MaskGuide: Efficient Distillation for Deployable Lightweight Segmentation in Marine Environments

Xinrui Wu, Ziqiang Zheng, Yiwei Chen, Zeyu Ma, Yang Yang, *Senior Member, IEEE*, and Sai-Kit Yeung, *Senior Member, IEEE*

***Abstract*—The growing demand for efficient image segmentation in marine ecological studies is currently constrained by two key factors: the high computational requirements of models such as the Segment Anything Model (SAM) and the degraded accuracy of lightweight models in underwater environments. To overcome these challenges, we introduce a Three-stage Iterative Optimization (TIO) training framework and a decoupled knowledge distillation strategy, termed MaskGuide, both of which facilitate the development of our optimized Tiny-MSAM model on the edge device. This model significantly improves frame processing speed while maintaining high accuracy, achieving an optimal tradeoff between segmentation precision and inference speed in practical underwater and marine applications. Our research provides a feasible direction for deploying advanced segmentation models in marine and other resource-constrained scenarios. The Tiny-MSAM model, even when trained from scratch, contains only 0.5% of the parameters of the SAM model and 58% of those in MobileSAM. In existing underwater image segmentation benchmarks (*e.g.*, UIIS dataset), it outperforms MobileSAM by a large margin and reaches 99.4% of the performance of the SAM ViT-H variant.**

***Index Terms*—Object segmentation, deep learning for visual perception, marine robotics.**

I. Introduction

MARINE biological scene understanding [1], [2], [3] is essential for advancing marine ecological research [4], resource management [5], and conservation efforts [6]. The rapid development of marine observation technologies has led to an exponential growth in marine biological data, highlighting the urgent need for efficient and accurate segmentation algorithms [7], [8], [9], [10], [11]. The Segment Anything Model (SAM) series [8], [9] has demonstrated remarkable zero-shot segmentation performance and versatility across various vision applications, positioning it as a powerful backbone for marine segmentation tasks. However, its heavyweight image encoder presents a formidable challenge for deployment on resource-constrained edge devices such as underwater robots and autonomous mobile monitoring systems.

Although recent works attempt to extend SAM to the marine domain, Zhang et al. [10] propose Dual-SAM to enhance marine animal segmentation performance. Lian et al. [2] introduced a large-scale underwater salient instance segmentation dataset accompanied by a SAM-guided baseline, showcasing the feasibility of leveraging foundation models for marine scenarios. However, these solutions still rely on the original ViT-H image encoder [8] and thus inherit its computational burden. Moreover, the vanilla SAM experiences substantial accuracy degradation under underwater conditions characterized by severe light attenuation [12], color distortion [13], and complex background clutter [1]. Therefore, developing a lightweight, marine-adapted segmentation model that achieves SAM-level precision while meeting real-time processing requirements remains an unresolved and critical research challenge, as shown in Fig. 1.

Despite the advancements in developing lightweight models such as MobileSAM [14] and FastSAM [15], their application to marine biological segmentation has revealed notable limitations. Although these models demonstrate strong efficiency, they fail to achieve satisfactory accuracy in marine environments. This gap arises primarily from two factors. First, their training data largely comprises daily scenes, offering *limited exposure to the unique visual challenges of underwater imagery*. Second, existing distillation strategies are *not well-suited for handling complex tasks* like scene segmentation under specific and challenging underwater conditions. As a result, these lightweight models exhibit substantially lower segmentation performance [16] in marine settings compared to general scene segmentation tasks.

To address persistent segmentation challenges in marine environments and the limitations of current lightweight models, we propose an efficient feature-level distillation method called MaskGuide that enables more effective knowledge transfer. Our approach enhances teacher features directly within the feature space to improve robustness under underwater conditions, enabling the development of an ultra-lightweight model, Tiny-MSAM, optimized for real-time deployment. This model achieves a 133% increase in FPS compared to MobileSAM during inference while retaining 99.4% of the original SAM's accuracy. The core innovation of MaskGuide lies in its feature-level enhancement paradigm tailored for marine imagery, setting it apart from prior methods through three key aspects:

- **Robust feature extraction** in challenging marine environments. MaskGuide mitigates the effects of underwater degradations (*e.g.*, low contrast, illumination shifts,

Received 23 October 2025; accepted 14 March 2026. Date of publication 13 April 2026; date of current version 17 April 2026. This article was recommended for publication by Associate Editor A. Quattrini Li and Editor G. Loianno upon evaluation of the reviewers' comments. This work was supported in part by Sichuan Science and Technology Program under Grant 2026NSFSC1451, in part by the Sichuan Province Innovative Talent Funding Project for Postdoctoral Fellows under Grant BX202405, and in part by Marine Conservation Enhancement Fund under Grant MCEF20107 and Grant MCEF22112. *(Xinrui Wu and Ziqiang Zheng contributed equally to this work.) (Corresponding author: Zeyu Ma.)*

Xinrui Wu, Zeyu Ma, and Yang Yang are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China (e-mail: xinruiwu.wxr@gmail.com; mazeyu@uestc.edu.cn; yang.yang@uestc.edu.cn).

Ziqiang Zheng, Yiwei Chen, and Sai-Kit Yeung are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong (e-mail: zhengziqiang1@gmail.com; ychenmb@connect.ust.hk; saikit@ust.hk).

Digital Object Identifier 10.1109/LRA.2026.3683334

turbidity) by distilling knowledge directly at the feature level, ensuring stable and reliable representation learning.

- **Feature-level enhancement strategy**. Unlike image-level pre-processing algorithms, our approach enhances representations at the feature level by decomposing teacher features into foreground and background components and reintegrating them through adaptive fusion, which strengthens discriminative information while preserving semantic integrity that are crucial for accurate marine segmentation.
- **Efficient deployment-oriented design**. Leveraging the MaskGuide framework and TIO strategy, we developed Tiny-MSAM, an ultra-lightweight model that achieves an optimal balance between efficiency and accuracy, making it well-suited for deployment in resource-constrained marine robotics applications.

