SAM2-MAS: Boundary-Aware Frequency-Adaptive Framework for Marine Animal Segmentation

Ziyi Bao, Jichao Jiao*, Ning Li, Yingchao Zeng, Yingjian Zhang, Peiyuan Zhao,
Yifan Li, Chunyang Wu, Tianxiang Zhang, Jiajie Huang
Beijing University of Posts and Telecommunications
Beijing, China
{baoziyi, jiaojichao, lnmmdsy, yingchaoz, gallagher, cywu, ztx_0727, huangjiajie}@bupt.edu.cn
pyuan_zhao126@126.com, 1575816307@qq.com

***Abstract*—Marine animal segmentation (MAS) in real underwater environments remains challenging due to optical degradation (scattering, attenuation, blur) and target–background ambiguity caused by camouflage and low contrast. We repurpose the SAM2 image encoder as a strong generic backbone, but its geometry priors alone can struggle to disambiguate category-level semantics and boundaries under severe degradations. We propose SAM2-MAS, a boundary-aware frequency-adaptive framework that couples SAM2 with DINOv2 through residual semantic fusion, enhances encoder features via a gated band-pass feature enhancement (GBFE) module in the Fourier domain, and refines ambiguous edges with an auxiliary boundary head and boundary-driven contrastive learning. Experiments on MAS3K and RMAS demonstrate state-of-the-art performance. In particular, SAM2-MAS achieves 0.864 mIoU on MAS3K and 0.791 mIoU on RMAS, yielding relative improvements of 1.3% and 2.2% over the previous best results, while reducing MAE to 0.013 and 0.018, respectively.**

***Index Terms*—marine animal segmentation, foundation models, SAM2, DINOv2, frequency learning, boundary supervision, contrastive learning**

I. Introduction

Marine animal segmentation (MAS) is central to ecological monitoring and autonomous underwater operations, yet real-world imagery suffers from scattering, attenuation, blur, and low contrast [1]–[3]. In typical MAS datasets (e.g., MAS3K and RMAS), variable turbidity and illumination can reduce local contrast and blur object–background transitions, making boundaries ambiguous and dense prediction brittle under distribution shifts.

Foundation models such as SAM and SAM2 provide strong geometry priors [4], [5], but they may lack category-level semantics that are critical when target and background textures look similar. Self-supervised vision transformers such as DINOv2 offer complementary semantic cues [6], and recent surveys highlight the growing role of foundation models across vision and other domains [7]. Meanwhile, most MAS adaptations focus on spatial tuning while overlooking frequency-domain characteristics that govern underwater degradations.

From a signal-processing view, underwater scattering and defocus attenuate mid/high-frequency components and distort spectral statistics, which directly affects boundary localization and fine-structure recovery. This motivates integrating learnable frequency adaptation into the segmentation pipeline rather than treating enhancement as a fixed pre-processing step [9]. In addition, since boundary regions often concentrate errors under blur and low contrast, boundary-aware objectives can help reduce leakage and improve boundary-sensitive metrics; similar ambiguity has also been studied in camouflage-related settings [11]–[13]. Finally, SAM2's geometry priors are effective for delineating shapes once targets are identifiable, but disambiguating camouflaged animals often requires stronger semantic cues. We therefore combine SAM2 with DINOv2 features under a dual-resolution setting, and design a residual fusion mechanism that injects semantics while protecting SAM2's boundary detail [14].

We propose SAM2-MAS, a boundary-aware frequency-adaptive framework that (i) performs gated band-pass feature enhancement (GBFE) in the Fourier domain, (ii) couples SAM2 with DINOv2 through deep residual fusion, and (iii) refines ambiguous edges with an auxiliary boundary head and boundary-driven contrastive learning. On MAS3K and RMAS, SAM2-MAS achieves state-of-the-art results, reaching 0.864 mIoU on MAS3K and 0.791 mIoU on RMAS with consistently improved boundary-sensitive metrics. Our contributions are:

- A dual-foundation residual fusion design that injects DINOv2 semantics while preserving SAM2 boundary detail.
- A gated band-pass feature enhancement (GBFE) module with content-adaptive filter mixing.
- Boundary-aware training with an auxiliary boundary head and boundary-driven mixed contrastive learning.

