian et al. (2023) or adapter-based tuning Lian and others. (2024), enhancing segmentation performance. Techniques such as RefineMask Zhang et al. (2021) and WaterMask Lian et al. (2023) exploit multi-scale features for improved contextual representation. However, RefineMask's multi-branch design



(a) Input (b) WaterMask

(c) USIS-SAM (d) **Ours**

Figure 1: Comparison of our approach with state-of-the-art methods on the UIIS dataset. USIS-SAM Lian and others. (2024) uses a ViT-H backbone, while all other methods adopt Swin-B. Our BARD-ERA method achieves the best performance across all AP metrics.

is parameter-heavy, and WaterMask aggregates features from limited pyramid levels without stage-wise refinement. Both lack mechanisms for progressive refinement and efficient upsampling, which can hinder boundary precision and detail recovery in cluttered scenes. In contrast, USIS-SAM Lian and others. (2024) utilizes a salient feature prompt generator to generate prompts via multi-scale fusion. While this guides saliency-aware segmentation, it does not explicitly refine mask boundaries or perform stage-wise feature enhancement.

To address these challenges, we introduce the Boundary-Aware Refinement Decoder (BARDecoder), aimed at enhancing instance segmentation through progressive feature refinement and boundary enhancement. Unlike conventional feature pyramid networks, BARDecoder incorporates a Multi-Stage Gated Refinement Network (MSGRN) for hierarchical refinement and Depthwise Separable Upsampling (DSU) for efficient multi-scale feature fusion, leading to more precise mask delineation with fewer parameters. In addition, we propose the Environment-Robust Adapter (ERA), a plug-and-play adapter tuning strategy tailored for underwater imagery. As illustrated in Fig. 2, ERA is inserted after each transformer or convolutional block, using lightweight modules to capture environment-specific priors. This design effectively mitigates underwater degradations such as scattering and color shifts, while reducing trainable parameters by over 90% compared to full fine-tuning. To further improve boundary localization, we propose the Boundary-Aware Cross-Entropy (BACE) loss, which explicitly enhances mask quality by refining object contours. By integrating BARDecoder, ERA, and BACE loss, our full model BARD-ERA achieves robust and efficient underwater instance segmentation, dynamically adapting to challenging visual conditions without incurring excessive inference overhead.

As shown in Fig. 1, BARD-ERA achieves consistent improvements over existing methods in both qualitative and quantitative evaluations on the UIIS dataset. Integrating BARD-ERA into Mask R-CNN achieves AP gains of 3.4, 1.5, and 5.4 in mAP, AP_{75}, and AP_{50}, respectively, over the baseline Mask R-CNN He et al. (2017) with Swin-B backbone. These results highlight the effectiveness of our approach in addressing underwater imaging challenges. In summary, the main contributions of this work are summarized as follows:

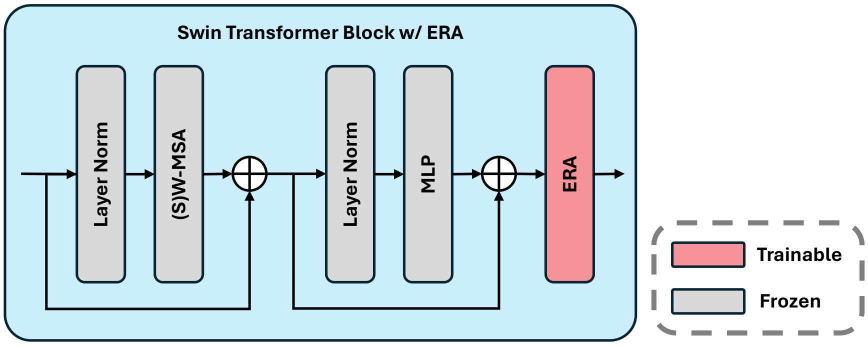


Figure 2: The Environment-Robust Adapter (ERA) integrated into a Swin Transformer block. This design enables efficient adaptation to underwater distortions without modifying the core model architecture.

- To enhance mask precision and boundary quality, we introduce Boundary-Aware Refinement Decoder (BARDecoder), which refines multi-scale features progressively using a gated refinement network and depthwise separable upsampling.
- To enable efficient adaptation to underwater distortions, Environment-Robust Adapter (ERA) is designed as a lightweight tuning strategy that captures environment-specific priors with over 90% fewer trainable parameters than full fine-tuning.
- In order to improve mask quality, Boundary-Aware Cross-Entropy (BACE) loss introduces boundary supervision to sharpen object contours and reduce ambiguity.
- Extensive experiments validate the effectiveness of BARD-ERA, establishing it as a state-of-the-art approach for underwater instance segmentation.

2. Related Work

2.1. Underwater Image Segmentation

Underwater image segmentation remains challenging due to environmental distortions such as light attenuation, scattering, and color degradation, which obscure object boundaries and diminish feature contrast. Early benchmarks like EUVP Islam et al. (2020) and SAUD Jiang et al. (2022) focus on image enhancement and color correction, while datasets such as USIS10K Lian and others. (2024), UIIS Lian et al. (2023), and DeepFish Garcia-D'Urso et al. (2022) emphasize biodiversity and fine-grained segmentation.

Recent methods explore multi-scale feature refinement to boost segmentation accuracy under underwater conditions. WaterMask Lian et al. (2023) fuses features across pyramid levels to improve context, but lacks stage-wise refinement, limiting its ability to capture fine details. Such approaches improve segmentation but struggle with boundary preservation in degraded scenes. USIS-SAM Lian and others. (2024) integrates semantic priors via prompt-based learning in a transformer framework. Although effective, its reliance on high-level semantic information and large backbones (e.g., ViT-H) results in high computational cost and slower inference, limiting real-time applicability. Thus, achieving accurate and efficient segmentation across diverse underwater environments remains an open challenge.

