Language-Guided Reasoning Segmentation for Underwater Images

Mingde Yao[a], King Man Tam[b], Menglu Wang[c], Lingen Li[a], Rei Kawakami[b]

[a] *The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR*
[b] *Institute of Science Tokyo, Tokyo, Japan*
[c] *University of Science and Technology of China, Hefei, China*

Abstract

In this paper, we introduce Language-Guided Reasoning Segmentation (LGRS), a framework that leverages human language instructions to guide underwater image segmentation. Unlike existing methods, that rely solely on visual cues or predefined categories, LGRS enables segmentation at underwater images based on detailed, context-aware textual descriptions, allowing it to tackle more challenging scenarios, such as distinguishing visually similar objects or identifying species from complex queries. To facilitate the development and evaluation of this approach, we create an underwater image-language segmentation dataset, the first of its kind, which pairs underwater images with detailed textual descriptions and corresponding segmentation masks. This dataset provides a foundation for training models capable of processing both visual and linguistic inputs simultaneously. Furthermore, LGRS incorporates reasoning capabilities through large language models, enabling the system to interpret complex relationships between objects in the scene and perform accurate segmentation in dynamic underwater environments. Notably, our method also demonstrates strong zero-shot segmentation capabilities, enabling the model to generalize to unseen categories without additional training. Experimental results show that LGRS outperforms existing underwater image segmentation methods in both accuracy and flexibility, offering a foundation for further advancements.

Email addresses: mingdeyao@foxmail.com (Mingde Yao), tam.k.4f46@m.isct.ac.jp (King Man Tam), vault@mail.ustc.edu.cn (Menglu Wang), lgli@link.cuhk.edu.hk (Lingen Li), kawakami.r.abb9@m.isct.ac.jp (Rei Kawakami)

Keywords: underwater image segmentation, reasoning segmentation, zero-shot segmentation, large language model

1. Introduction

Underwater images are crucial for applications such as marine species identification [1], habitat mapping [2], and pollution monitoring [3], providing vital insights into complex environments. In this context, accurate image segmentation is essential for isolating and identifying objects, enabling precise analysis. For autonomous underwater robots, interactive and adaptive segmentation [4] is particularly important. By incorporating human input or contextual instructions, robots can dynamically adjust their tasks, improving their ability to navigate, detect obstacles, map terrain, and assist in environmental monitoring. This interactive approach enhances the flexibility and efficiency of underwater exploration, making image segmentation a key tool in advancing underwater robotics.

Existing underwater image segmentation methods [5, 6, 7, 8, 9] primarily focus on leveraging visual features, such as patterns, textures, and other visual cues, to identify and delineate objects in underwater scenes. These approaches, often based on advanced convolutional and transformer-based networks, aim to improve segmentation accuracy despite challenging underwater conditions. However, these methods are typically designed to output fixed masks for specific objects, such as in salient or instance segmentation [5, 9]. Additionally, while effective for predefined tasks, these methods lack interactivity and cannot adapt to dynamic conditions or user input. Without the ability to incorporate higher-level guidance, like natural language instructions, these methods are limited in their flexibility. They struggle to adjust their outputs based on changing underwater environments or specific user needs, making them less versatile and reducing their effectiveness in real-world applications [6, 8].

On the other hand, natural image segmentation methods [10, 11, 12, 13, 14, 15, 16, 17, 18, 19] have started using language instructions to improve performance. By incorporating language instructions, these methods can understand higher-level context and adjust their segmentation decisions based on more complex reasoning. For example, natural language can help the model focus on specific objects or relationships between objects [10, 17, 18, 19], making the segmentation more accurate and contextually relevant. However, these techniques are not directly applicable to underwater images due

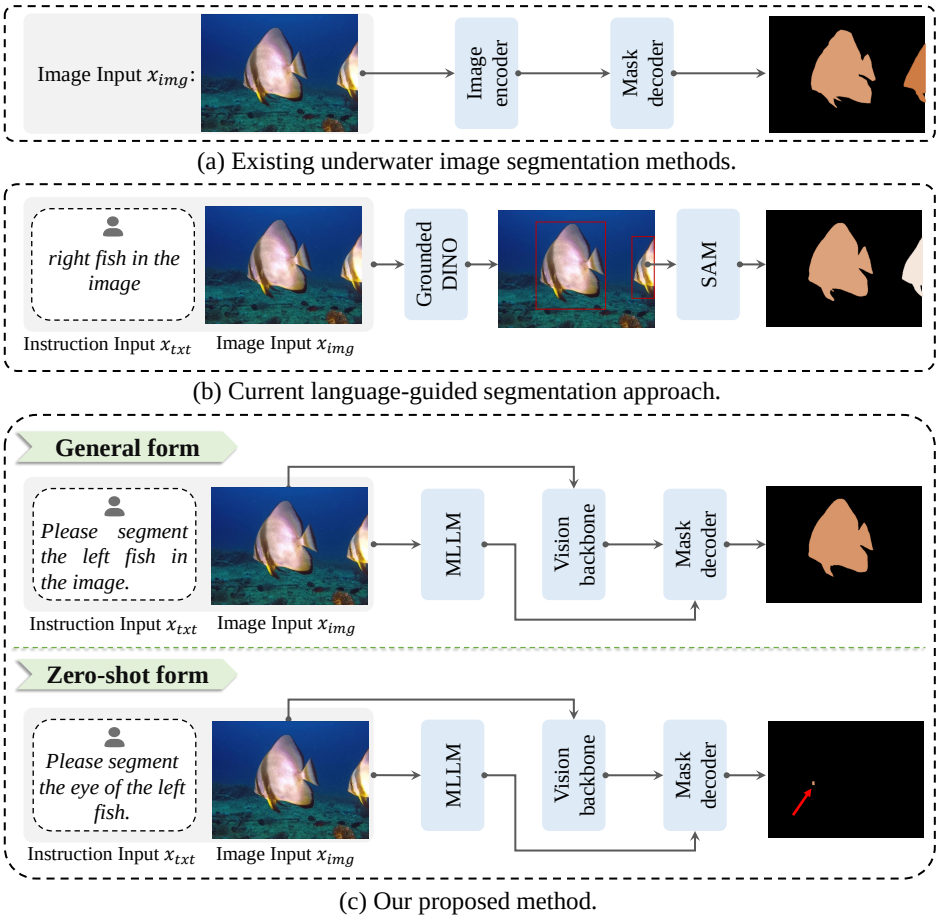

Figure 1: (a) Existing underwater image segmentation methods rely on fixed processing pipelines, limiting interactivity and adaptability. (b) Current language-guided segmentation approaches perform poorly on underwater images, struggling with nuanced reasoning and handling only simple, deterministic descriptions. (c) Our proposed method leverages language models to perform reasoning segmentation, effectively identifying objects such as fish based on complex descriptors (e.g., specific patterns or behaviors). Additionally, it exhibits robust zero-shot segmentation capabilities, demonstrating exceptional generalization performance.

to the unique challenges of the underwater environment [9, 20, 21, 22], where light attenuation, color shifts, and reduced visibility distort images, making it harder to identify and separate objects. Additionally, there is a significant lack of datasets specifically designed for language-guided segmentation in underwater images.