II. Related Work

A. Marine Animal Segmentation

Early MAS systems mainly relied on CNN-based encoder–decoder designs and multi-scale feature aggregation to cope with scale variations and cluttered backgrounds [1], [2]. With the rise of attention mechanisms and vision transformers [15], [16], transformer-based architectures further improved long-range modeling and global context reasoning for dense prediction [17]–[20]. Recent underwater benchmarks and instance-level studies further push MAS toward larger-scale evaluations

*Corresponding author

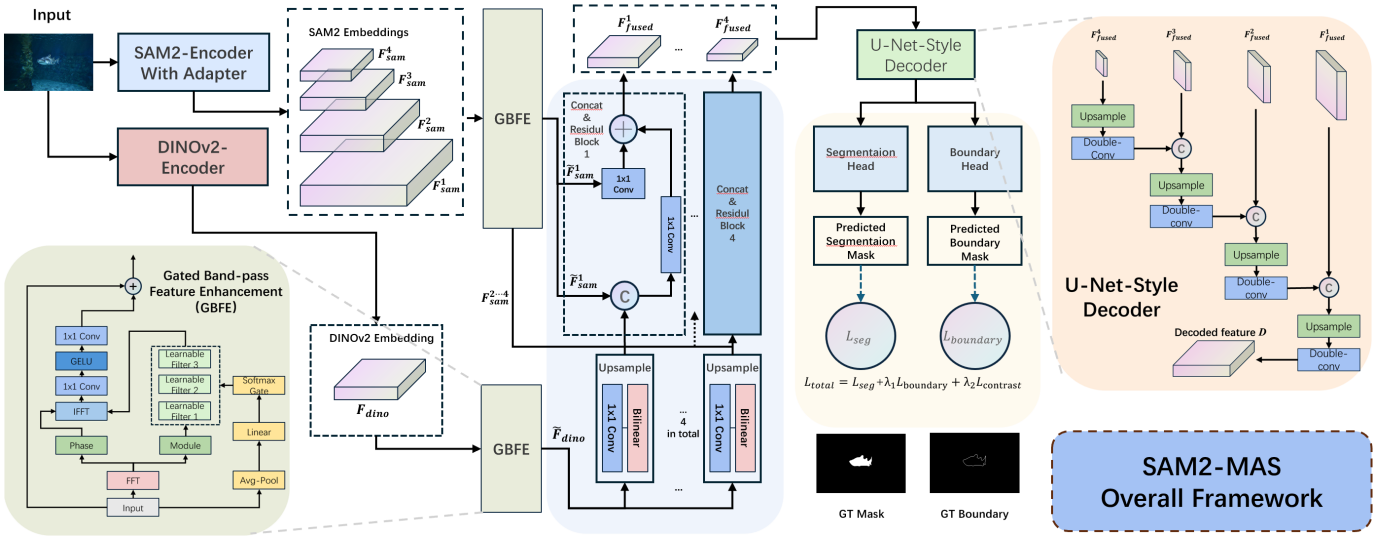

Fig. 1. SAM2-MAS overall framework.

[3], [21], [22]. Despite progress, both CNN and transformer segmenters often require careful domain-specific training and still struggle under severe underwater degradations and camouflage, which motivates leveraging stronger priors from foundation models.

B. Foundation Models and Adaptation

SAM [4] and SAM2 [5] provide strong segmentation priors and robust visual representations, while self-supervised vision transformers such as DINOv2 [6] excel at learning transferable semantics. Surveys have highlighted the rapid expansion of foundation models across tasks and domains [7], alongside continuous advances in general-purpose visual backbones [23]. Recent studies explore adapting SAM-like models to specific domains via parameter-efficient tuning, adapter layers, prompt learning, preference optimization, or feature matching [24]–[28]. In MAS, methods such as MAS-SAM [29] demonstrate that foundation models can be effective but typically rely on single-model adaptation or straightforward feature concatenation. SAM2-UNeXT [30] further combines SAM2 and DINOv2, yet its fusion design does not explicitly preserve SAM2's geometry-sensitive representations, which can weaken SAM2's boundary localization ability. In contrast, we emphasize *dual-foundation synergy* between geometry and semantics with a resize–concat–residual reinjection mechanism that explicitly protects SAM2 features while injecting DINOv2 semantics for improved discrimination under camouflage and low contrast.

C. Frequency Learning in Vision

Frequency-domain analysis has a long history in image restoration and enhancement, and is particularly relevant for degradations caused by scattering and blur. Recent learning-based approaches incorporate Fourier transforms, frequency selection, or learnable filters, suggesting that frequency cues provide complementary information to spatial features; frequency-domain restoration has also been used in low-light enhancement [8]. Frequency-aware designs have also been explored in data-scarce segmentation settings [9], yet frequency-aware adaptation remains under-explored in MAS, where underwater conditions can alter frequency characteristics and weaken boundary cues. We bridge this gap by introducing a gated mixture of band-pass filters that is *learnable* and *content-adaptive* within an end-to-end segmentation framework.