II. Related Works

A. Marine Scene Understanding

Marine and underwater scene understanding focuses on explaining the visual and semantic relationships between objects, environments, and dynamic processes occurring beneath the water's surface [2], [4], [5], [17], [18]. The reliable segmentation [2] faces unique challenges from light scattering, color distortion, and low contrast [12], [13], [16], [19], [20]. Furthermore, marine specimens exhibit substantial variability in shape, color, and texture, while available datasets remain far more limited than aerial counterparts, such as UAVid [21] and MarineInst20M [1]. This scarcity of high-quality, well-annotated marine datasets hinders the training of robust segmentation models. Moreover, synthetic data generation [22] and the direct application of terrestrial models have proven suboptimal due to domain shifts [23]. These challenges emphasize the importance of developing dedicated domain adaptation strategies [24] to better align model learning with the unique characteristics of underwater imagery. To mitigate these limitations, prior studies have employed a combination of enhancement techniques [25], [26] and deep learning approaches [27]. Among them, Semi-UIR [28] represents a promising direction by first applying advanced underwater image enhancement before segmentation, leading to significant improvements in marine organism segmentation accuracy. This line of research underscores the growing recognition that effective underwater scene understanding requires addressing both visual degradation and task-specific adaptation simultaneously.

B. Efficient Segmentation Model Design

The Segment Anything Model (SAM) [8], [9] has revolutionized prompt-based segmentation with exceptional zero-shot generalization, yet its heavy transformer-based encoder makes deployment on edge or resource-constrained devices challenging. This limitation has motivated research into lightweight optimizations [29], leading to variants such as EfficientSAM [30], FastSAM [15], and MobileSAM [14], which reduce parameters and computation but often sacrifice segmentation precision. To better balance this trade-off, knowledge distillation [31] has emerged as an effective approach for transferring semantic and structural knowledge from large teacher models to compact student networks. Parallel to these advances, efficient architectures designed for edge deployment—such as MobileNet [32], ShuffleNet [33], GhostNet [34], and Vision Transformer hybrids [35], have demonstrated that compact models can maintain strong representational power with minimal overhead. Recent solutions like EdgeNeXt [36], FastSAM [15], and PIDNet [37] exemplify this synergy between knowledge distillation and efficient design, advancing the development of real-time, high-performance segmentation models suitable for edge applications. In this work, we aim to propose a MaskGuide training framework that integrate the feature-level augmentation and adaptive feature fusion into the knowledge distillation process for a better tradeoff between segmentation precision and inference speed.

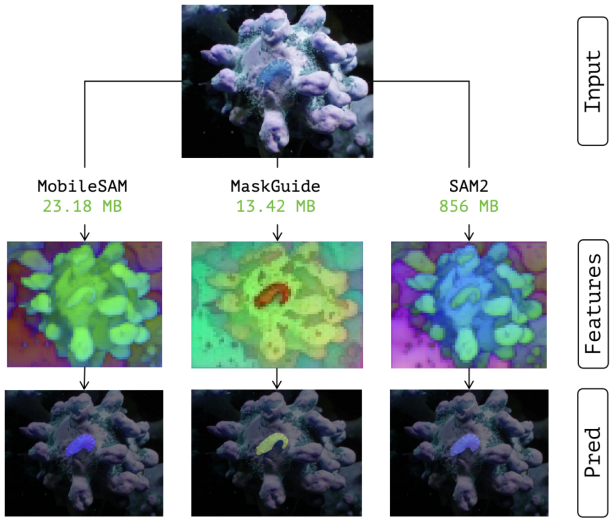


Fig. 1. The comparison of feature maps and model performance among our method, MobileSAM, and SAM 2 shows that our model achieves an IoU accuracy comparable to SAM 2 while utilizing fewer parameters than MobileSAM, making it highly suitable for deployment on edge devices. Furthermore, the feature maps clearly illustrate the superior feature extraction capability of our model.

III. Method

A. MaskGuide: Targeted Feature Enhancement

Conventional knowledge distillation approaches [31] for semantic segmentation often suffer from two critical limitations in complex marine environments: the transfer of excessive background noise and the provision of insufficiently nuanced guidance signals [23]. These issues are particularly pronounced in underwater scenarios characterized by low contrast and cluttered backgrounds, where distinguishing foreground objects from their environment presents a significant challenge [16]. To overcome these fundamental limitations, we introduce **MaskGuide**, a novel feature enhancement methodology that operates on the core principle of feature fusion. As illustrated in Fig. 2, our approach integrates two complementary augmentations with an advanced feature fusion mechanism to enable targeted, effective knowledge transfer in challenging marine environments.

1) Foreground Augmentation: Foreground Feature Purification: To mitigate background noise in complex marine scenes [16], we introduce the **Masked-Augmentation** strategy. This approach implements explicit information gating through element-wise multiplication:

$$\mathbf{I}_{\text{Foreground}} = \mathbf{I} \odot \mathbf{M}, \tag{1}$$

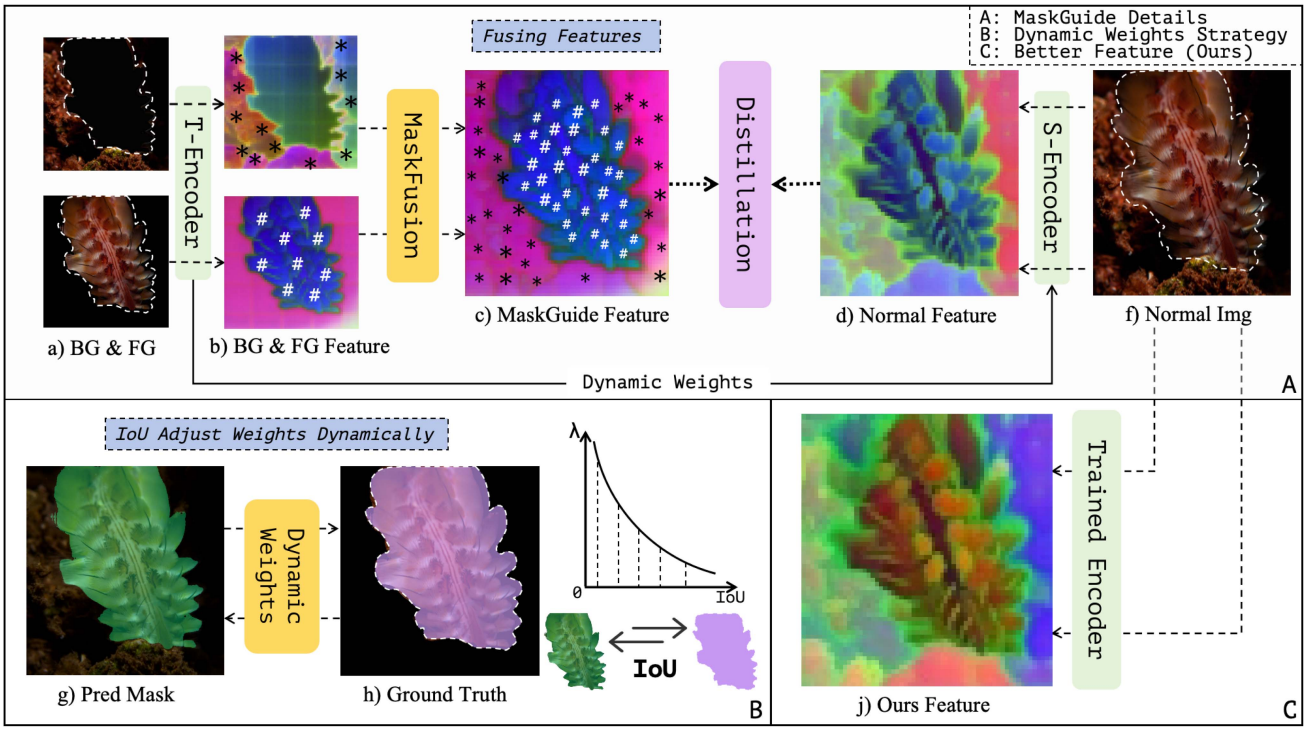


Fig. 2. Detailed schematic of **MaskGuide**. By performing fusion operations between feature maps derived from inverse mask application and those from positive mask application, high-quality feature maps are obtained for knowledge distillation. The notation '#' denotes using element-wise multiplication to preserve the foreground feature map, and the notation '$*$' denotes using element-wise multiplication to preserve the background feature map.

where I denotes the input image feature map, M represents the binary foreground mask generated by the teacher model, and $\odot$ indicates element-wise multiplication. This operation retains target-region features while suppressing the background, guiding the student model to focus on the teacher's intrinsic foreground representations. This "feature focus mode" enhances discriminative learning and prevents the student from capturing spurious background correlations.

2) Background Augmentation: Contextual Awareness Enhancement: While foreground purification is essential, over-reliance on target features can lead to inadequate contextual understanding and imprecise segmentation boundaries [2]. To address this complementary challenge, we introduce the **Background-Augmentation** strategy, defined as:

$$\mathbf{I}_{\text{Background}} = \mathbf{I} \odot (1 - \mathbf{M}), \tag{2}$$

where (1 - M) is the inverted mask used to extract and amplify contextual background features. This operation explores a crucial yet underused aspect of knowledge distillation, highlighting background context by suppressing primary targets. It encourages more balanced representation learning, improving boundary discrimination and contextual reasoning [38], both vital for accurate segmentation in marine environments.

The combination of (1) and (2) forms the core innovation of our MaskGuide methodology. Our core idea is to disentangle the teacher's features into complementary foreground and background components. As shown in Fig. 2, combining these orthogonal augmentations provides a balanced learning signal that enhances both object characterization and contextual understanding, achieving observed performance gains.

3) MaskFusion: Synergistic Feature Integration: The foreground and background augmentations offer complementary perspectives—target-focused and context-aware. Our key insight is that optimal distillation arises from their intelligent fusion rather than isolated use. Since simple operations (*e.g.*, concatenation or addition) can cause feature imbalance, we propose **MaskFusion**, a novel feature integration mechanism and core contribution of this work:

$$\mathbf{F}_{\text{fused}} = \text{Fusion}(\mathbf{F}_{\text{original}}, \mathbf{F}_{\text{foreground}}, \mathbf{F}_{\text{background}}). \tag{3}$$

The primary goal of MaskFusion is to construct a teacher feature that provides higher-quality supervisory signals, rather than being used directly for inference in the student model. To ensure the efficiency of the distillation process, we have adopted the simplest feasible scheme. This design lets MaskFusion balance semantic richness and efficiency, providing precise teacher guidance without additional inference cost. In detail, we perform a weighted element-wise averaging of the three feature maps from the *original*, *foreground*, and *background* branches (shown in (3)). This simple yet effective fusion aggregates multi-perspective information while preserving spatial semantics, ensuring accurate teacher guidance during distillation. Building on this module, we further design a structured training framework to systematically enhance student capability. In summary, our MaskFusion provides a powerful mechanism for feature-level enhancement. To effectively integrate this module into a holistic distillation pipeline, we further propose a structured training framework that systematically builds student capabilities.