To address this gap, we develop the first-ever dataset tailored for language-guided reasoning in underwater image segmentation. This dataset incorporates both visual data and language descriptions, enabling models to leverage semantic guidance and reasoning to achieve robust segmentation performance for underwater images. To generate these instructions, we use a two-stage algorithm. First, we employ Large Language and Vision Assistant (LLaVA) [23] to generate natural language descriptions, highlighting key elements like spatial positioning or environmental cues. The generated instructions include both simple descriptions for object (*e.g.,* "a fish near the coral") and more complex ones that require reasoning and contextual understanding (*e.g.,* "an instrument capable of detecting underwater metals"). After generating these initial annotations, a rigorous human filtering and refinement process is applied, ensuring that the language aligns with real-world scenarios and provides useful guidance for segmentation tasks. Therefore, our dataset enables segmentation models to combine visual information with semantic cues in natural language, allowing them to adapt to changing underwater conditions and dynamically adjust for underwater conditions.

Moreover, we introduce Language-Guided Reasoning Segmentation (LGRS), a framework that leverages natural language as a semantic guide to enhance underwater image segmentation, as shown in Fig. 1. By incorporating pre-trained vision-language models [23], LGRS effectively processes language inputs and leverages external knowledge to interpret complex instructions and improve segmentation accuracy. However, two significant challenges arise: first, VLMs are designed to output text, creating a gap between language understanding and generating segmentation masks. Second, these models are primarily trained on natural images, whereas underwater images present unique challenges, such as severe visibility degradation, color shifts, and inconsistent lighting, which demand tailored solutions to bridge this domain gap.

To generate the segmentation mask, a key component of our approach is the inclusion of an additional token, $\langle SEG \rangle$, into the vocabulary of the model, which is designed to link language instructions with segmentation outputs. Specifically, when the $\langle SEG \rangle$ token is generated, its hidden embedding is passed to a mask decoder along with image features to produce the corresponding segmentation mask. By representing segmentation masks as tokens, the framework benefits from end-to-end training and achieves seamless integration between language and vision. To address the unique challenges of underwater imagery—such as color shifts, visibility degradation, and the

presence of occlusions—we utilize a pre-trained encoder [24] for robust feature extraction. Furthermore, we enhance the encoder by adding a bypass encoder that processes underwater-specific image features. This design ensures that the framework adapts effectively to the complexities of underwater environments while maintaining strong segmentation performance.

We evaluate our method on multiple benchmark underwater datasets and demonstrate its superiority over state-of-the-art methods in terms of accuracy and robustness. The results show that the inclusion of language-guided reasoning provides greater interpretability, making it a promising approach for addressing the complexities of underwater image analysis.

2. Related Work

2.1. Underwater Image Segmentation

The development of underwater image segmentation has seen significant progress, driven by the emergence of notable datasets. For underwater scene recognition and target detection, the SUN dataset [25] and the WildFish dataset [26] are widely utilized. Islam *et al.* [8] introduce the first underwater semantic segmentation dataset, consisting of 1,500 annotated images, which has been widely adopted in the research community. Following this, Nahuel *et al.* [27] create the DeepFish dataset, designed for instance segmentation of various fish species. More recently, the UIIS dataset [5] has been developed to address the shortage of general multiclass instance segmentation datasets for underwater images. In addition, the USIS10K dataset [9] is introduced to tackle the challenge of underwater salient instance segmentation. These datasets provide essential resources for advancing underwater image segmentation research and the development of more accurate models for complex underwater environments; however, there is currently no dataset specifically designed for language-guided underwater image segmentation.

Various methodologies have been proposed for underwater image segmentation, spanning traditional techniques to deep learning-based approaches. Early methods use image preprocessing and pixel clustering, such as CLAHE histogram equalization [28] and Particle Swarm Optimization [29]. Later, techniques like K-Means clustering with HOG feature extraction [30], parametric kernel graph cuts [31], and active contour models [32] were introduced to enhance segmentation accuracy. More recent work leverages deep learning, with fully convolutional networks applied to segmentation in underwater

images [32]. WaterMask [5], improves segmentation by using the graph attention to recover lost details due to image degradation and downsampling. Recently, USIS-SAM [9] takes advantage of the Segment Anything Model (SAM) [24], leveraging its pre-trained image segmentation capabilities to enhance underwater salient instance segmentation. Despite their advancements, these methods rely predominantly on visual information, which can be limiting in complex scenarios. Integrating natural language adds semantic context, enabling more nuanced segmentation and improving situational awareness and task flexibility, such as in practical robotic interactions.

2.2. Prompt-based Segmentation

Prompt-based segmentation methods [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 24, 33, 34] have evolved significantly, transitioning from traditional interactive approaches to modern, multi-modal paradigms. Early interactive methods, such as GrabCut [35] and Livewire [36, 37], rely on user-provided clicks or bounding boxes to refine segmentation boundaries through graph cuts or dynamic programming. Recent advancements leverage prompts in diverse forms, including points, boxes, and language. For example, SAM [24] and DINO [38] excel at general-purpose segmentation through spatial prompts, while CLIPSeg [39], and BLIPSeg [40] integrate vision-language models [41, 42] for text-guided segmentation. Unified frameworks like Pix2Seq [43], UNINEXT [44], and OFA [45] explore task generalization by harmonizing diverse visual modalities with sequential or multi-task outputs. Additionally, ControlNet [46], MaskCLIP [47], and methods like LISA [10] and GSVA [48] enhance segmentation flexibility and precision, employing innovative mechanisms to leverage prompt information effectively. However, these methods are primarily designed for natural images. Underwater images, characterized by a lack of large-scale datasets and inherent challenges such as color distortion and low visibility, have yet to benefit from language-guided semantic segmentation approaches.