III. Method

A. Overall Architecture

Fig. 1 presents the overall pipeline. Given an underwater image $I \in \mathbb{R}^{H\times W\times 3}$, a SAM2 image encoder (input size 1024×1024) augmented with lightweight adapters extracts four-resolution features $\{F^i_{\text{sam}}\}_{i=1}^4$. Then, we remove the prompt/memory components and freeze the SAM2 trunk, and wrap each transformer block with a small bottleneck MLP adapter that injects domain-specific cues while keeping most parameters fixed. In parallel, DINOv2 extracts a semantic feature F_{dino} from a lower-resolution input (448×448) to reduce computation; the DINOv2 encoder is also kept frozen. We apply the gated band-pass feature enhancement (GBFE) module to obtain frequency-enhanced features $\{\tilde{F}^i_{\text{sam}}\}_{i=1}^4$ and $\tilde{F}_{\text{dino}}$. Next, we inject $\tilde{F}_{\text{dino}}$ into each resolution and fuse it with $\tilde{F}^i_{\text{sam}}$ via deep residual fusion, producing fused features $\{F^i_{\text{fused}}\}_{i=1}^4$. A U-Net-style decoder takes $\{F^i_{\text{fused}}\}_{i=1}^4$ and outputs a decoded feature map D, from which the segmentation head predicts the mask $\hat{Y}$ and the boundary head predicts boundary maps $\hat{B}$.

Here, i indexes the four resolutions from shallow to deep. Each adapter is a two-layer MLP with a bottleneck dimension 32: $\text{Adapter}(x) = \text{GELU}(W_2\,\text{GELU}(W_1 x))$, where $W_1 \in \mathbb{R}^{\text{dim}\times 32}$ and $W_2 \in \mathbb{R}^{32\times\text{dim}}$. The decoder uses bilinear interpolation for upsampling and standard Double-Conv blocks composed of two consecutive conv layers, each followed by

batch normalization and ReLU. This layout keeps the foundation encoders intact and concentrates learnable parameters in the adapters, the fusion layers, the decoder, and lightweight auxiliary heads.

B. Gated Band-Pass Feature Enhancement

Underwater degradations often distort frequency distributions and suppress high-frequency details. We introduce the gated band-pass feature enhancement (GBFE) module that enhances a feature map $x \in \mathbb{R}^{C\times H\times W}$. It is applied once to each final-scale encoder output from SAM2 and DINOv2 before residual fusion. Let $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the 2D Fourier transform and its inverse, and let $\mathrm{Shift}(\cdot)$ be the standard frequency-center shift. We compute a shifted spectrum $X = \mathrm{Shift}(\mathcal{F}(x))$, and decompose it into log-magnitude and phase: $A = \log(1 + |X|)$ and $\Phi = \angle X$.

We adopt the log-magnitude representation to stabilize the dynamic range of the spectrum and to make band selection less sensitive to absolute activation scales. The enhancement is performed through magnitude modulation while preserving phase, since phase often carries structural information that is important for spatial alignment. The radial masks are defined on the frequency plane and are shared across channels, which keeps the module lightweight and encourages coherent frequency emphasis for the entire feature map.

We construct a normalized radius map $\rho(u, v) \in [0, 0.5]$ centered at the frequency origin and define K Gaussian radial masks with learnable centers and bandwidths:

$$G_k(\rho) = \exp\left(-\frac{1}{2}\left(\frac{\rho - \mu_k}{\sigma_k}\right)^2\right), \qquad \mu_k = \mathrm{Softplus}(\tilde{\mu}_k), \quad \sigma_k = \mathrm{Softplus}(\tilde{\sigma}_k) + \epsilon. \tag{1}$$

A lightweight gating network predicts sample-wise mixture weights from the input content:

$$\boldsymbol{\alpha} = \mathrm{Softmax}\big(\mathrm{Linear}(\mathrm{GAP}(x))\big), \qquad \sum_{k=1}^{K} \alpha_k = 1. \tag{2}$$

The mixed mask is $M = \sum_{k=1}^{K} \alpha_k G_k(\rho)$, and we apply a depthwise 1×1 refinement $\hat{M} = M + \mathrm{DWConv}_{1\times1}(M)$. We then enhance the spectrum by modulating the log-magnitude:

$$\hat{A} = \mathrm{Softplus}(A) \odot \hat{M}, \quad |\hat{X}| = \exp(\hat{A}) - 1, \qquad \hat{x} = \mathrm{Re}\Big(\mathcal{F}^{-1}\big(\mathrm{Shift}^{-1}(|\hat{X}| \cdot e^{j\Phi})\big)\Big). \tag{3}$$