As shown in Fig. 3, directly extracting features from raw images in complex environments often leads to background interference, object confusion, and indistinct boundaries. In

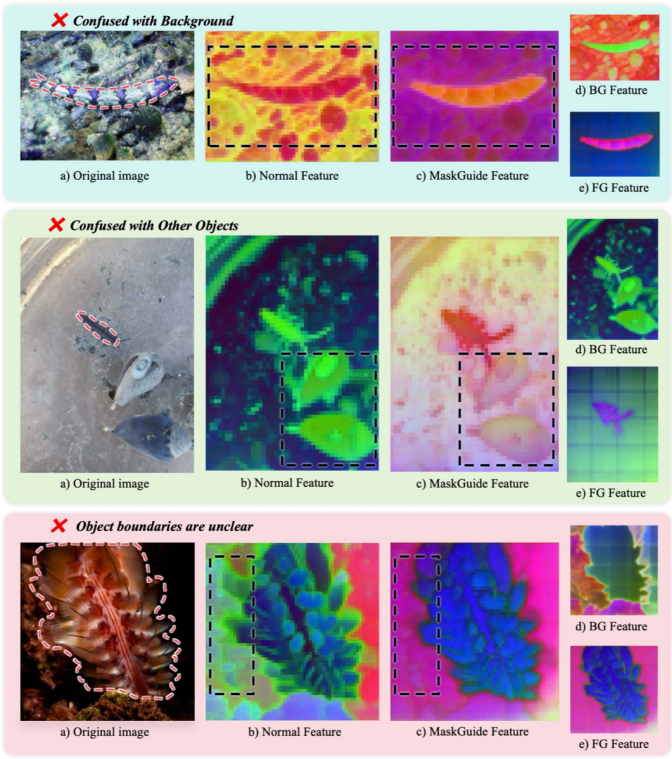


Fig. 3. Comparative analysis of feature maps between MaskGuide and baseline methods. **Blue** regions represent foreground–background confusion, **green** regions denote inter-class feature overlap, and **pink** regions correspond to ambiguous object boundaries. Our method can effectively alleviate these issues.

contrast, our MaskGuide method alleviates these issues by disentangling foreground and background information, producing feature maps that are more target-focused, exhibit lower inter-class overlap, and yield clearer, more discriminative object boundaries.

B. TIO Strategy

Framework overview: We first detail our framework overview. Standard knowledge distillation methods often use static strategies that inadequately transfer complex capabilities from advanced teachers [8], [31]. We propose the **TIO** method, a progressive training regimen with three phases: **Primary Capability Formation**, **Intensive Feature Distillation**, and **Refinement Optimization**. This structured approach mitigates catastrophic forgetting while enabling progressive specialization, based on the philosophy that effective distillation requires a strategically sequenced training curriculum.

1) Stage I: Primary Capability Formation: This stage uses large-scale vision datasets (*e.g.*, COCO 2017 [39]) via feature distillation [36]. The student learns broad visual patterns directly from SAM [8], building foundational representations and avoiding premature overspecialization to marine data.

2) Stage II: Intensive Feature Distillation: In this stage, we conduct **MaskGuide** to overcome insufficient guidance and background noise. The teacher model produces three supervisory streams:

- **Original** branch keeps target-background interactions.
- **Foreground** branch isolates target features.
- **Background** branch preserves context.

Through effective feature fusion, MaskGuide could yield richer features, improving learning for complex marine segmentation.

3) Stage III: Refinement Optimization: The final stage uses an **IoU-guided dynamic weighting** mechanism, emphasizing challenging samples (lower IoU) which contain more learning value [24]:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{distill}}, \quad \lambda = \begin{cases} 10/(1+\text{IoU}), & \text{if IoU} < 0.9 \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

This prioritizes difficult cases, fostering robust segmentation in complex marine scenes [37]. We provide the ablation studies in Section V-C.

IV. Implementation

A. Model Overview

In this section, we present the detailed implementation of our approach. Two models are trained: the marine-domain-tuned MobileSAM and our ultra-lightweight Tiny-MSAM (from SAM). MobileSAM is trained only in Stages 2 and 3, whereas Tiny-MSAM undergoes all three stages. When performing the knowledge distillation, we adopt an adaptive temperature mechanism as described in Section IV-B. Then we detail our proposed TIO training flow in Section IV-C. Finally, we provide the training details and computational cost analysis in Section IV-D and Section IV-E, respectively.

B. Adaptive Temperature Mechanism

An adaptive temperature mechanism was used during the knowledge distillation procedure, varying with training progress $R \in [0, 1]$:

$$T(R) = 0.8 + 3.2 \cdot \frac{1 + \cos(\pi R)}{2}, \quad T \in [0.8, 4.0] \tag{5}$$

The distillation loss combines KL divergence and cosine similarity:

$$\mathcal{L}_{\text{kd}} = \alpha \cdot \text{KL loss} + \beta \cdot \text{Cosine Loss} \tag{6}$$

with $\alpha = 0.8$ and $\beta = 0.2$. The training has two stages:

- **Early stage** ($R < 0.5$): Focuses on feature alignment with coefficients $A = 0.1$ and $B = 0.04$.
- **Late stage** ($R \geq 0.5$): Focuses on attention matching with $A = 0.08$ and $B = 0.02$.

via the two-stage distillation training, we preserve the teacher model's capability while efficiently adapting to the target data distribution.

C. TIO Training Flow

MaskGuide consists of three sequential stages: 1) The first stage establishes **foundational segmentation capabilities** by training the model, initialized with Kaiming initialization, on COCO [39] while freezing the mask decoder, achieving 75.6% and 69.8% mIoU on the marine data. Meanwhile, adding IMC and UIIS further improves IoU to 81.2%and 76.1%, effectively preventing terrestrial overfitting. 2) The second stage employs MaskGuide for intensive **domain-specific feature distillation**, using only marine datasets (UIIS and IMC) to focus the student model on relevant biological features via temperature-scaled

TABLE I
COMPARISON OF DIFFERENT MODELS BASED ON PARAMETER COUNT, MODEL SIZE, AND INFERENCE SPEED (FPS).