3. Dataset

We construct a novel dataset Reference Under Water Segmentation (RefUDS) that integrates both visual and linguistic information, aiming to harness the complementary strengths of both modalities. This multimodal dataset is specifically designed to enable language-guided reasoning, improv-

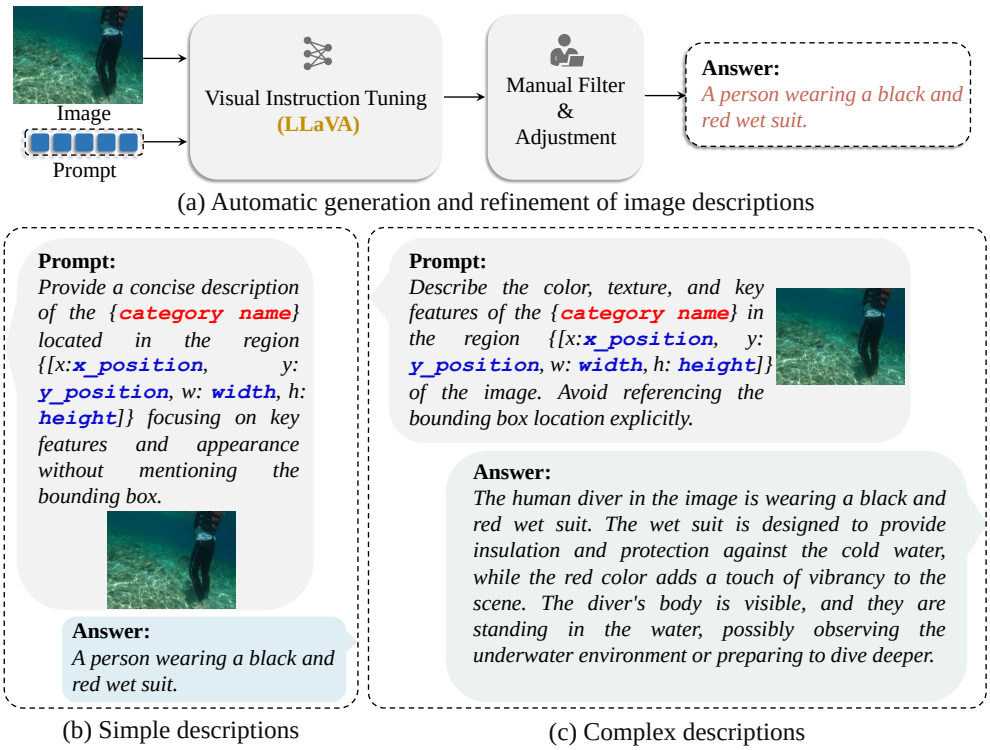

Figure 2: Overview of our constructed underwater dataset. (a) We initially utilize Visual Instruction Tuning (LLaVA) to automatically generate coarse descriptions for the images. These descriptions are then refined by human annotators through filtering and adjustment to ensure both accuracy and clarity. (b)&(c)We generate two types of descriptions: simple descriptions for straightforward identification and complex descriptions requiring reasoning capabilities.

ing the robustness and accuracy of underwater image segmentation in complex and degraded environments.

Our dataset builds upon two existing underwater segmentation datasets, USIS10K [9] and UIIS [5], which feature diverse annotated underwater images, including marine species, coral reefs, underwater structures, and other elements commonly found in underwater environments. While these datasets are useful for conventional segmentation tasks, they lack linguistic annotations essential for enabling models to reason about object relationships, contextual information, and spatial configurations.

To address this, we employed a two-step annotation process. First, we used LLaVA [23] to generate coarse textual descriptions for each image.

Specifically, as shown in Fig. 2, we applied two types of prompts to produce simple and complex descriptions. For images containing multiple objects, we further localized descriptions to specific objects using bounding box coordinates. However, due to the unique characteristics of underwater images, such as color distortion, low contrast, and complex scenes, the generated descriptions were often insufficient for training and testing.

To ensure high-quality annotations, we further conduct a rigorous human filtering and adjustment process guided by the following principles:

- ***Relevance***: *Descriptions must provide meaningful semantic guidance for segmentation. Irrelevant or overly generic descriptions should be removed.*
- ***Clarity***: *Annotations are revised to be concise and unambiguous, avoiding overly complex language.*
- ***Completeness***: *Descriptions comprehensively covered the object's appearance, position, and relationships with other objects.*
- ***Accuracy***: *Annotations were carefully validated to align with the visual properties of the image.*

As a result, each segmentation mask is annotated with both simple and complex descriptions:

- ***Simple Descriptions***: *Focus on essential object-level attributes such as shape, size, color, and texture. For example, "The fish is small, with a silver body and a long tail."*
- ***Complex Descriptions***: *Incorporate contextual and relational information, such as spatial arrangement, interactions, and scene context. For example, "The fish is located near the center of the image, swimming above a coral reef. Its silver body contrasts with the greenish coral and blue water background."*

Our dataset comprises 40,798 annotations paired with 12,972 images, offering both simple and complex descriptions for training language-guided reasoning segmentation models. This dual-level annotation strategy enables models to integrate fine-grained visual feature recognition with high-level contextual reasoning, making the dataset an essential resource for advancing multimodal segmentation research. Our whole dataset can be divided into two subsets of RefUIIS and RefUSIS, derived from the UIIS and USIS10K dataset, respectively.

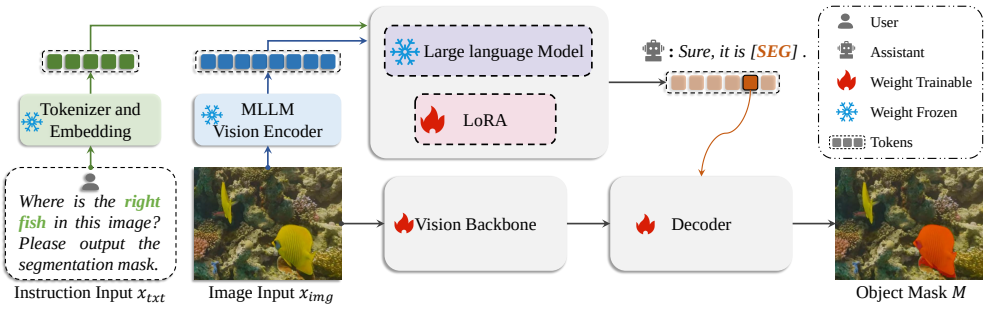

Figure 3: Framework of our proposed underwater segmentation method. It integrates hierarchical feature extraction, multi-modal contextual reasoning, and a task-specific inference module. By combining domain-specific priors with a scalable architecture, the model achieves accurate and robust segmentation without additional labeled data for new environments.