Finally, a lightweight spatial MLP produces a residual update $y = \mathrm{MLP}(\hat{x})$, and the output is $\mathrm{LN}(x + \gamma y)$ with a learnable scalar gate γ initialized to a small value. This design enables content-adaptive enhancement while remaining close to identity at initialization. By predicting mixture weights conditioned on the input, GBFE can adaptively emphasize frequency bands that are informative under different underwater conditions, providing complementary cues for boundary localization.

In practice, the learnable centers $\{\mu_k\}$ and bandwidths $\{\sigma_k\}$ allow the filters to cover low, mid, and high-frequency regions, while the simplex-constrained weights $\boldsymbol{\alpha}$ choose an effective combination for each sample. The depthwise refinement adds a small local adjustment to the radial mask, which can compensate for discretization artifacts on the frequency grid. The final layer normalization keeps the enhanced feature distribution consistent with subsequent fusion and decoding modules.

C. SAM2–DINOv2 Deep Residual Fusion

SAM2 provides strong geometry-driven segmentation, while DINOv2 provides robust semantic abstractions. We fuse them carefully to avoid diluting SAM2's precise boundary features. For each resolution i, we align the DINOv2 feature to the SAM2 feature in spatial resolution and channel dimension. Let $\tilde{F}^i_{\text{sam}} \in \mathbb{R}^{h_i\times w_i\times c_s}$ and $\tilde{F}_{\text{dino}} \in \mathbb{R}^{h'\times w'\times c_d}$. We resize $\tilde{F}_{\text{dino}}$ to (h_i, w_i) and apply a 1×1 convolution projection:

$$\bar{F}^i_{\text{dino}} = \mathrm{Conv}_{1\times1}\big(\mathrm{Resize}(\tilde{F}_{\text{dino}})\big). \tag{4}$$

We then concatenate and fuse the features, followed by residual reinjection:

$$F^i_{\text{fused}} = \mathrm{Conv}\Big(\mathrm{Cat}\big(\tilde{F}^i_{\text{sam}}, \bar{F}^i_{\text{dino}}\big)\Big) + \tilde{F}^i_{\text{sam}}. \tag{5}$$

Here, $\mathrm{Conv}(\cdot)$ denotes a lightweight 1×1 convolution that mixes channels after concatenation and produces a residual update in the SAM2 channel space. This formulation treats semantic injection as a refinement branch, while the skip connection guarantees that the fused representation can fall back to the original SAM2 features when semantic cues are weak or noisy. The residual term $+\tilde{F}^i_{\text{sam}}$ preserves SAM2's fine boundary details and prevents coarse semantic cues from overwhelming geometry-sensitive representations, leading to more reliable segmentation in camouflage and low-contrast regions. This structure keeps an identity path for SAM2 features while injecting semantics through the fused branch, which improves compatibility between geometry priors and category-level cues.

D. Boundary-Aware Decoder and Training Objective

We attach a boundary head to the decoder to predict boundary probability maps $\hat{B}$ alongside segmentation masks $\hat{Y}$. The total objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}}(\hat{Y}, Y) + \lambda_1 \mathcal{L}_{\text{boundary}}(\hat{B}, B) + \lambda_2 \mathcal{L}_{\text{contrast}}. \tag{6}$$

Here, $\mathcal{L}_{\text{seg}}$ is the pixel-wise binary cross-entropy loss for segmentation, and $\mathcal{L}_{\text{boundary}}$ is also implemented as a pixel-wise binary cross-entropy loss computed between predicted boundaries and target boundaries extracted from ground-truth masks.

The boundary head is a lightweight layer trained jointly with the segmentation head to provide explicit supervision for transition regions. For the contrastive term, we use a projection head to map features to an embedding space with normalized embeddings.