Model	Parameters (M) ↓	Weights ↓	FPS ↑
SAM (ViT-H) [8]	637	2.43 GB	4.15
SAM 2 (Large) [9]	224.4	856 MB	11.9
SAM 2 (Tiny) [9]	38.9	149 MB	31.4
FastSAM [15]	68	138 MB	37.2
EfficientSAM [30]	9.8	39.1 MB	112.3
MobileSAM [14]	6	23.18 MB	127.4
Ours	**3.5**	**13.42 MB**	**164.7**

knowledge transfer from the teacher, thereby enhancing feature extraction efficiency. 3) The final stage refines optimization by extending training to the **full segmentation pipeline** and integrating dynamic IoU-guided dynamic weighting, which accelerates convergence and mitigates boundary and internal inconsistencies.

D. Training Details

Our Tiny-MSAM model's IMG-Encoder is derived from MobileSAM's IMG-Encoder, retaining the TinyViT architecture. The structural adjustments include reducing the model depth from [2, 2, 6, 2] to [1, 2, 4, 1], modifying the number of heads from [2, 4, 5, 10] to [2, 3, 4, 8], and updating the embedding dimensions from [64, 128, 160, 320] to [64, 96, 128, 320]. As a result, the Tiny-MSAM IMG-Encoder contains approximately 3.5 million parameters—substantially fewer than MobileSAM's 6 million and the SAM ViT-H model's 637 million parameters.

For finetuning MobileSAM based on our MaskGuide approach, we only employ Stage 3, utilizing SAM ViT-H as the teacher model. Training is conducted for 20 epochs on the combined UIIS and IMC datasets, with sampling rates adjusted according to their data size to ensure balanced distribution. Resource consumption totals 24 GPU hours using 2 A800 GPUs (40 GB memory each). For optimizing our Tiny-MSAM, we detail the training procedures of all three stages as follows:

- Stage 1: 5 epochs on COCO2017 datasets.
- Stage 2: 20 epochs on integrated COCO2017, VOC, IMC, and UIIS datasets.
- Stage 3: 20 epochs as finetuning MobileSAM.

This staged optimization approach ensures progressive capability enhancement through efficient knowledge transfer and optimization.

E. Computational Cost Analysis

We finally report computational cost analysis in Table I. The proposed Tiny-MSAM model demonstrates a favorable Pareto optimality on efficiency–accuracy trade-off curve, significantly outperforming all existing open-source alternatives.

V. EXPERIMENT

A. Experimental Setup

Datasets: Models are trained and evaluated on three datasets: COCO [39], UIIS [40], and a self-collected IMC dataset, which was constructed by the biologist team, including 18,948 images from 1,000 marine species. It features 37,804 taxonomically labeled instance masks and 41,944 non-taxonomic salient masks for boundary delineation.

Evaluation settings: We designed a series of rigorous experiments to verify the robustness and effectiveness of our MaskGuide over existing methods. Our performance evaluation is based on the following established metrics:

- **1-P (single-point accuracy).** This metric evaluates model robustness with minimal visual cues, using a single randomly sampled pixel as the prompt. The score is the mean IoU over all test instances.
- **5-P (5-point accuracy).** It extends 1-P to evaluate stability under richer cues by providing five randomly sampled foreground pixels as prompts.
- **BBox (bounding-box accuracy).** It serves as an upper-bound in which the model is given ground-truth bounding boxes (a fully known spatial context) for all objects.
- **Avg. (average accuracy)** is the arithmetic mean of the mIoU scores from the 1-P, 5-P, and BBox settings. It provides a single composite score summarizing the model's performance across different prompt granularities.

To ensure a rigorous and fair comparison, the identical sets of point coordinates (for 1-P/5-P) are used across all evaluated models on a given dataset. This is controlled by fixing the random seed during point generation, eliminating variance from the evaluation process.

B. Comparison With SOTAs

We present a detailed comparison with existing algorithms (SAM and SAM 2) as well as specifically designed knowledge distillation methods. The quantitative results of three datasets are provided in Table II, providing compelling evidence for the efficacy of our proposed method. On the IMC benchmark, MobileSAM's IoU performance (BBox setting) improved by 1.5% after Stage 2 and reached a total gain of 2.5% after Stage 3. Our Tiny-MSAM model, despite using only 58% of MobileSAM's parameters, achieved 99.4% accuracy on the UIIS benchmark, closely approaching the performance of the much larger SAM model. On the UIIS dataset, MobileSAM's IoU performance increased by 2.5% after Stage 2, surpassing its teacher model, SAM (ViT-H), and achieving a total improvement of 3% after Stage 3. Tiny-MSAM further outperformed MobileSAM while maintaining a compact architecture, highlighting its strong efficiency and adaptability to complex marine scenes. Please note that Tiny-MSAM was fully optimized from scratch, which can reveal the lower bound of the proposed method.

Qualitative comparisons of different algorithms under various underwater scenarios: As illustrated in Fig 4, our method demonstrates consistently superior segmentation quality across a variety of challenging underwater scenarios, including *severe light attenuation*, *low contrast*, and *cluttered backgrounds*. The qualitative comparisons highlight that the model trained with the MaskGuide framework produces more precise and reliable full-scenario segmentation outputs than MobileSAM, while achieving competitive, or even better performance compared with the much larger SAM model. These observations validate the effectiveness of the proposed feature disentanglement and fusion strategies in enhancing segmentation robustness for complex marine imagery.

Furthermore, we provide the convergence analysis of our proposed MaskGuide in Fig. 5. As demonstrated, MaskGuide

TABLE II
THE PERFORMANCE COMPARISON OF DIFFERENT MODELS ON UIIS, IMC, AND COCO2017 BENCHMARKS. RESULTS IN **LIGHT PURPLE** INDICATE THE RESULTS OF OUR APPROACH.