2.2. Adapter-Tuning

Adapter-tuning is an efficient transfer learning technique that introduces small, trainable modules into frozen pretrained networks, reducing the need for full fine-tuning. Originally developed for natural language processing Houlsby et al. (2019); Tinn et al. (2023), this approach has gained traction in vision tasks through methods like AdaptFormer Chen et al. (2022), Polyhistor Liu et al. (2022a), and Mona-tuning Yin et al. (2023). These techniques have demonstrated success in classification and dense prediction tasks by enabling models to adapt to new domains with fewer trainable parameters. In underwater segmentation, USIS-SAM Lian and others. (2024) incorporates adapter-based tuning to integrate domain-specific priors. However, existing adapter methods primarily focus on feature modulation and do not explicitly counteract underwater-specific distortions, such as scattering and wavelength-dependent attenuation. While adapter-tuning efficiently reduces training costs, its effectiveness in handling complex underwater degradations and segmentation challenges remains an area requiring further exploration.

3. Method

This section introduces our method with three components: BARDecoder (Section 3.1), ERA-Tuning (Section 3.2), and Boundary-Aware Cross-Entropy Loss (Section 3.3).

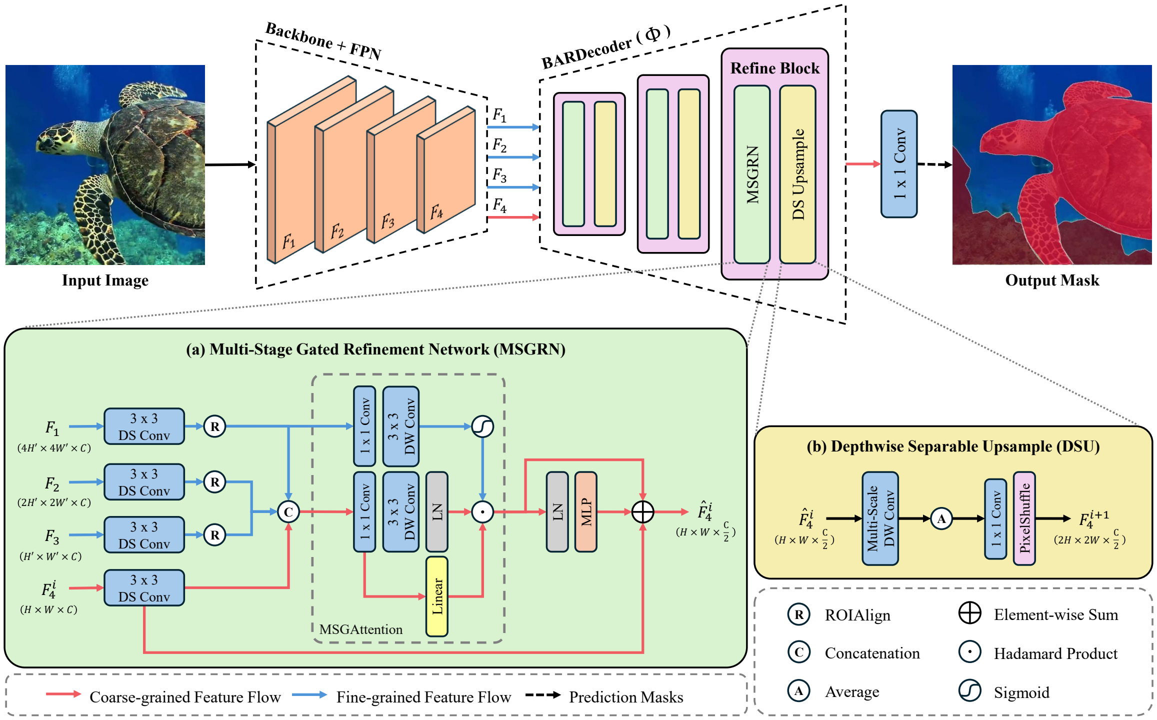

Input Image
Backbone + FPN
F_1
F_2
F_3
F_4
BARDecoder (Φ)
Refine Block
MSGRN
DS Upsample
1 x 1 Conv
Output Mask
(a) Multi-Stage Gated Refinement Network (MSGRN)
F_1 $(4H' \times 4W' \times C)$
F_2 $(2H' \times 2W' \times C)$
F_3 $(H' \times W' \times C)$
F_4^i $(H \times W \times C)$
3 x 3 DS Conv
R
C
1 x 1 Conv
3 x 3 DW Conv
LN
Linear
MSGAttention
LN
MLP
$\hat{F}_4^i$ $(H \times W \times \frac{C}{2})$
(b) Depthwise Separable Upsample (DSU)
$\hat{F}_4^i$ $(H \times W \times \frac{C}{2})$
Multi-Scale DW Conv
A
1 x 1 Conv
PixelShuffle
F_4^{i+1} $(2H \times 2W \times \frac{C}{2})$
R ROIAlign
C Concatenation
A Average
⊕ Element-wise Sum
⊙ Hadamard Product
Sigmoid
Coarse-grained Feature Flow
Fine-grained Feature Flow
Prediction Masks

Figure 3: The architecture of the proposed BARDecoder for underwater instance segmentation. BARDecoder consists of (a) Multi-Stage Gated Refinement Network (defined in Section 3.1.1) and (b) Depthwise Separable Upsample (defined in Section 3.1.2).