To refine ambiguous boundaries, we introduce a boundary-driven contrastive loss. Let z_p denote the decoder feature at pixel p derived from D via a projection head. Given a

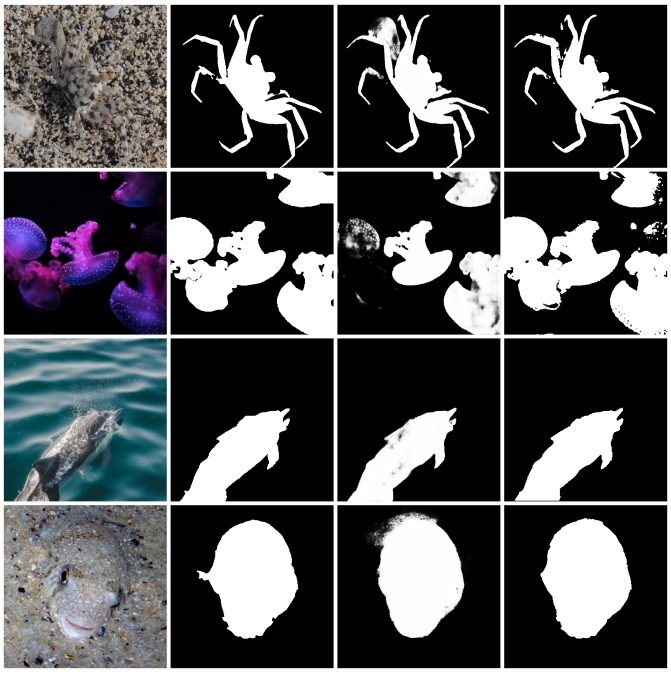

Fig. 2. Qualitative comparisons on MAS3K. We select examples with challenging boundary delineation.

ground-truth mask Y, we construct a thin boundary band B via morphological operations with radius $r = 1$: $B = \text{Dilate}(Y, 1) \setminus \text{Erode}(Y, 1)$. We also tested Sobel and Laplacian operators to obtain boundary targets; when using the same 1-pixel band width, the impact on mIoU is below 10^{-3}. We sample anchors $\mathcal{A}$ from B, positives from an inner band $\mathcal{P}(a) \subseteq \text{Erode}(Y, 1)$, and hard negatives from an outer band $\mathcal{N}(a) \subseteq \text{Dilate}(Y, 1) \setminus Y$, which encourages separation between boundary-adjacent foreground and background. If the expanded outer bands overlap, we keep only one negative sample for that region; if an outer band falls into the foreground interior due to thin structures, we instead replace it by randomly sampling negatives from the background to maintain valid supervision. For each anchor $a \in \mathcal{A}$, we optimize an InfoNCE-style objective:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \log \frac{\sum_{p \in \mathcal{P}(a)} \exp(\text{sim}(z_a, z_p)/\tau)}{\sum_{u \in \mathcal{P}(a) \cup \mathcal{N}(a)} \exp(\text{sim}(z_a, z_u)/\tau)}, \tag{7}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity and τ is a temperature (set to 0.07). This objective encourages boundary features to be closer to object interior while being far from background, which improves boundary alignment and reduces leakage in low-contrast transitions.

IV. Experiments

A. Experimental Setup

Datasets and evaluation metrics. We evaluate our method on two public MAS benchmarks: MAS3K [10] and RMAS [2]. Following common practice in prior work [2], [10], [29], we use the default splits: MAS3K contains 3,103 marine images (including 193 background images), with 1,769 images for training and 1,141 images for testing; RMAS consists of 3,014 marine animal images with 2,514 images for training and 500 images for testing.

Metrics. We report five metrics: mean intersection-over-union (mIoU), structure measure (S_α), weighted F-measure (F_β^w), mean enhanced-alignment measure (mE_ϕ), and mean absolute error (MAE). Higher is better for mIoU, S_α, F_β^w, and mE_ϕ, while lower is better for MAE.

Implementation details. Our method is implemented in PyTorch and trained on an NVIDIA RTX A6000 GPU. We use the AdamW optimizer with an initial learning rate of 2×10^{-4} and cosine learning-rate decay. We train all models for 20 epochs with a batch size of 1. Data augmentation includes random horizontal and vertical flipping. Unless otherwise specified, we use the large versions of SAM2 and DINOv2 initialized from publicly available pretrained weights. We remove the prompt encoder, mask decoder, and memory components from SAM2 and reuse its image encoder trunk as a frozen backbone, training only the inserted adapters, the decoder, and auxiliary modules. The input resolutions are set to 1024×1024 for the SAM2 branch and 448×448 for the DINOv2 branch. We set loss weights $\lambda_1 = 0.2$ and $\lambda_2 = 0.05$ for all experiments and use $\tau = 0.07$ across datasets.