Method	UIIS Val				IMC Val				COCO Val
	1-P	5-P	BBox	Avg.	1-P	5-P	BBox	Avg.	BBox
Tiny-MSAM (stage 3)	17.63	21.56	83.51	*40.9*	70.26	85.26	89.72	*81.7*	76.97
MobileSAM (ViT-Tiny) [14]	15.54	20.65	82.16	*39.5*	71.96	87.31	90.40	*83.2*	77.84
$\text{Ours}_{\text{MobileSAM}}$ (stage 2)	19.57	23.84	84.61	*42.7*	79.06	88.43	91.67	*86.4*	78.88
$\text{Ours}_{\text{MobileSAM}}$ (stage 3)	**21.03**	**24.06**	**85.17**	***43.4***	**79.29**	**89.94**	**92.48**	***87.2***	**79.13**
SAM (ViT-B) [8]	11.73	17.33	81.32	*36.8*	66.32	89.56	90.91	*82.3*	77.41
SAM (ViT-H) [8]	13.83	21.72	84.03	*39.8*	72.70	92.66	94.06	*86.5*	80.43
SAM 2 (Large) [9]	16.92	22.09	84.00	*41.0*	72.33	91.30	93.13	*85.6*	**80.82**
SAM 2 (Tiny) [9]	12.41	17.59	81.93	*37.3*	71.45	90.93	92.02	*84.8*	78.18
Ours_{SAM} (stage 3)	**21.82**	**24.21**	**86.73**	***44.3***	**78.52**	**93.84**	**94.92**	***89.0***	80.55

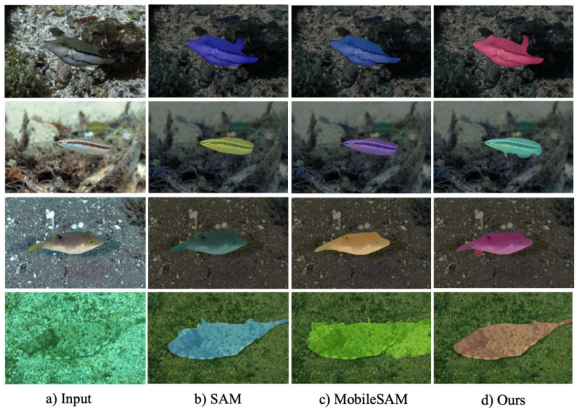
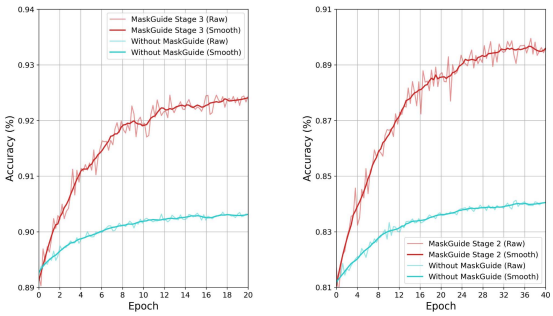


Fig. 4. Qualitative comparisons of different algorithms under various underwater scenarios.



MobileSAM (Stage 3) Tiny-MSAM (Stage 2)

Fig. 5. The IoU convergence curves on the validation set demonstrate that the training process using the MaskGuide achieves significantly improved convergence efficiency.

demonstrates more stable convergence and superior final performance, directly contributing to the IoU accuracy improvements as reported in Table II. This consistent training behavior highlights the effectiveness of MaskGuide in facilitating balanced knowledge transfer and improved model generalization across marine scenes.

C. Ablation Studies

In this section, we first provide the detailed ablation studies regarding our **TIO training strategy**. We conducted a dedicated ablation study comparing the full TIO pipeline against three single-stage baselines trained on: (1) COCO only, (2) UIIS + IMC only, and (3) a mixture of the three datasets (COCO + UIIS + IMC). All models were trained for the same total number of epochs to ensure a fair comparison. We provide the quantitative results in Table III, which clearly demonstrate the superiority of our three-stage optimization design. Specifically, the single-stage baselines suffer from lower performance across all key metrics (1-P, 5-P, BBox, and Avg.) on both UIIS and IMC validation sets. This performance gap validates that a single-stage training scheme fails to resolve the inherent conflicts between learning general representations, domain adaptation, and task-specific fine-tuning. Our proposed TIO strategy effectively decouples and sequentially optimizes these objectives, leading to a more robust and accurate segmentation model.

Furthermore, we conduct ablation studies regarding different **feature fusion strategies**. We compared the impact of different branch combinations (*e.g.*, using only "original + foreground", "original + background", or "foreground + background" pairs) on the supervisory effectiveness of the teacher model. The experimental results indicate that the teacher model's performance decreases when any two of the three branches are used, compared to the full three-branch fusion scheme. This finding underscores the importance of integrating complementary information from all three branches to achieve optimal feature representation and supervision quality. We further compared the effectiveness of our MaskFusion module against simpler fusion strategies (*e.g.*, unweighted element-wise addition and element-wise multiplication). All the experimental results are reported in Table IV. The results clearly demonstrate that the proposed weighted element-wise averaging scheme attains the optimal balance (on UIIS validation dataset, our MaskFusion method achieved **a leading average score of 46.2%**, outperforming alternative fusion strategies such as element-wise addition (43.8%) and "original + foreground" (45.9%)) between performance and efficiency.

Finally, we evaluate the effectiveness of proposed IoU-guided dynamic weighting in Table V. Our formulation systematically assigns higher weights to low-IoU samples, concentrating optimization efforts on the model's weak areas while reducing the influence of well-handled cases to prevent gradient dominance. As demonstrated in Table V, this adaptive focusing accelerates convergence by delivering stronger learning signals where most needed, thereby systematically enhancing final model robustness and segmentation quality.