4. Method

4.1. Overview

We introduce a framework designed to address the challenges of underwater image segmentation with language guidance. Our approach introduces specific design elements to improve segmentation performance for underwater environments. Current language-based segmentation methods suffer from several limitations. Two-stage approaches, such as Grounded SAM [49], rely on first detecting objects and then segmenting them, often leading to inefficiencies and errors due to cascading failures between the stages. On the other hand, recent methods in natural image segmentation, such as LISA [10] and GSVA [48], leverage pre-trained vision encoders [24] and large language models (LLMs) to perform segmentation by combining semantic understanding with vision features. However, these models struggle to capture the unique characteristics of underwater imagery, such as color distortion, low contrast, and varying visibility, due to the domain-specific nature of underwater scenes.

Our framework takes an image and a language instruction as input and outputs a segmentation mask guided by the instruction. To address the limitations of existing methods, our approach seamlessly integrates multimodal reasoning with precise segmentation to tackle underwater segmentation challenges. The framework comprises three core components: a visual backbone, a multimodal language model, and a segmentation decoder. Specifically, the multimodal language model leverages the reasoning capabilities of pre-trained LLMs to process underwater images alongside corresponding instructions and

outputs tokens, including a $\langle SEG \rangle$ token enriched with semantic information about the segmentation target. Subsequently, image features extracted by the vision backbone and the $\langle SEG \rangle$ tokens are fed into a segmentation decoder to produce the final segmentation mask. Below, we describe the process in detail.

4.2. Text and Visual Feature Encoding

Textual Tokens. Existing vision-language models [41, 42], while capable of interpreting images and generating meaningful textual responses, face limitations in segmentation tasks. One issue is their inability to directly output actionable segmentation information, as their outputs are restricted to plain text. Moreover, these models, typically trained on terrestrial datasets, are ill-equipped to handle the unique properties of underwater images, such as light absorption and scatter, which significantly alter visual features.

To address these problems, inspired by LISA [10], we adopt the "embedding as mask" strategy. In this way, the multimodal-language model outputs a specialized $\langle SEG \rangle$ token that encapsulates both semantic and spatial information, enabling precise segmentation by aligning visual and textual features within a shared embedding space. As shown in Fig. 3, the $\langle SEG \rangle$ token serves as a bridge between high-level multimodal reasoning and low-level spatial predictions. It captures the contextual meaning of objects described in the language prompt, such as their relationships and attributes, while simultaneously encoding the spatial structure and appearance from the visual input.

However, training a multimodal-language model from scratch would be prohibitively resource-intensive, requiring vast datasets and computational power. To achieve this, we employ LoRA (Low-Rank Adaptation) [50] for fine-tuning the multimodal language model. LoRA allows efficient fine-tuning by introducing learnable low-rank adaptations to the pre-trained model, enabling it to better adapt to underwater imagery. Through LoRA fine-tuning on our curated dataset, we enhance the model's ability to process underwater scenes effectively.

Visual Features. For visual feature extraction, we use the pre-trained image encoder from SAM, which has demonstrated robustness in extracting strong visual features across diverse tasks. However, due to the unique characteristics of underwater images, such as uneven lighting and reduced clarity, the extracted features may lack sufficient representation of local and

global variations. To address this, we introduce an auxiliary feature extraction pathway, designed to complement the pre-trained encoder. Specifically, we employ a lightweight convolutional layer to capture local features and a fully connected layer to extract global variation features. The outputs of the encoder are concatenated with the features extracted by SAM, to form a comprehensive representation of the image, which is then passed to the mask decoder.

Finally, the segmentation decoder integrates enriched image features and the $\langle SEG \rangle$ token to produce the final segmentation mask tailored for underwater environments. We show the overall pipeline in Fig. 3.

4.3. Loss Functions

To ensure robust performance across multimodal reasoning and precise segmentation tasks, we employ multiple losses, which are designed to jointly optimize the textual understanding capabilities of the multimodal language model and the segmentation accuracy of the decoder. The total loss function combines three components: Textual Loss, Binary Cross-Entropy Loss, and Dice Loss. Each component targets a specific aspect of the model's performance.

Textual Loss The textual loss L_{txt} is applied to fine-tune the language model's ability to generate semantically rich and accurate textual embeddings, particularly the $\langle SEG \rangle$ token, which carries critical segmentation-related semantic information. We use the cross-entropy loss for this purpose:

$$L_{txt} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{txt}^{(i,j)} \log \hat{y}_{txt}^{(i,j)}, \tag{1}$$

where N is the number of samples, M is the number of token classes, $y_{txt}^{(i,j)}$ is the true label for the j-th token in the i-th sample, and $\hat{y}_{txt}^{(i,j)}$ is the predicted probability for that token. This loss function helps minimize the difference between predicted and true labels, improving the model's embedding accuracy.

Binary Cross-Entropy Loss The Binary Cross-Entropy Loss L_{bce} is employed to supervise pixel-level mask predictions. It ensures that the predicted segmentation mask aligns with the ground-truth binary mask by penalizing incorrect pixel classifications. The loss is defined as:

$$L_{bce} = -\frac{1}{P} \sum_{k=1}^{P} \left[y_k \log \hat{y}_k + (1 - y_k) \log(1 - \hat{y}_k) \right], \tag{2}$$

where P is the total number of pixels, y_k and $\hat{y}_k$ denote the ground-truth and predicted probabilities for pixel k, respectively.

Dice Loss To complement the Binary Cross-Entropy Loss, we include the Dice Loss L_{dice}, which is particularly effective in handling imbalanced segmentation tasks. The Dice Loss measures the overlap between the predicted and ground-truth masks:

$$L_{dice} = 1 - \frac{2\sum_{k=1}^{P} y_k \hat{y}_k}{\sum_{k=1}^{P} y_k + \sum_{k=1}^{P} \hat{y}_k}, \tag{3}$$

where y_k and $\hat{y}_k$ are the same as in the Binary Cross-Entropy Loss.

Combined Loss Function The total loss function L integrates the three loss terms, with each weighted by a hyperparameter λ to control its relative contribution:

$$L = \lambda_{txt} L_{txt} + \lambda_{bce} L_{bce} + \lambda_{dice} L_{dice}. \tag{4}$$

Here, λ_{txt}, λ_{bce}, and λ_{dice} are scalar coefficients chosen through hyperparameter tuning. These weights balance the importance of accurate textual embedding generation and precise segmentation mask prediction.