3.1. BARDecoder

The BARDecoder (Fig. 3) refines instance segmentation masks by progressively fusing multi-scale features using gated attention and efficient upsampling. While prior works such as RefineMask Zhang et al. (2021) and WaterMask Lian et al. (2023) leverage multi-level features, they either rely on parameter-heavy branches or limited pyramid aggregation without stage-wise refinement. BARDecoder overcomes these limitations by introducing two novel components: 1) the Multi-Stage Gated Refinement Network (MSGRN) for selective multi-scale fusion and 2) the Depthwise Separable Upsample (DSU) module for high-resolution mask reconstruction. These design choices lead to improved boundary precision and segmentation accuracy while maintaining computational efficiency. Given multi-scale features $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4\}$ from the backbone, the final segmentation mask $\boldsymbol{M}_{out}$ is produced as:

$$\boldsymbol{M}_{out} = Conv_{1\times1}(\Phi(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4)), \tag{1}$$

where $\Phi(\cdot)$ represents BARDecoder, which processes multi-scale features using sequential refinement blocks. Each block applies MSGRN and DSU to refine feature quality:

$$\begin{aligned} \hat{\boldsymbol{F}}_4^i &= \mathcal{M}_{MSGRN}^i(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4^i), \\ \boldsymbol{F}_4^{i+1} &= \mathcal{M}_{DSU}^i(\hat{\boldsymbol{F}}_4^i), \end{aligned} \tag{2}$$

where $\boldsymbol{F}_4^i$ are the features from the i-th refinement stage.

3.1.1. Multi-Stage Gated Refinement Network

The Multi-Stage Gated Refinement Network (MSGRN) progressively refines multi-scale features to enhance spatial details, as illustrated in Fig.3 (a). Unlike conventional fusion methods, it uses Multi-Scale Gated Attention (MSGAttention) to selectively emphasize informative regions and suppress redundancy, improving boundary precision.

Inspired by High-Order Spatial Attention (HSA) from SegAdapter Peng and Kameyama (2024), which modulates global features via self-gating, MSGAttention adaptively adjusts feature weights at multiple scales to refine object boundaries. The process begins with depthwise separable convolutions (DSConv) for multi-scale feature extraction:

$$\begin{aligned} \boldsymbol{X}_n &= DSConv_{3\times3}(\boldsymbol{F}_n),\ n \in \{1,2,3,4\}, \\ \boldsymbol{X}'_n &= ROIAlign(\boldsymbol{X}_n),\ n \in \{1,2,3\}, \\ \boldsymbol{X} &= Concat(\boldsymbol{X}_4, \{\boldsymbol{X}'_n\}_{n=1}^3) \end{aligned} \tag{3}$$

We then apply MSGAttention to enhance the fused features:

$$\begin{aligned} \hat{\boldsymbol{X}} &= Conv_{1\times1}(\boldsymbol{X}), \\ \boldsymbol{Y} &= LN(Conv_{3\times3}(\hat{\boldsymbol{X}})), \\ \boldsymbol{V} &= Linear(\hat{\boldsymbol{X}}), \\ \boldsymbol{W} &= DSConv_{3\times3}(\boldsymbol{X}'_1), \\ \hat{\boldsymbol{Z}} &= MSGAttention(\boldsymbol{X}, \boldsymbol{X}'_1) = \sigma_{sig}(\boldsymbol{W}) \odot (\boldsymbol{Y} \odot \boldsymbol{V}), \\ \boldsymbol{Z} &= FFN(\hat{\boldsymbol{Z}}) = MLP(LN(\hat{\boldsymbol{Z}})) + \hat{\boldsymbol{Z}} \end{aligned} \tag{4}$$

Here, σ_{sig} represents the sigmoid activation, $\odot$ denotes the Hadamard product, and LN denotes layer normalization Xu et al. (2019). While SegAdapter's HSA globally adjusts features using high-level semantic priors, MSGAttention locally refines multi-scale features to enhance spatial details. To stabilize the refinement and preserve previously learned representations, we integrate a residual connection:

$$\hat{\boldsymbol{F}}_4^i = \boldsymbol{Z} + \boldsymbol{X}_4, \tag{5}$$

This refinement strategy enhances boundary accuracy while maintaining model compactness and efficiency, which is well suited for resource-constrained underwater applications.

3.1.2. Depthwise Separable Upsample

The Depthwise Separable Upsample (DSU) module, shown in Fig. 3 (b), enhances spatial resolution while preserving feature integrity. Unlike bilinear interpolation, DSU combines multi-scale depthwise convolutions with pixel shuffle, capturing fine-grained details efficiently. This approach enables effective multi-level feature fusion ($\boldsymbol{F}_1$ to $\boldsymbol{F}_4$) while reducing computational overhead. We first extract multi-scale features from $\hat{\boldsymbol{F}}_4^i$:

$$\hat{\boldsymbol{F}}_4^{i,j} = DWConv_{j\times j}(\hat{\boldsymbol{F}}_4^i), \ j \in \{3, 5, 7\}. \tag{6}$$

Then, aggregated features are passed through upsampling:

$$\boldsymbol{F}_4^{i+1} = PS(Conv_{1\times 1}(\text{Average}(\{\hat{\boldsymbol{F}}_4^{i,j}\}_{j\in\{3,5,7\}}))), \tag{7}$$

Here, PS denotes pixel shuffle. This design enables efficient detail recovery and supports progressive refinement, improving underwater mask quality with minimal overhead.

Together, MSGRN and DSU enable BARDecoder to progressively refine features with high spatial fidelity, which is crucial for accurate and detail-preserving segmentation in underwater scenes with complex degradations. This design achieves strong performance with significantly fewer parameters than prior multi-branch decoders, as demonstrated in our ablation study (Section 4.4).