B. Comparison with State of the Art

Table I reports quantitative comparisons on MAS3K and RMAS against representative MAS methods, ranging from CNN/transformer segmenters (e.g., ZoomNet and H2Former) to foundation-model baselines and SAM2-based adaptations. SAM2-MAS achieves the best overall performance, with the highest mIoU and the lowest MAE on both datasets; compared with the previous best results, it yields 1.3% and 2.2% relative mIoU improvements on MAS3K and RMAS, respectively. It also consistently improves boundary-sensitive metrics; on RMAS, S_α increases from 0.883 to 0.895 and F_β^w increases from 0.841 to 0.864. Fig. 2 further provides qualitative comparisons against representative baselines, showing that SAM2-MAS produces cleaner masks and sharper boundaries under camouflage and severe degradations.

C. Ablation Study: Fusion Strategy

We conduct the following ablations on RMAS under the same training setting to isolate the contribution of fusion design, boundary-aware objectives, and GBFE. We study fusion strategies on RMAS in Table II. We use a dual-foundation baseline that combines a SAM2 encoder (frozen, with lightweight adapters) and a DINOv2 encoder (frozen) under a dual-resolution setting (input sizes 1024×1024 and 448×448, respectively), aligns DINOv2 cues to four SAM2 resolutions via spatial resizing and 1×1 channel projections, and uses a U-Net-style decoder. On top of this baseline, we compare several ways to inject aligned DINOv2 cues into SAM2 features. Specifically, we compare Concat, Gated

TABLE I
COMPARISON WITH REPRESENTATIVE METHODS ON MAS3K AND RMAS.

Method	MAS3K					RMAS				
	mIoU ↑	S_α ↑	F_β^w ↑	mE_ϕ ↑	MAE ↓	mIoU ↑	S_α ↑	F_β^w ↑	mE_ϕ ↑	MAE ↓
C2FNet [13]	0.717	0.851	0.761	0.894	0.038	0.721	0.858	0.788	0.923	0.026
OCENet [11]	0.667	0.824	0.703	0.868	0.052	0.680	0.836	0.752	0.900	0.030
ZoomNet [12]	0.736	0.862	0.780	0.898	0.032	0.728	0.855	0.795	0.915	0.022
ECDNet [1]	0.711	0.850	0.766	0.901	0.036	0.664	0.823	0.689	0.854	0.036
MASNet [2]	0.742	0.864	0.788	0.906	0.032	0.731	0.862	0.801	0.920	0.024
SETR [17]	0.715	0.855	0.789	0.917	0.030	0.654	0.818	0.747	0.933	0.028
H2Former [20]	0.748	0.865	0.810	0.925	0.028	0.717	0.844	0.799	0.931	0.023
SAM [4]	0.566	0.763	0.656	0.807	0.059	0.445	0.697	0.534	0.790	0.053
SAM-Adapter [24]	0.714	0.847	0.782	0.914	0.033	0.656	0.816	0.752	0.927	0.027
MAS-SAM [29]	0.788	0.887	0.840	0.938	0.025	0.742	0.865	0.819	0.948	0.021
SAM2-UNet [31]	0.799	0.903	0.848	0.943	0.021	0.738	0.874	0.810	0.944	0.022
SAM2-UNeXT [30]	0.853	0.926	0.900	0.960	0.014	0.774	0.883	0.841	0.949	0.019
SAM2-MAS	**0.864**	**0.933**	**0.907**	**0.964**	**0.013**	**0.791**	**0.895**	**0.864**	**0.960**	**0.018**

TABLE II
ABLATION ON FUSION STRATEGY ON RMAS.

Variant	mIoU ↑	S_α ↑	F_β^w ↑	mE_ϕ ↑	MAE ↓
Concat	0.774	0.883	0.841	0.949	0.019
Gated	0.769	0.885	0.848	0.943	0.020
Res	0.781	**0.891**	0.853	**0.956**	**0.018**
Concat-Res	**0.782**	0.889	**0.854**	0.954	0.019

TABLE III
ABLATION ON BOUNDARY-AWARE TRAINING OBJECTIVES ON RMAS.

Variant	mIoU ↑	S_α ↑	F_β^w ↑	mE_ϕ ↑	MAE ↓
No boundary objective	0.782	0.889	0.854	0.954	0.019
+ Boundary supervision	0.785	**0.893**	0.857	0.957	**0.018**
+ Boundary contrastive	0.784	0.891	0.857	0.957	0.019
+ Boundary supervision + contrastive	**0.787**	**0.893**	**0.861**	**0.958**	**0.018**

Fig. 3. Boundary target construction. From left to right: mask, morphological, Sobel, and Laplacian. With the same 1-pixel band width, the impact on mIoU is below 10^{-3}.

TABLE IV
ABLATION ON GBFE ON RMAS.