5. Experiments

In this section, we provide a comprehensive evaluation of our proposed underwater reasoning segmentation framework. We begin by describing the datasets used for training and testing, followed by a detailed explanation of the experimental setup. Next, we present both quantitative and qualitative results, highlighting the framework's performance. Additionally, we demonstrate the zero-shot capabilities of our method on unseen categories. Finally, we discuss the limitations of our approach and analyze failure cases to provide insights for future improvements.

5.1. Datasets

As metioned in Sec. 3, we build the RefUDS using a combination of existing datasets, including USIS10K [9] and UIIS [5], to address the specific challenges of underwater image segmentation. Specifically, the training set and test set are carefully curated to ensure diverse underwater scenes, including coral reefs, marine life, and man-made underwater structures. The training set consists of 11,377 images with 37,982 language annotations, while

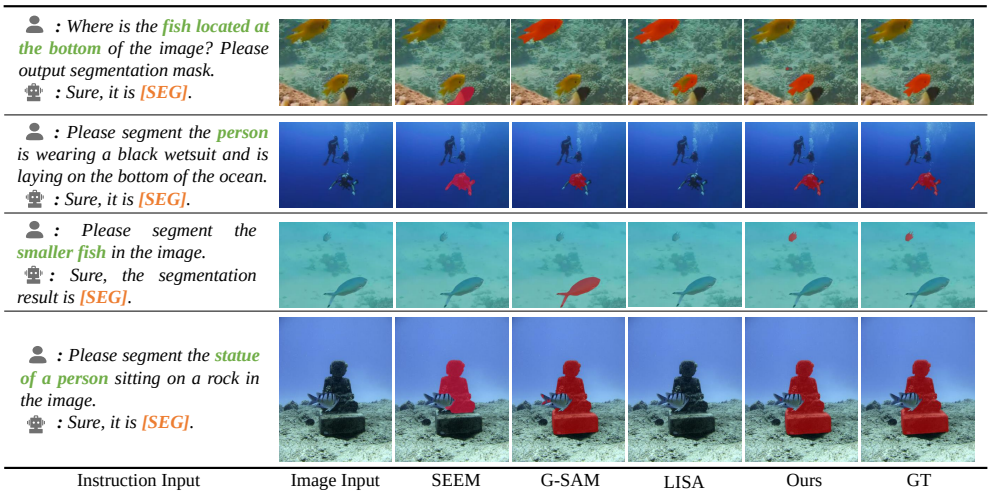

Figure 4: Visual comparisons of segmentation results. It can be seen that, while SEEM and G-SAM can segment objects, they often exhibit semantic misunderstanding, leading to over-segmentation or incorrect segmentation of objects. Similarly, LISA also struggles with segmentation errors and, in certain cases, fails to generate valid segmentation masks (e.g., the third scenario). In contrast, our method demonstrates reliable and accurate underwater image segmentation.

the test set includes 1,595 images and 2,870 language annotations, providing a comprehensive benchmark for evaluation.

Moreover, following LISA [10], we utilize training data from semantic segmentation, referring segmentation, and visual question answering (VQA) datasets. Semantic datasets like ADE20K [51, 52], COCO-Stuff [53], and LVIS-PACO [54] are reformatted into question-answer pairs with binary masks as ground truth. Referring segmentation datasets, such as refCOCO [55] and refCLEF [56], provide images with textual descriptions, which were similarly converted for visual-text alignment. VQA datasets, including LLaVA-Instruct-150k [23], enriched the model's general visual reasoning capabilities by incorporating diverse queries and answers.

5.2. Experimental Setup

Our framework integrates a vision backbone, a multimodal language model, and a segmentation decoder. We use the SAM image encoder as the visual backbone to extract robust visual features, while leveraging a pretrained language model LLaVA-v1.5 [23] fine-tuned with LoRA (Low-Rank Adaptation) [50] to efficiently adapt the language model to the underwater domain.

The model is trained using the AdamW optimizer with a learning rate of 1e-4, and a batch size of 8 on 4 GTX4090Ti NVIDIA GPUs.

To validate our framework, We compare our method against a variety of state-of-the-art approaches, including Underwater segmentation methods like USIS [9], which rely solely on visual features, and language-guided methods such as Grounded SAM [49] (G-SAM in short), SEEM [12], and LISA [10], which integrate multimodal reasoning. For language-guided methods, we directly evaluate their zero-shot performance using official pretrained models and code without further fine-tuning. We follow previous methods [57] to set evaluation metrics, including IoU (Intersection over Union) [58] for mask overlap, and GIoU (Generalized Intersection over Union) [59] and CIoU (Complete Intersection over Union) [60] for spatial consistency and completeness. GIoU improves upon IoU by considering the area outside the bounding box, while CIoU further incorporates the aspect ratio and center distance to evaluate the alignment between predicted and ground truth bounding boxes, providing a comprehensive assessment of performance and robustness in comparison to state-of-the-art approaches.

5.3. Results

Quantitative Results. Our framework achieved state-of-the-art performance across a variety of tasks, significantly outperforming existing methods in both general language-guided segmentation and underwater-specific segmentation methods, as shown in Tab. 1.

When compared to language-guided segmentation methods such as Grounded SAM [49], SEEM [12], and LISA [10], our framework demonstrated superior performance in metrics like IoU, GIoU, and CIoU, particularly excelling in queries that require fine-grained object understanding and relational reasoning. For instance, it surpassed methods that struggle with capturing complex relationships between objects. These results highlight the model's ability to effectively integrate linguistic context with visual inputs for improved segmentation accuracy and reasoning.

Since existing underwater image segmentation methods like USIS [9] focus on global segmentation tasks that aim to segment all possible objects in an image, their objectives differ fundamentally from our language-based segmentation approach. As a result, direct numerical comparisons are not feasible. Our framework is designed to segment specific objects or regions based on linguistic guidance, addressing a different and more fine-grained task.

Table 1: Comparison of our method with existing methods on the RefUDS datasets, which consists of two sub-datasets RefUIIS and RefUSIS.