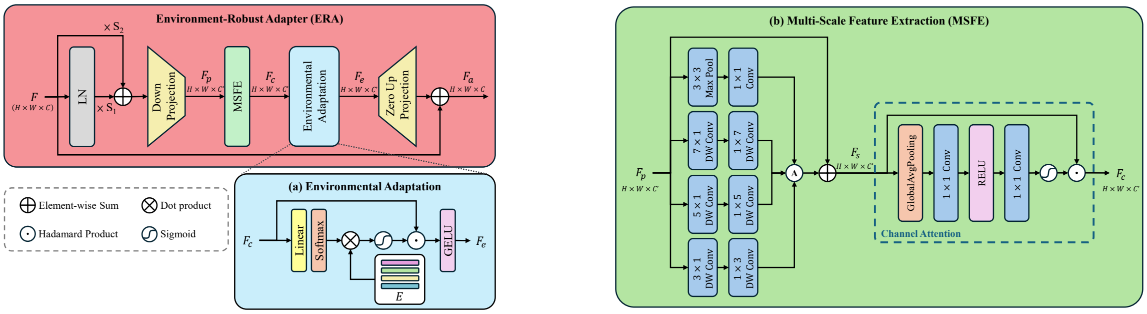


Figure 4: The architecture of the Environment-Robust Adapter (ERA). ERA enhances feature representations through multi-scale feature extraction and environmental adaptation.

3.2. ERA-Tuning

ERA-tuning extends the adapter-based paradigm to address underwater degradation, inspired by Mona-Tuning Yin et al. (2023) and recent prompt-based techniques like USIS-SAM Lian and others. (2024). Unlike USIS-SAM, which integrates general semantic priors, ERA learns environmental embeddings to capture underwater degradation patterns. As shown in Fig. 4, the input features first pass through a normalization layer, followed by two learnable scaling factors, S_1 and S_2, which adaptively modulate the feature representation. ERA then applies a down-projection, mapping features from $\mathbb{R}^{H \times W \times C}$ to $\mathbb{R}^{H \times W \times C'}$, where $C' = C/\gamma$. The ratio γ controls feature compression and influences adaptability, as further analyzed in the supplementary material. Beyond environmental adaptation, effective feature extraction is crucial for robust segmentation. To further enhance spatial representations, we introduce the Multi-Scale Feature Extraction (MSFE) module.

3.2.1. Multi-Scale Feature Extraction

The Multi-Scale Feature Extraction (MSFE) enhances spatial representations by capturing information at multiple receptive fields. Inspired by iFormer Si et al. (2022), where diverse kernel sizes enable robust feature learning. Fig. 4 (b) illustrates the architecture of MSFE, showcasing the combination of multi-scale depthwise convolutions, max pooling, and channel attention for robust feature learning. Specifically, MSFE applies multiple depthwise separable convolutions and max-pooling layers to improve feature discrimination:

$$\begin{aligned}
\boldsymbol{F}_{s,max} &= Conv_{1\times1}(MaxPooling(\boldsymbol{F}_p)), \\
\boldsymbol{F}^{j}_{s,conv} &= DWConv_{1\times j}(DWConv_{j\times1}(\boldsymbol{F}_p)), j \in \{3,5,7\}, \\
\boldsymbol{F}_s =& Average(\boldsymbol{F}_{s,max}, \{\boldsymbol{F}^{j}_{s,conv}\}_{j\in\{3,5,7\}}) + \boldsymbol{F}_p,
\end{aligned} \tag{8}$$

To further enhance feature representation, MSFE integrates a Channel Attention (CA) mechanism following USIS-SAM Lian and others. (2024). The CA module dynamically reweights feature channels to emphasize discriminative spectral information:

$$\begin{aligned}
\boldsymbol{S} &= Conv_{1\times1}(\delta(Conv_{1\times1}(GAP(\boldsymbol{F}_s)))), \\
\boldsymbol{F}_c &= CA(\hat{\boldsymbol{F}}_s) = \hat{\boldsymbol{F}}_s \odot \sigma_{sig}(\boldsymbol{S}),
\end{aligned} \tag{9}$$

where δ is the RELU activation, σ_{sig} is the sigmoid function, and $\odot$ denotes element-wise multiplication. CA improves feature discriminability by emphasizing informative channels, which is beneficial in underwater conditions with spectral distortion.

3.2.2. Environmental Adaptation

Fig. 4 (a) illustrates the Environmental Adaptation module, which leverages learned environmental priors to modulate features based on underwater conditions. The environmental adaptation module employs learnable embeddings $\boldsymbol{E} \in \mathbb{R}^{N \times C}$, where N represents predefined underwater conditions, to model degradation variations. By learning distinct embeddings, the module dynamically modulates features based on the observed scene. A per-pixel environmental descriptor is first computed by projecting feature maps $\boldsymbol{F}_c$ into the environmental embedding space:

$$\boldsymbol{E}_{adapted} = \sigma_{soft}(Linear(\boldsymbol{F}_c)) \otimes \boldsymbol{E}, \tag{10}$$

where σ_{soft} is the Softmax function, ensuring each pixel receives a probabilistic weighting over environmental types. This allows the model to emphasize features relevant to specific conditions, such as light absorption, scattering, and turbidity. The computed priors then modulate feature representations through a weighted gating mechanism:

$$\boldsymbol{F}_e = \phi(\boldsymbol{F}_c \odot \sigma_{sig}(\boldsymbol{E}_{adapted})), \tag{11}$$

where σ_{sig} denotes the Sigmoid function and ϕ is the GELU activation. This formulation enhances relevant features while mitigating underwater degradations like light attenuation and color distortion. Finally, a zero-initialized up-projection step follows Zhang et al. (2023) to stabilize early training while preserving environmental priors:

$$\boldsymbol{F}_a = \boldsymbol{F} + \mathcal{Z}(\boldsymbol{F}_e), \tag{12}$$

where $\mathcal{Z}$ is a projection initialized to zero. This stabilizes early training while preserving adaptation priors. Combining environment-specific modulation with efficient parameter tuning, ERA enables adaptation to underwater scenes without retraining the backbone.

3.3. Boundary-Aware Cross-Entropy Loss

Boundary-Aware Cross-Entropy (BACE) loss improves segmentation quality by guiding the model to better focus on learning accurate boundary information. BACE loss leverages range-null space decomposition, a fundamental concept in linear algebra widely applied in inverse problems Wang et al. (2023a,b). We observe that when applied to segmentation, this decomposition effectively preserves non-boundary structures while refining ambiguous edges, facilitating clearer and more accurate boundary representations.