D. Further Discussions

Besides the detailed ablation studies, we further explored the comparison with the enhancement-then-segmentation pipeline to address the specific challenge of underwater images. We incorporate the underwater image enhancement algorithm [28]

TABLE III
ABLATION STUDIES REGARDING THE EFFECTIVENESS OF OUR PROPOSED TIO TRAINING STRATEGY.

Training strategy	UIIS Val				IMC Val			
	1-P	5-P	BBox	Avg.	1-P	5-P	BBox	Avg.
1-stage (COCO only)	13.74	17.76	76.79	*36.1*	65.37	81.11	85.92	*77.5*
1-stage (UIIS + IMC)	14.18	18.01	77.99	*36.7*	64.13	80.88	85.32	*76.8*
1-stage (All three datasets)	16.02	19.87	82.43	*39.4*	68.61	83.67	88.15	*80.1*
TIO (Ours)	**17.63**	**21.56**	**83.51**	***40.9***	**70.26**	**85.26**	**89.72**	***81.7***

TABLE IV
PERFORMANCE COMPARISON OF FEATURE FUSION STRATEGIES USING MOBILESAM ON UIIS AND IMC DATASETS.

Training strategy	UIIS Val				IMC Val			
	1-P	5-P	BBox	Avg.	1-P	5-P	BBox	Avg.
Original + Foreground	22.00	**24.58**	91.12	*45.9*	81.19	93.01	96.13	*90.1*
Original + Background	21.91	24.13	90.88	*45.6*	80.92	92.66	96.01	*89.9*
Foreground + Background	21.95	24.01	89.99	*45.3*	80.31	92.85	94.99	*89.4*
Element-wise Multiplication	14.67	18.11	79.30	*37.4*	70.17	85.68	88.71	*81.5*
Element-wise Add	20.83	22.67	88.01	*43.8*	80.07	90.46	94.14	*88.2*
Ours	**22.13**	24.37	**92.23**	***46.2***	**82.12**	**94.87**	**96.75**	***91.3***

TABLE V
IOU COMPARISON OF MODELS UNDER DIFFERENT SETTINGS ON UIIS, IMC, AND COCO DATASETS. WE DEMONSTRATE THE PERFORMANCE IMPROVEMENTS OF THE PROPOSED IOU-GUIDED DYNAMIC WEIGHTING.

Method	UIIS Val				IMC Val				COCO2017 Val
	1-P	5-P	BBox	Avg.	1-P	5-P	BBox	Avg.	BBox
MobileSAM (ViT-Tiny) [14]	15.54	20.65	82.16	*39.5*	71.96	87.31	90.40	*83.2*	77.84
Ours$_{\text{MobileSAM}}$	**22.13**	**24.37**	**92.23**	***46.2***	**82.12**	**94.87**	**96.75**	***91.3***	**86.74**
SAM (ViT-H) [8]	13.83	21.71	84.00	*39.8*	72.70	92.66	94.06	*86.5*	80.40
Ours$_{\text{SAM}}$	**24.02**	**26.42**	**94.37**	***48.3***	**84.73**	**96.59**	**98.82**	***93.4***	**89.29**

TABLE VI
PERFORMANCE COMPARISON UNDER DIFFERENT SETTINGS. "W/ SEMI-UIR" INDICATES THAT WE USE SEMI-UIR [28] FOR THE UNDERWATER IMAGE ENHANCEMENT BEFORE PERFORMING THE SEGMENTATION.

Backbone	Method	UIIS Val				IMC Val			
		1-P	5-P	BBox	Avg.	1-P	5-P	BBox	Avg.
MobileSAM [14]	Baseline	15.54	20.65	82.16	*39.5*	71.96	87.31	90.40	*83.2*
	w/ Semi-UIR	15.31	20.89	82.49	*39.6*	73.81	88.31	91.25	*84.5*
	Ours	**21.03**	**24.06**	**85.17**	***43.4***	**79.29**	**89.94**	**92.48**	***87.2***
SAM [8]	Baseline	13.83	21.72	84.00	*39.8*	72.70	92.66	94.06	*86.5*
	w/ Semi-UIR	16.31	22.09	83.89	*40.8*	72.81	92.91	94.37	*86.7*
	Ours	**21.82**	**24.21**	**86.73**	***44.3***	**78.50**	**93.81**	**94.94**	***89.0***

for image restoration first, followed by segmentation models such as SAM and MobileSAM. The experiments were conducted across two datasets (UIIS and IMC) to assess performance under varying image quality conditions. As reported in Table VI, our method consistently surpasses both the baseline model and Semi-UIR across all metrics on both UIIS and IMC validation sets, regardless of the backbone (MobileSAM or SAM). Notably, on the more challenging IMC Val with MobileSAM backbone, our method outperforms Semi-UIR by a clear margin **(Avg. score: 87.2% vs. 84.5%)**. We attribute the performance improvements to that knowledge distillation enables direct transfer of semantic and structural priors from a powerful teacher model, allowing the student network to learn robust representations invariant to underwater degradations, whereas enhancement-then-segmentation approaches may amplify artifacts and disrupt downstream feature consistency.

VI. CONCLUSION

In this work, we propose MaskGuide, which advances marine biological image segmentation by balancing accuracy, efficiency, and deployability. The core innovation is to introduce a feature-level enhancement paradigm that enables robust feature extraction under visual degradations and efficient fusion through foreground–background decomposition. Combined with the Three-stage Iterative Optimization (TIO) scheme, it enhances model learning stability and capability. The ultra-lightweight Tiny-MSAM achieves 99.4% of SAM's accuracy with only 3.5 M parameters, 13.42 MB of weights, and 164.7 FPS, effectively bridging the gap between high accuracy and real-time on-device processing for marine robotics. Despite these advances, staged training requires careful tuning, and performance may degrade under extreme underwater conditions. Future work will focus on improving robustness, adaptive training, and further model compression to enhance efficiency without sacrificing accuracy.