Dataset	Method	mIoU	gIoU	cIoU
RefUSIS	G-SAM [49]	61.64	60.89	59.43
	SEEM [12]	31.49	30.16	31.49
	Lang-Seg [61]	45.56	44.38	42.94
	LISA [10]	11.85	10.76	10.50
	G-tag2text [49]	70.01	69.56	68.39
	G-RAM [49]	71.73	71.22	70.17
	Ours	**75.81**	**75.24**	**77.92**
RefUIIS	G-SAM [49]	34.87	33.57	30.40
	SEEM [12]	15.05	13.94	15.05
	Lang-Seg [61]	19.39	18.27	15.16
	LISA [10]	8.67	7.54	6.36
	G-tag2text [49]	42.81	41.75	38.73
	G-RAM [49]	43.62	42.24	39.39
	Ours	**48.21**	**45.85**	**46.77**

Qualitative Results. The qualitative results further emphasize the strengths of our framework in understanding and segmenting complex underwater scenes. Fig. 4 presents results that our method excels compared to baseline approaches. For instance, in a query like "Segment the coral closest to the camera," traditional methods such as UIIS [5] often misidentify the target, focusing on the largest coral regardless of proximity. In contrast, our framework precisely identifies and segments the correct coral by leveraging textual guidance and spatial reasoning.

Another example is the query "Segment the fish swimming above the coral." Methods like Grounded SAM [49] and SEEM [12] often struggle with relational queries, resulting in over-segmentation or missed details. In contrast, our approach successfully distinguishes the fish from nearby objects, producing more accurate masks. As shown in Fig. 4, our method precisely captures the intended regions, highlighting its strength in handling complex relational queries.

5.4. Zero-Shot Performance

As shown in Fig. 5, further analysis on unseen categories and zero-shot tasks demonstrated the robustness of our framework. For example, when

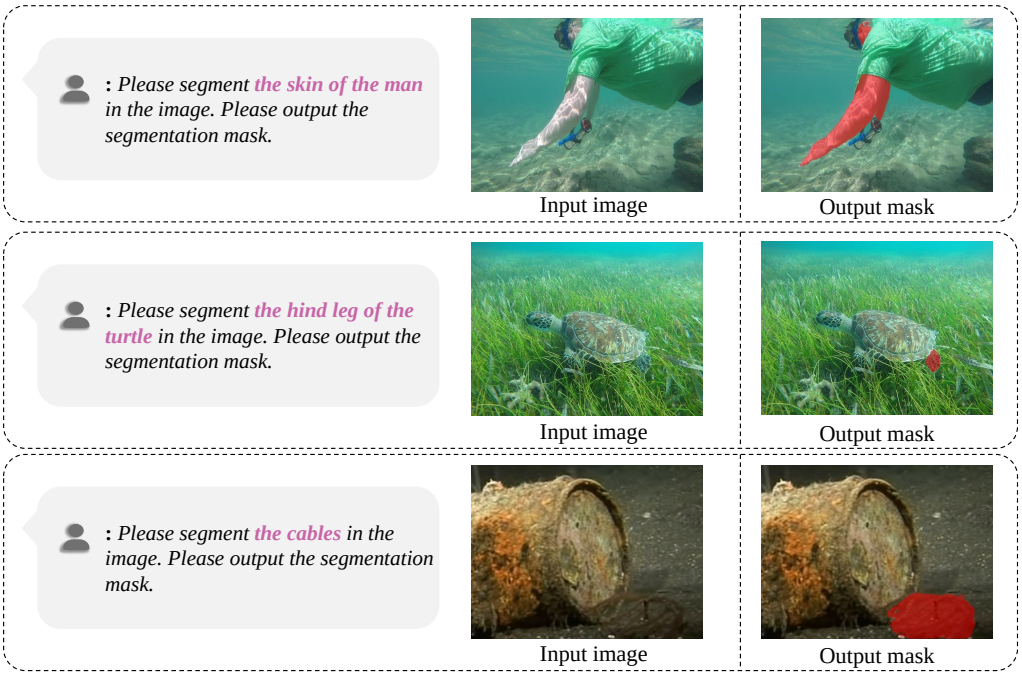

Figure 5: Zero-shot segmentation results on unseen underwater scenarios. Our method effectively generalizes to diverse environments and object types without additional training, showcasing robust adaptability to diverse underwater conditions and maintaining high segmentation accuracy.

evaluated on reasoning-based queries such as "Why can the fish see the environment?" or "Segment the fish closest to the coral," our approach consistently outperformed baseline methods. Such queries require the model to reason about functional relationships or spatial hierarchies, areas where traditional or unimodal approaches falter. These results affirm that our multimodal framework generalizes well to complex, unseen scenarios, demonstrating its utility for real-world underwater applications.

5.5. Limitations

Despite the strong performance of our approach, certain failure cases reveal areas for improvement. First, our model currently handles segmentation tasks on single scenes within an image but struggles when multiple scenes are described in a single instruction. For example, if the input query includes instructions like "segment the eye of the left fish and the fish on the right," the model fails to segment both objects correctly. This limitation arises because