Variant	mIoU ↑	S_α ↑	F_β^w ↑	mE_ϕ ↑	MAE ↓
w/o GBFE	0.787	0.893	0.861	0.958	0.018
SENet (spatial attention)	0.782	0.890	0.858	0.955	0.019
Single band-pass (no gate)	0.780	0.890	0.864	0.952	0.019
Gated filters ($K = 2$)	0.789	0.894	0.863	0.959	0.018
Gated filters ($K = 3$)	**0.791**	**0.895**	**0.864**	**0.960**	**0.018**

(reweighting both features before concatenation), and two residual variants. Res injects the resized DINOv2 feature into SAM2 through a residual connection after channel alignment. Concat-Res concatenates the aligned SAM2 and DINOv2 features, projects them with a 1×1 convolution to match the SAM2 channel dimension, and then injects the fused representation via a residual connection, achieving the best mIoU of 0.782. Unless otherwise specified, we use Concat-Res as the default fusion strategy in the following ablations. Compared with direct concatenation, residual reinjection maintains an identity path for SAM2 features and makes the fused branch behave as a bounded refinement, which is less likely to overwrite geometry-sensitive cues and yields more stable boundary quality.

D. Ablation Study: Boundary Supervision

We first construct boundary targets as a 1-pixel band derived from the ground-truth masks. We adopt morphological operators (dilation/erosion with $r = 1$) by default and also evaluate Sobel and Laplacian operators; as shown in Fig. 3, these alternatives yield similar boundary maps and their impact on mIoU is below 10^{-3} under the same band width. We then quantify the contribution of boundary supervision and boundary-driven contrastive learning on RMAS in Table III. All variants use aligned dual-foundation features and the default fusion strategy (Concat-Res); starting from the base mIoU of 0.782, enabling boundary supervision and contrastive learning improves the best mIoU to 0.787. This is consistent with the observation that most errors concentrate near object boundaries under blur and low contrast, and explicitly emphasizing boundary regions can reduce foreground leakage and improve boundary-sensitive measures.

E. Ablation Study: GBFE

Table IV evaluates GBFE on RMAS. All variants are built on the best setting from Table III (default fusion with boundary supervision and contrastive learning), and we only change GBFE. Starting from the base mIoU of 0.787, using gated filters improves performance to 0.789 with $K = 2$ and to 0.791 with $K = 3$. We compare removing GBFE entirely, using a single band-pass filter without gating, and the full gated design with different numbers of learnable filters K. Taken together, these ablations indicate complementary benefits from fusion, frequency adaptation, and boundary-aware training.

The gated design consistently outperforms a fixed single band-pass filter, suggesting that adapting the emphasized frequency range to input content is beneficial under diverse underwater degradations. Increasing K yields a modest gain, indicating that a small number of learnable bands is sufficient in practice.

We also compare GBFE with SENet [32], a representative spatial attention mechanism that adaptively recalibrates channel-wise feature responses. Replacing GBFE with SENet yields 0.782 mIoU, which is 0.005 lower than the baseline without GBFE and 0.009 lower than the full GBFE design. This indicates that purely spatial attention struggles to address underwater degradations that fundamentally alter frequency characteristics, whereas GBFE's frequency-domain adaptation directly compensates for scattering and blur effects.

V. Conclusion

We presented a frequency-adaptive and boundary-aware framework for marine animal segmentation. By integrating SAM2's geometry priors with DINOv2's semantic representations through deep residual fusion, our method improves semantic discrimination in underwater scenes. The proposed GBFE module improves robustness under diverse degradations by dynamically emphasizing informative frequency bands. Finally, boundary supervision with mixed contrastive learning refines ambiguous boundaries and improves boundary-sensitive metrics. Future work includes extending dual-foundation synergy to temporal underwater video segmentation and exploring stronger physical priors for frequency modeling.

References

[1] L. Li, B. Dong, E. Rigall, T. Zhou, J. Dong, and G. Chen, "Marine Animal Segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2303–2314, 2021.

[2] Z. Fu, R. Chen, Y. Huang, *et al.*, "MASNet: A Robust Deep Marine Animal Segmentation Network," *IEEE J. Oceanic Eng.*, 2023.

[3] S. Lian, H. Li, R. Cong, *et al.*, "Watermask: Instance Segmentation for Underwater Imagery," in *Proc. ICCV*, 2023, pp. 1305–1315.

[4] A. Kirillov *et al.*, "Segment Anything," in *Proc. ICCV*, 2023.