ACKNOWLEDGMENT

The authors would also like to express their sincere gratitude to the "Sustainable Smart Campus as a Living Lab" (SSC) program at HKUST for its vital support. This work was conducted during Xinrui Wu internship at HKUST.

REFERENCES

[1] Z. Zheng, Y. Chen, H. Zeng, T.-A. Vu, B.-S. Hua, and S.-K. Yeung, "MarineInst: A foundation model for marine image analysis with instance visual description," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 239–257.

[2] S. Lian et al., "Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 29545–29559.

[3] P. Li, Y. Fan, Z. Cai, Z. Lyu, and W. Ren, "Detection method of marine biological objects based on image enhancement and improved YOLOv5S," *J. Mar. Sci. Eng.*, vol. 10, no. 10, 2022, Art. no. 1503. [Online]. Available: https://www.mdpi.com/2077-1312/10/10/1503

[4] S. Raine, R. Marchant, B. Kusy, F. Maire, and T. Fischer, "Point label aware superpixels for multi-species segmentation of underwater imagery," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 8291–8298, Jul. 2022.

[5] B. Kiefer et al., "1st workshop on maritime computer vision (MaCVi) 2023: Challenge results," in *Proc. Winter Conf. Appl. Comput. Vis. Workshops*, 2023, pp. 265–302.

[6] Z. Zheng et al., "CoralSCOP: Segment any COral image on this planet," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 28170–28180.

[7] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1290–1299.

[8] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[9] N. Ravi et al., "SAM 2: Segment anything in images and videos," in *Proc. 13th Int. Conf. Learn. Representations*, 2024.

[10] P. Zhang, T. Yan, Y. Liu, and H. Lu, "Fantastic animals and where to find them: Segment any marine animal with dual SAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition*, 2024, pp. 2578–2587.

[11] Z. Ziqiang, H. Liang, F. H. Wut, Y. H. Wong, A. P.-Y. Chui, and S. -K. Yeung, "HKcoral: Benchmark for dense coral growth form segmentation in the wild," *IEEE J. Ocean. Eng.*, vol. 50, no. 2, pp. 697–713, Apr. 2024.

[12] D. Akkaynak and T. Treibitz, "Sea-Thru: A method for removing water from underwater images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1682–1691.

[13] Y. Xie et al., "UVEB: A large-scale benchmark and baseline towards real-world underwater video enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22358–22367.

[14] C. Zhang et al., "Faster segment anything: Towards lightweight SAM for mobile applications," 2023, *arXiv:2306.14289*.

[15] X. Zhao et al., "Fast segment anything," 2023, *arXiv:2306.12156*.

[16] X. Cong, Y. Zhao, J. Gui, J. Hou, and D. Tao, "A comprehensive survey on underwater image enhancement based on deep learning," 2024, *arXiv:2405.19684*.

[17] F. Yu et al., "Segmentation of side scan sonar images on AUV," in *Proc. IEEE Underwater Technol.*, 2019, pp. 1–4.

[18] B. Arain, C. McCool, P. Rigby, D. Cagara, and M. Dunbabin, "Improving underwater obstacle detection using semantic image segmentation," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 9271–9277.

[19] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robot. Autom. Lett.*, vol. 3, no. 1, pp. 387–394, Jan. 2018, doi: 10.1109/LRA.2017.2730363.

[20] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3227–3234, Apr. 2020.

[21] Y. Lyu, G. Vosselman, G. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS j. photogrammetry remote sens.*, vo. 165, pp. 108–119, 2020.

[22] M. O'Byrne, V. Pakrashi, F. Schoefs, and B. Ghosh, "Semantic segmentation of underwater imagery using deep networks trained on synthetic imagery," *J. Mar. Sci. Eng.*, vol. 6, 2018, Art. no. 93.

[23] M. J. Islam et al., "Semantic segmentation of underwater imagery: Dataset and benchmark," in *IEEE/RSJ int. conf. intell. robots syst*, 2020, pp. 1769–1776.

[24] Z. Wang, L. Shen, M. Yu, K. Wang, Y. Lin, and M. Xu, "Domain adaptation for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 32, pp. 1442–1457, 2023.

[25] C. Wu et al., "Underwater image enhancement via dehazing and color restoration," 2025, *arXiv:2409.09779*.

[26] M. A. B. Siddique, J. Wu, I. Rekleitis, and M. J. Islam, "AquaFuse: Waterbody fusion for physics-guided view synthesis of underwater scenes," *IEEE Robot. Autom. Lett.*, vol. 10, no. 5, pp. 4316–4323, May 2025.

[27] S. Anwar, C. Li, and F. Porikli, "Deep underwater image enhancement," 2018, *arXiv:1807.03528*.

[28] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li, "Contrastive semi-supervised learning for underwater image restoration via reliable bank," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18145–18155.

[29] C. Zhang et al., "A survey on segment anything model (SAM): Vision foundation model meets prompt engineering," 2024, *arXiv:2306.06211*.

[30] Y. Xiong et al., "EfficientSAM: Leveraged masked image pretraining for efficient segment anything," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16111–16121.

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[32] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[33] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[34] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2020, pp. 1580–1589.

[35] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Representations*, 2022.

[36] M. Maaz et al., "EdgeNeXt: Efficiently amalgamated CNN-transformer architecture for mobile vision applications," in *European Conf. Comput. Vis.*, 2022, pp. 3–20.

[37] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19529–19539.

[38] Y. Liu, C. Shun, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7035–7049, 2020.

[39] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[40] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, and S. Kwong, "WaterMask: Instance segmentation for underwater imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 1305–1315.