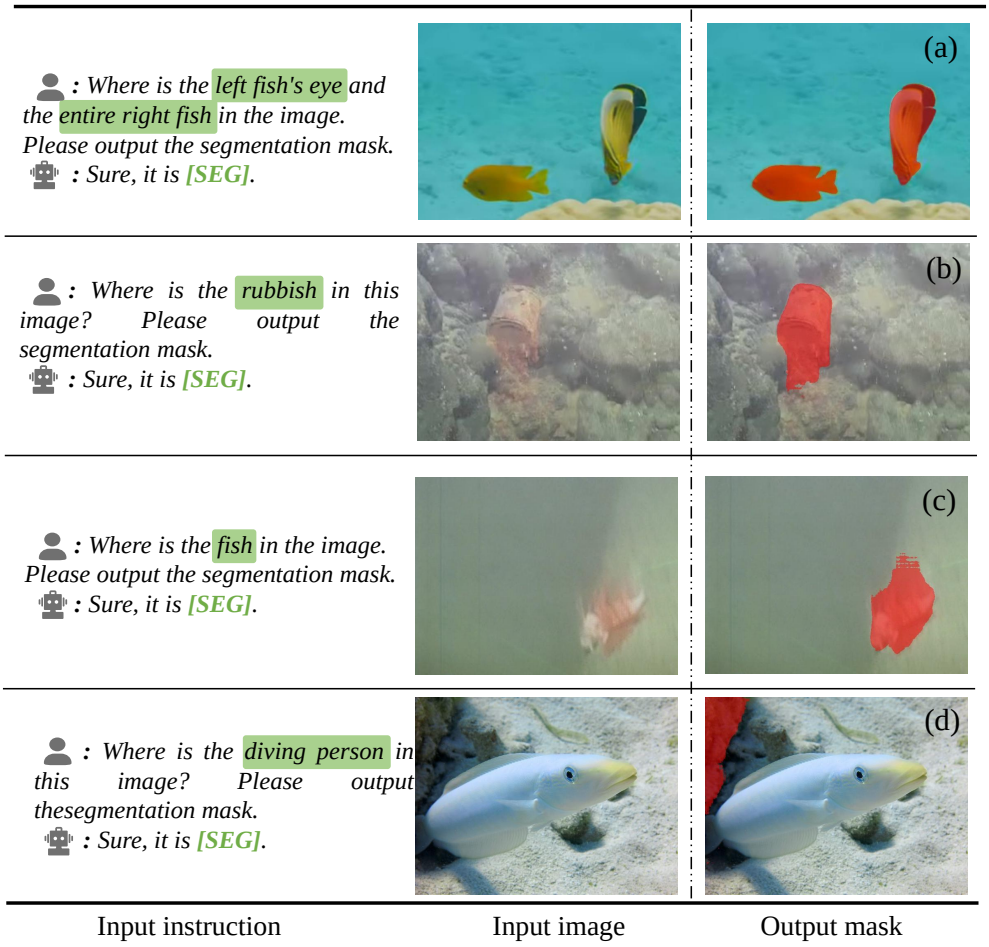

Figure 6: Failure cases. Although our method successfully handles most underwater image segmentation scenarios, it still encounters challenges in certain extreme cases: (a) When segmenting multiple objects as instructed, the results may deviate from the language instructions, with segmentation granularity being insufficient. (b) Poor segmentation performance occurs in extreme low-quality underwater images where objects are unclear. (c) Dynamic scenes with motion blur can lead to inaccurate segmentation results. (d) When the instructions specify objects absent in the image, segmentation may fail or produce incorrect results. We provide a detailed analysis of these issues in Sec. 5.4 and discuss potential improvements.

our approach is designed to handle one target object per instruction. However, this work provides a baseline, and future improvements could include adjusting the $\langle SEG \rangle$ token to allow multiple instances for handling complex

Table 2: Ablation study on visual encoder.

Model	mIoU	GIoU	CIoU
Pretrained-SAM	74.77	74.35	75.28
Ours	**75.81**	**75.24**	**77.92**

queries with multiple scenes.

Additionally, dynamic underwater scenes, such as schools of fish in rapid motion, present challenges in maintaining temporal consistency for segmentation, as the model currently processes static images effectively but struggles with video or multi-frame tasks.

Furthermore, when the instruction refers to objects that are not present in the image, such as "segment the unicorn", the model may produce incorrect results. To address these limitations, we plan to develop and incorporate more diverse datasets that include a wider range of instructions, as well as introduce rejection tokens to enhance flexibility and generalization. These improvements aim to make our method more robust, adaptable, and capable of handling a broader spectrum of segmentation tasks. We show these visual results in Fig. 6.

5.6. Ablation Study

To thoroughly evaluate the contributions of individual components in our proposed model as shown in Tab. 2, we conducted an ablation study on the USIS subset of the RefUDS dataset. We explore additional pathways in vision encoder, where we test the effect of incorporating an extra pathway, analyzing whether this fusion improves the model's ability to handle complex underwater images. Incorporating additional pathways within the vision encoder, which allowed for better integration of underwater features, resulted in substantial improvements in underwater image segmentation task. This fusion allowed the model to better align visual content with textual queries, leading to more accurate segmentation boundaries. These results demonstrate the effectiveness of our method.

6. Conclusion

In this paper, we introduce Language-Guided Reasoning Segmentation (LGRS), a novel framework that integrates natural language instructions with underwater image segmentation, marking a significant advancement

over traditional methods that rely solely on visual cues. A key contribution of this work is the development of the first-ever underwater image-language segmentation dataset, which pairs underwater images with detailed textual descriptions and corresponding segmentation masks. This dataset enables the model to process both visual and linguistic inputs simultaneously, facilitating better handling of complex queries, spatial relationships, and fine-grained object segmentation in challenging underwater environments. Our approach, which leverages multimodal inputs, outperforms existing methods in terms of accuracy and robustness, while also demonstrating strong zero-shot capabilities for generalization to unseen categories. Although challenges remain—such as handling ambiguous queries, segmenting fine-scale details, and dealing with dynamic scenes—our dataset and framework provide a solid foundation for further advancements in underwater image segmentation.

Our contributions pave the way for advancements in underwater computer vision and multimodal reasoning, with potential applications in marine biology, ocean exploration, and autonomous underwater vehicles. Future work will focus on enhancing the framework's robustness to extreme underwater conditions, exploring temporal coherence for video data, and adapting the model to unseen domains for broader applicability.

References

[1] M. Pedersen, J. Bruslund Haurum, R. Gade, T. B. Moeslund, Detection of marine animals in a new underwater dataset with varying visibility, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[2] H. Mohamed, K. Nadaoka, T. Nakamura, Automatic semantic segmentation of benthic habitats using images from towed underwater camera in a complex shallow water environment, Remote Sensing 14 (8) (2022). doi:10.3390/rs14081818.

[3] S. Saji, M. S. Manikandan, J. Zhou, L. R. Cenkeramaddi, Underwater debris detection using visual images and yolov8n for marine pollution monitoring, in: 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA), 2024, pp. 1–6. doi:10.1109/ICIEA61579.2024.10664718.

[4] J. Wang, B. Li, Y. Zhou, E. Rocco, Q. Meng, Compact and fast underwater segmentation network for autonomous underwater vehicles, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2020.

[5] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, S. Kwong, Watermask: Instance segmentation for underwater imagery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1305–1315.

[6] M. O'Byrne, V. Pakrashi, F. Schoefs, B. Ghosh, Semantic segmentation of underwater imagery using deep networks trained on synthetic imagery, Journal of Marine Science and Engineering 6 (3) (2018). doi:10.3390/jmse6030093.

[7] C. Guo, R. Wu, X. Jin, L. Han, Z. Chai, W. Zhang, C. Li, Underwater ranker: Learn which is better and how to be better, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[8] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, J. Sattar, Semantic Segmentation of Underwater Imagery: Dataset and Benchmark, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE/RSJ, 2020.

[9] S. Lian, Z. Zhang, H. Li, W. Li, L. T. Yang, S. Kwong, R. Cong, Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset, in: Forty-first International Conference on Machine Learning, 2024.

[10] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, J. Jia, Lisa: Reasoning segmentation via large language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 9579–9589.

[11] S. Yang, M. Xia, G. Li, H.-Y. Zhou, Y. Yu, Bottom-up shift and reasoning for referring image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11266–11275.

[12] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, Y. J. Lee, Segment everything everywhere all at once, in: Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 19769–19782.