3.3.1. Range-Null Space Decomposition

Given a transformation matrix $\boldsymbol{A} \in \mathbb{R}^{d\times D}$, its pseudo-inverse $\boldsymbol{A}^\dagger \in \mathbb{R}^{D\times d}$ satisfies:

$$\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{A}. \tag{13}$$

Any vector $\boldsymbol{x} \in \mathbb{R}^D$ can be decomposed into range-space and null-space components:

$$\boldsymbol{x} = \boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{A}^\dagger\boldsymbol{A})\boldsymbol{x}. \tag{14}$$

The first term projects $\boldsymbol{x}$ onto the range space of $\boldsymbol{A}$, preserving coarse structural content, while the second captures residual information in the null space, often associated with high-frequency boundary information. This decomposition allows us to selectively emphasize different components of the learning signal during training.

3.3.2. Application in Instance Segmentation

To apply this to segmentation, construct a boundary-aware target Γ by combining structural guidance from ground truth $\boldsymbol{M}_{\text{gt}}$ and boundary emphasis from model predictions $\boldsymbol{M}_\theta$:

$$\Gamma(\boldsymbol{M}_\theta, \boldsymbol{M}_{\text{gt}}) = \boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{M}_{\text{gt}} + (\boldsymbol{I} - \boldsymbol{A}^\dagger\boldsymbol{A})\boldsymbol{M}_\theta. \tag{15}$$

Here, $\boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{M}$gt emphasizes stable non-boundary regions, while $(\boldsymbol{I} - \boldsymbol{A}^\dagger\boldsymbol{A})\boldsymbol{M}\theta$ encourages the model to refine focus on more ambiguous boundary areas. This blended supervision leads the model to sharpen object contours without relying on explicit boundary annotations.

In practice, $\boldsymbol{A}$ is implemented as a max-pooling operator that extracts dominant features, while $\boldsymbol{A}^\dagger$ is a nearest-neighbor upsampling that restores spatial resolution. This design induces the model to learn boundary-aware representations through structured supervision.

3.3.3. Final Loss Function

The BACE loss is defined as:

$$\mathcal{L}_{BACE}(\boldsymbol{M}_\theta, \boldsymbol{M}_{gt}) = \frac{1}{N}\sum_{i=1}^{N} BCE(\boldsymbol{M}_{gt}^{i}, \Gamma(\boldsymbol{M}_\theta, \boldsymbol{M}_{gt})^{i}), \tag{16}$$

and integrated into the total training objective:

$$\mathcal{L}_{Total}(\boldsymbol{M}_\theta, \boldsymbol{M}_{gt}) = \mathcal{L}_{CE}(\boldsymbol{M}_\theta, \boldsymbol{M}_{gt}) + \lambda \cdot \mathcal{L}_{BACE}(\boldsymbol{M}_\theta, \boldsymbol{M}_{gt}), \tag{17}$$

where $\lambda = 1$ balances the contribution of standard cross-entropy and BACE loss. This formulation helps better localize object boundaries, improving segmentation precision.

4. Experiments

We conduct extensive experiments to evaluate the effectiveness of BARDecoder, ERA-tuning, and BACE loss across multiple instance segmentation benchmarks. Section 4.1 details the experimental setup, including datasets, baselines, and metrics. Section 4.2 compares with state-of-the-art methods, and Section 4.3 analyzes ERA-tuning efficiency and adaptability. Section 4.4 presents ablation studies on individual modules, boundary-aware losses, and refinement strategies. Supplementary material includes extended ablations, detailed ERA analyses, and comparisons of boundary-aware loss variants such as Laplacian filtering and range-null space decomposition.

4.1. Implementation Details

We evaluate our approach on the Underwater Image Instance Segmentation (UIIS) dataset Lian et al. (2023) and the Underwater Salient Instance Segmentation (USIS10K) dataset Lian and others. (2024). The UIIS dataset consists of 3,937 training images and 691 validation images, covering diverse underwater visibility conditions. USIS10K, a larger dataset, includes 10,632 images with more complex underwater environments.

We compare BARD-ERA against leading instance segmentation frameworks, including Mask R-CNN He et al. (2017), Cascade Mask R-CNN Cai and Vasconcelos (2018), PointRend Kirillov et al. (2020), SOLOv2 Wang et al. (2020), Mask2Former Cheng et al. (2022), WaterMask Lian et al. (2023), and USIS-SAM Lian and others. (2024). Additionally, we evaluate ERA-tuning against mainstream parameter-efficient tuning methods, including BitFit Zaken et al. (2021); Cai et al. (2020), NormTuning Giannou et al. (2023), PARTIAL-1 Yosinski et al. (2014), Adapter Houlsby et al. (2019), LoRA Hu et al. (2021), AdapterFormer Chen et al. (2022), and MONA Yin et al. (2023). To ensure a fair comparison, all methods use either the Swin Transformer Liu et al. (2021) or ConvNeXt V2 Liu et al. (2022b) backbone, both pre-trained on ImageNet-22k Deng et al. (2009), with the exception of USIS-SAM, which employs a ViT-H backbone. All models are implemented in PyTorch Paszke et al. (2017) using the OpenMMLab framework Chen et al. (2019), and training is conducted on an NVIDIA Titan RTX GPU. For evaluation, we report mask AP Lin et al. (2014) metrics, including mAP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L, to ensure a comprehensive assessment across various IoU thresholds and object sizes.