[5] N. Ravi *et al.*, "SAM 2: Segment Anything in Images and Videos," arXiv preprint arXiv:2408.00714, 2024.

[6] M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," arXiv preprint arXiv:2304.07193, 2023.

[7] M. Awais, M. Naseer, S. Khan, *et al.*, "Foundation Models Defining a New Era in Vision: A Survey and Outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

[8] D. Feijoo, J. C. Benito, A. Garcia, and M. V. Conde, "DarkIR: Robust Low-Light Image Restoration," in *Proc. CVPR*, 2025, pp. 10879–10889.

[9] Y. Bo, Y. Zhu, L. Li, *et al.*, "FAMNet: Frequency-Aware Matching Network for Cross-Domain Few-Shot Medical Image Segmentation," in *Proc. AAAI*, vol. 39, no. 2, 2025, pp. 1889–1897.

[10] L. Li, E. Rigall, J. Dong, and G. Chen, "MAS3K: An Open Dataset for Marine Animal Segmentation," in *Proc. International Symposium on Benchmarking, Measuring and Optimization*, Springer, 2020, pp. 194–212.

[11] J. Liu, J. Zhang, and N. Barnes, "Modeling Aleatoric Uncertainty for Camouflaged Object Detection," in *Proc. WACV*, 2022, pp. 1445–1454.

[12] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom In and Out: A Mixed-Scale Triplet Network for Camouflaged Object Detection," in *Proc. CVPR*, 2022, pp. 2160–2170.

[13] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-Aware Cross-Level Fusion Network for Camouflaged Object Detection," arXiv preprint arXiv:2105.12555, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[15] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention Is All You Need," in *Proc. NeurIPS*, vol. 30, 2017.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021.

[17] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, *et al.*, "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," in *Proc. CVPR*, 2021, pp. 6881–6890.

[18] H. Cao, Y. Wang, J. Chen, *et al.*, "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation," in *Proc. ECCV*, 2022, pp. 205–218.

[19] J. Chen *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, 2021.

[20] A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 42, no. 9, pp. 2763–2775, 2023.

[21] S. Lian, Z. Zhang, H. Li, *et al.*, "Diving into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation and a Large-Scale Dataset," arXiv preprint arXiv:2406.06039, 2024.

[22] H. Li, S. Lian, Z. Li, *et al.*, "UWSAM: Segment Anything Model Guided Underwater Instance Segmentation and a Large-Scale Benchmark Dataset," arXiv preprint arXiv:2505.15581, 2025.

[23] M. Lou, S. Zhang, H.-Y. Zhou, *et al.*, "TransXNet: Learning Both Global and Local Dynamics with a Dual Dynamic Token Mixer for Visual Recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, 2025.

[24] X. Zhang *et al.*, "SAM-Adapter: Adapting Segment Anything in Under-Resourced Domains," arXiv preprint arXiv:2304.09148, 2023.

[25] T. Chen, A. Lu, L. Zhu, *et al.*, "SAM2-Adapter: Evaluating and Adapting Segment Anything 2 in Downstream Tasks: Camouflage, Shadow, Medical Image Segmentation, and More," arXiv preprint arXiv:2408.04579, 2024.

[26] A. Konwer, Z. Yang, E. Bas, *et al.*, "Enhancing SAM with Efficient Prompting and Preference Optimization for Semi-Supervised Medical Image Segmentation," in *Proc. CVPR*, 2025, pp. 20990–21000.

[27] Y. Liu, M. Zhu, H. Li, *et al.*, "Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching," in *Proc. ICLR*, 2024.

[28] J. Wei, X. Zhao, J. Woo, *et al.*, "Mixture-of-Shape-Experts (MoSE): End-to-End Shape Dictionary Framework to Prompt SAM for Generalizable Medical Segmentation," in *Proc. CVPR*, 2025, pp. 6448–6458.

[29] Y. Liu *et al.*, "MAS-SAM: Adapting Segment Anything Model for Marine Animal Segmentation," arXiv preprint, 2024.

[30] X. Xiong, Z. Wu, L. Zhang, L. Lu, M. Li, and G. Li, "SAM2-UNeXT: An Improved High-Resolution Baseline for Adapting Foundation Models to Downstream Segmentation Tasks," arXiv preprint arXiv:2508.03566, 2025.

[31] X. Xiong, Z. Wu, S. Tan, W. Li, F. Tang, Y. Chen, S. Li, J. Ma, and G. Li, "Sam2-Unet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation," *Visual Intelligence*, vol. 4, no. 1, p. 2, 2026.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. CVPR*, 2018, pp. 7132–7141.