[13] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, et al., Visionllm: Large language model is also an open-ended decoder for vision-centric tasks, Advances in Neural Information Processing Systems 36 (2024).

[14] V. K. Nagaraja, V. I. Morariu, L. S. Davis, Modeling context between objects for referring expression understanding, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 792–807.

[15] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, N. Peng, L. Wang, Y. J. Lee, J. Gao, Generalized decoding for pixel, image, and language, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15116–15127.

[16] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, Advances in neural information processing systems 35 (2022) 23716–23736.

[17] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, X. Jin, Pixellm: Pixel reasoning with large multimodal model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26374–26383.

[18] R. Yang, L. Song, Y. Li, S. Zhao, Y. Ge, X. Li, Y. Shan, Gpt4tools: Teaching large language model to use tools via self-instruction, Advances in Neural Information Processing Systems 36 (2024).

[19] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, Y. Zhuang, Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, Advances in Neural Information Processing Systems 36 (2024).

[20] R. Liu, X. Fan, M. Zhu, M. Hou, Z. Luo, Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light, IEEE transactions on circuits and systems for video technology 30 (12) (2020) 4861–4875.

[21] R. Liu, Z. Jiang, S. Yang, X. Fan, Twin adversarial contrastive learning for underwater image enhancement and beyond, IEEE Transactions on Image Processing 31 (2022) 4922–4936.

[22] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, C. Li, Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement, IEEE Transactions on Image Processing 31 (2022) 3997–4010.

[23] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning (2023).

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, R. Girshick, Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4015–4026.

[25] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3485–3492.

[26] P. Zhuang, Y. Wang, Y. Qiao, Wildfish: A large benchmark for fish recognition in the wild, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 1301–1309.

[27] N. E. Garcia-D'Urso, A. Galan-Cuenca, P. Climent-Pérez, M. Saval-Calvo, J. Azorin-Lopez, A. Fuster-Guillo, Efficient instance segmentation using deep learning for species identification in fish markets, in: 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–8.

[28] M. S. Hitam, E. A. Awalludin, W. N. Jawahir Hj Wan Yussof, Z. Bachok, Mixture contrast limited adaptive histogram equalization for underwater image enhancement, in: 2013 International Conference on Computer Applications Technology (ICCAT), 2013, pp. 1–5. doi:10.1109/ICCAT.2013.6522017.

[29] R. Zhang, J. Liu, Underwater image segmentation with maximum entropy based on particle swarm optimization (pso), in: First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06), Vol. 2, 2006, pp. 360–636. doi:10.1109/IMSCCS.2006.280.

[30] M. Rajasekar, C. Aruldoss, M. Anto Bennet, Underwater k-means clustering segmentation using svm classification, Middle-East J Sci Res 23 (2015) 2166–2172.

[31] M. B. Salah, A. Mitiche, I. B. Ayed, Multiregion image segmentation by parametric kernel graph cuts, IEEE Transactions on Image Processing 20 (2) (2011) 545–557. doi:10.1109/TIP.2010.2066982.

[32] M. Chicchon, H. Bedon, C. R. Del-Blanco, I. Sipiran, Semantic segmentation of fish and underwater environments using deep convolutional neural networks and learned active contours, IEEE Access 11 (2023) 33652–33665. doi:10.1109/ACCESS.2023.3262649.

[33] J. Hu, J. Lin, S. Gong, W. Cai, Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects, Proceedings of the AAAI Conference on Artificial Intelligence 38 (11) (2024) 12511–12518.

[34] J. Hu, J. Lin, J. Yan, S. Gong, Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[35] C. Rother, V. Kolmogorov, A. Blake, Grabcut -interactive foreground extraction using iterated graph cuts, in: ACM Transactions on Graphics (SIGGRAPH), ACM, 2004.

[36] W. A. Barrett, E. N. Mortensen, Interactive live-wire boundary extraction, Medical image analysis 1 (4) (1997) 331–341.

[37] A. X. Falcao, J. K. Udupa, S. Samarasekera, S. Sharma, B. E. Hirsch, R. d. A. Lotufo, User-steered image segmentation paradigms: Live wire and live lane, Graphical models and image processing 60 (4) (1998) 233–260.

[38] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the International Conference on Computer Vision (ICCV), 2021.

[39] T. Lüddecke, A. Ecker, Image segmentation using text and image prompts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7086–7096.

[40] O. Ülger, M. Kulicki, Y. Asano, M. R. Oswald, Self-guided open-vocabulary semantic segmentation, arXiv preprint arXiv:2312.04539 (2023).

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[42] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.

[43] T. Chen, S. Saxena, L. Li, D. J. Fleet, G. Hinton, Pix2seq: A language modeling framework for object detection, arXiv preprint arXiv:2109.10852 (2021).

[44] B. Yan, Y. Jiang, J. Wu, D. Wang, Z. Yuan, P. Luo, H. Lu, Universal instance perception as object discovery and retrieval, in: CVPR, 2023.

[45] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, CoRR abs/2202.03052 (2022).

[46] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models (2023).

[47] C. Zhou, C. C. Loy, B. Dai, Extract free dense labels from clip, in: European Conference on Computer Vision (ECCV), 2022.

[48] Z. Xia, D. Han, Y. Han, X. Pan, S. Song, G. Huang, Gsva: Generalized segmentation via multimodal large language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 3858–3869.

[49] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, L. Zhang, Grounded sam: Assembling open-world models for diverse visual tasks (2024). arXiv:2401.14159.

[50] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.

[51] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, International Journal of Computer Vision 127 (3) (2019) 302–321.

[52] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[53] H. Caesar, J. Uijlings, V. Ferrari, Coco-stuff: Thing and stuff classes in context, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1209–1218. doi:10.1109/CVPR.2018.00132.

[54] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian, A. Mousavi, Y. Song, A. Dubey, D. Mahajan, Paco: Parts and attributes of common objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7141–7151.

[55] S. Kazemzadeh, V. Ordonez, M. Matten, T. Berg, ReferItGame: Referring to objects in photographs of natural scenes, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 787–798. doi:10.3115/v1/D14-1086.

[56] S. Kazemzadeh, V. Ordonez, M. Matten, T. Berg, Referitgame: Referring to objects in photographs of natural scenes, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 787–798.

[57] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in:

Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 11–20.

[58] J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, Unitbox: An advanced object detection network, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 516–520.

[59] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.

[60] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-iou loss: Faster and better learning for bounding box regression, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 12993–13000.

[61] L. Medeiros, Lang segment anything, `https://github.com/luca-medeiros/lang-segment-anything`, accessed: 2024-12-14 (2024).