Underwater Image Instance Segmentation (UIIS)										
Method	**Backbone**	**mAP**	**AP$_{50}$**	**AP$_{75}$**	**Params**	**Backbone**	**mAP**	**AP$_{50}$**	**AP$_{75}$**	**Params**
Mask R-CNN	Swin-B	28.2	46.6	32.1	106.75 M	ConvNeXt V2-B	28.5	46.0	32.3	107.70 M
Cascade Mask R-CNN	Swin-B	29.4	48.0	32.7	139.79 M	ConvNeXt V2-B	28.2	45.2	32.4	140.74 M
Point Rend	Swin-B	29.7	47.7	32.2	118.84 M	ConvNeXt V2-B	30.0	47.7	32.3	119.79 M
SOLOv2	Swin-B	28.6	45.4	30.6	109.00 M	ConvNeXt V2-B	30.8	47.7	33.9	109.95 M
Mask2Former	Swin-B	30.3	45.6	32.4	106.75 M	ConvNeXt V2-B	25.1	38.9	26.7	107.70 M
WaterMask	Swin-B	30.1	49.0	33.5	110.40 M	ConvNeXt V2-B	30.1	48.3	34.4	111.35 M
USIS-SAM	ViT-H	29.4	45.0	32.3	698.12 M	-	-	-	-	-
BARD-ERA (Ours)	Swin-B	31.6	52.0	33.6	114.44 M	ConvNeXt V2-B	32.3	51.4	36.3	112.46 M

Table 1: Quantitative comparison with state-of-the-art methods on the UIIS dataset. USIS-SAM Lian and others. (2024) uses a ViT-H backbone, while all other methods adopt Swin-B and ConvNeXt V2-B backbones. Red indicates the best, blue indicates the second-best.

4.2. Comparison with State-of-the-Art Methods

We evaluate BARD-ERA on the UIIS and USIS10K datasets, comparing its performance against leading instance segmentation methods. As shown in Table 1, BARD-ERA consistently outperforms prior methods on the UIIS dataset. With the Swin-B backbone, our method improves mAP by 3.4, 1.3, and 1.5 over Mask R-CNN He et al. (2017), Mask2Former Cheng et al. (2022), and WaterMask Lian et al. (2023), respectively. With ConvNeXt V2-B, it surpasses Mask R-CNN, Mask2Former, and SOLOv2 Wang et al. (2020) by 3.8, 7.2, and 1.5 mAP. Table 2 further validates our method on the USIS10K dataset, where BARD-ERA outperforms WaterMask by 3.1 mAP and USIS-SAM Lian and others. (2024), which employs a ViT-H backbone, by 4.2 mAP. These results confirm the effectiveness of our approach across diverse underwater segmentation scenarios.

Underwater Salient Instance Segmentation (USIS10K)				
Method	**Backbone**	**Multi-Class**		
		mAP	**AP$_{50}$**	**AP$_{75}$**
WaterMask	ResNet-101	38.7	54.9	43.2
WaterMask	Swin-B	44.2	61.5	49.6
RSPrompter	ViT-H	40.2	55.3	44.8
USIS-SAM	ViT-H	43.1	59.0	48.5
BARD-ERA (Ours)	Swin-B	**47.3**	**65.1**	**53.7**

Table 2: Quantitative comparisons with state-of-the-arts methods on the USIS10K datasets. BARD-ERA follows the same hyperparameters and settings as in Table 1. **Bold:** best.

Figure 5 compares BARD-ERA with underwater-specific methods on the UIIS dataset. WaterMask sometimes fails to segment occluded objects, while USIS-SAM can produce fragmented masks. Our method preserves object structures and successfully recovers missing regions, even under turbidity. Figure 6 shows comparisons with terrestrial models such as SOLOv2 and Mask2Former. These models may produce coarse or misaligned masks on underwater images, whereas BARD-ERA yields cleaner and more accurate segmentations. Figure 8 visualizes predictions on USIS10K dataset, showing that BARD-ERA performs reliably across diverse underwater scenes, including marine life, divers, and artificial objects.

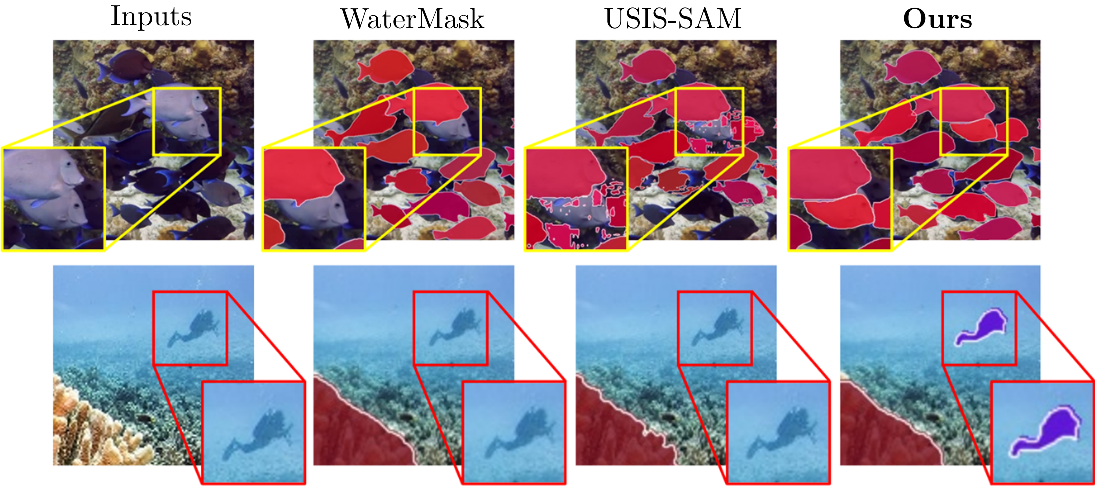


Figure 5: Qualitative comparison with state-of-the-art **underwater-specific** methods on the UIIS dataset, using Swin Transformer (top) and ConvNeXt V2 (bottom) backbones.

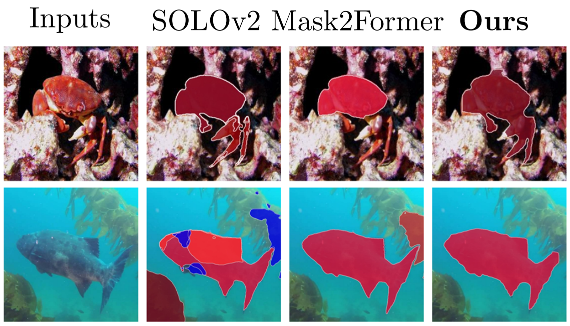


Figure 6: Qualitative comparison with **terrestrial** methods on the UIIS dataset.

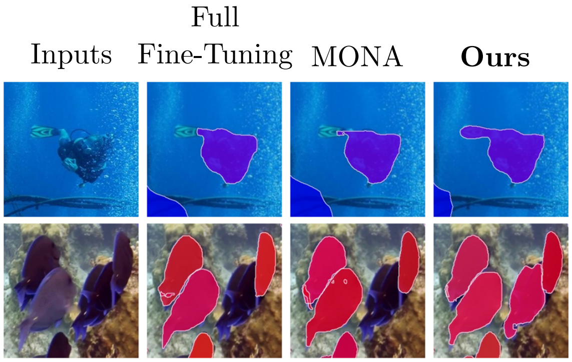


Figure 7: Qualitative comparison with different fine-tuning methods on the UIIS dataset.

4.3. Comparison with Fine-Tuning Methods

We evaluated ERA against various fine-tuning techniques using Swin Transformer backbones on the UIIS dataset. To ensure a fair comparison, we adjust the number of ERA parameters by modifying the compression ratio γ so that its trainable parameter count closely matches that of MONA. This adjustment ensures that the observed improvements come from the effectiveness of the ERA rather than differences in the parameter budget, highlighting the efficiency of our approach. As shown in Table 3, ERA achieves the highest mAP of 29.9, surpassing full fine-tuning by 1.7 mAP while using only 4.67% of the trainable parameters. Compared to MONA Yin et al. (2023), which achieves 28.9 mAP, ERA further improves performance by 1.0 mAP.

Fig. 7 qualitatively compares ERA with full fine-tuning and MONA. ERA better preserves boundaries and reduces segmentation errors, especially under challenging conditions with turbidity and lighting variations. Compared to others, ERA yields more complete segmentations and finer details, reinforcing its robustness in underwater instance segmentation.

Method	Trained Params*	%	Extra Structure	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
				Swin Transformer					
Full Fine-Tuning	86.75 M	100.00 %	✘	28.2	46.6	32.1	9.5	23.4	39.6
BitFit	0.20 M	0.23 %	✘	26.0	46.8	25.5	8.4	21.9	36.9
NormTuning	0.06 M	0.07 %	✘	25.5	45.8	26.0	8.5	21.3	35.0
PARTIAL-1	12.60 M	14.53 %	✘	25.3	47.3	24.5	9.1	20.4	35.2
Adapter	3.11 M	3.46 %	✓	24.3	43.9	23.9	8.7	20.4	33.8
LoRA	3.08 M	3.43 %	✓	25.9	46.8	26.8	9.2	21.2	36.0
AdapterFormer	1.55 M	1.76 %	✓	27.7	49.0	29.6	9.5	22.6	38.8
MONA	3.67 M	4.06 %	✓	28.9	48.7	32.5	10.0	22.5	41.4
ERA (Ours)	4.25 M	4.67 %	✓	29.9	50.5	32.5	10.1	23.5	42.2

Table 3: Quantitative comparison with different fine-tuning methods on UIIS dataset using Swin Transformer backbones. Red indicates the best, and blue indicates the second-best. * denotes the trainable parameters in backbones.

Method	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
Mask R-CNN	28.2	46.6	32.1	9.5	23.4	39.6	106.75 M
w/ BARDecoder	30.0	49.2	32.1	9.5	23.7	42.8	105.14 M
w/ ERA	30.2	51.6	32.0	10.8	23.6	41.9	116.06 M
w/ BACE Loss	29.3	48.4	32.4	10.6	23.5	39.7	106.75 M
Full model (Ours)	**31.6**	**52.0**	**33.6**	**10.7**	**24.0**	**45.0**	114.44 M

Table 4: Effectiveness of each component. Swin-Transformer backbone and 1× training schedule is adopted. **Bold:** best.

4.4. Ablation Studies

Effectiveness of Each Component. We analyze the contribution of each component in BARD-ERA using the Swin Transformer backbone, as shown in Table 4. The Mask R-CNN achieves an mAP of 28.2, serving as the baseline. Incorporating the BARDecoder improves mAP to 30.0, enhancing feature boundaries, refining details, and strengthening multi-scale fusion. ERA-tuning further increases mAP to 30.2, demonstrating its effectiveness in mitigating underwater degradations and improving feature adaptability. BACE Loss boosts boundary refinement, achieving 29.3 mAP. The full model, integrating all components, attains the highest mAP of 31.6, confirming their complementary benefits for underwater instance segmentation.

Effectiveness of Refinement Method. To justify the design of BARDecoder, we compare it with existing refinement modules, including RefineMask Zhang et al. (2021) and WaterMask Lian et al. (2023), as shown in Table 5. While all methods leverage multi-scale feature fusion, BARDecoder introduces a gated refinement mechanism that selectively enhances informative features while preserving structural details. Notably, BARDecoder achieves the highest mAP while maintaining the lowest parameter count among the compared methods. This demonstrates its superior trade-off between accuracy and efficiency, validating its effectiveness in refining object boundaries and improving feature aggregation under resource constraints.

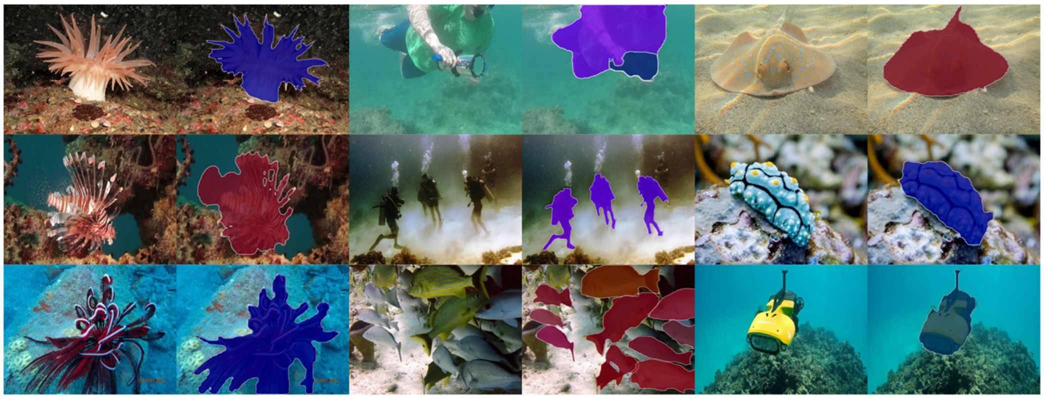


Figure 8: Visualization of prediction results on samples from the USIS10K dataset, using our BARD-ERA method with Swin Transformer backbones.

Method	mAP	AP_{50}	AP_{75}	Params
Mask R-CNN	28.2	46.6	32.1	106.75 M
w/ RefineMask	29.7	47.8	**32.8**	110.35 M
w/ WaterMask	29.3	46.7	32.5	110.40 M
w/ BARDecoder (Ours)	**30.0**	**49.2**	32.1	105.14 M

Table 5: Effectiveness of refinement method. **Bold:** best.

Method	mAP	AP_{50}	AP_{75}
Cross Entropy Loss (CE)	28.2	46.6	32.1
CE + b-awareness Loss	28.9	47.3	32.0
CE + AB Loss	28.5	47.5	32.3
CE + BACE Loss (Ours)	**29.3**	**48.4**	**32.4**

Table 6: Effectiveness of boundary-aware loss. **Bold:** best.

Effectiveness of Different Boundary-Aware Loss. Table 6 compares BACE Loss with other boundary-aware losses. Unlike b-awareness Loss from PIDNet, which applies weighted cross-entropy to emphasize edges, and Active Boundary Loss (ABL), which optimizes local boundary alignment, BACE Loss uses range-null space decomposition to refine boundary consistency while preserving global structure. It achieves an mAP of 29.3, outperforming prior losses and demonstrating effectiveness in challenging segmentation tasks.

5. Conclusion

In this work, we present BARD-ERA, a boundary-aware and environment-adaptive framework for underwater instance segmentation. The proposed BARDecoder refines multi-scale features through a Multi-Stage Gated Refinement Network (MSGRN) and Depthwise Separable Upsampling (DSU), achieving accurate and detail-preserving segmentation with fewer parameters than prior refinement designs. To address underwater degradation, we introduce the Environment-Robust Adapter (ERA), a lightweight tuning module that captures environmental priors with over 90% fewer trainable parameters compared to full fine-tuning. Additionally, the Boundary-Aware Cross-Entropy (BACE) loss guides the model to better learn boundary representations via range-null space decomposition. Extensive experiments on UIIS and USIS10K benchmarks validate the effectiveness of each component, with BARD-ERA achieving state-of-the-art performance in both accuracy and efficiency. In future work, we plan to further improve robustness under severe turbidity and lighting distortions, and extend our framework to broader underwater vision tasks and real-time deployment scenarios.

References

Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, and David Iluz. What is the space of attenuation coefficients in underwater computer vision? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4931–4940, 2017.

Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. *Advances in Neural Information Processing Systems*, 33: 11285–11297, 2020.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Nahuel E Garcia-D'Urso, Alejandro Galan-Cuenca, Pau Climent-Pérez, Marcelo Saval-Calvo, Jorge Azorin-Lopez, and Andres Fuster-Guillo. Efficient instance segmentation using deep learning for species identification in fish markets. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

Angeliki Giannou, Shashank Rajput, and Dimitris Papailiopoulos. The expressive power of tuning only the norm layers. *arXiv preprint arXiv:2302.07937*, 8, 2023.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020.

Qiuping Jiang, Yuese Gu, Chongyi Li, Runmin Cong, and Feng Shao. Underwater image enhancement quality evaluation: Benchmark dataset and objective metric. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5959–5974, 2022.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.

Shijie Lian and others. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In *ICML*, 2024.

Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1305–1315, 2023.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE transactions on circuits and systems for video technology*, 30(12):4861–4875, 2020.

Yen-Cheng Liu, Chih-Yao Ma, Junjiao Tian, Zijian He, and Zsolt Kira. Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. *Advances in Neural Information Processing Systems*, 35:36889–36901, 2022a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022b.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *Advances in neural information processing systems*, 2017.

Dingjie Peng and Wataru Kameyama. Simple and efficient vision backbone adapter for image semantic segmentation. In *Asian Conference on Machine Learning*, pages 1071–1086. PMLR, 2024.

Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *Advances in Neural Information Processing Systems*, 35:23495–23509, 2022.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4), 2023.

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33: 17721–17732, 2020.

Yinhuai Wang, Yujie Hu, Jiwen Yu, and Jian Zhang. Gan prior based null-space learning for consistent super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2724–2732, 2023a.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023b.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

Dongshuo Yin, Leiyi Hu Bin Li, and Youqun Zhang. Adapter is all you need for tuning visual tasks. *arXiv preprint arXiv:2311.15010*, 2023.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6861–6869, 2021.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.