Enhancing Vision-Language Models with Morphological and Taxonomic Knowledge: Towards Coral Recognition for Ocean Health

Hongyong Han[1], **Wei Wang**[1*], **Gaowei Zhang**[1], **Mingjie Li**[2,3], **Yi Wang**[1]
[1]Beijing University of Posts and Telecommunications
[2]Technology Innovation Center for South China Sea Remote Sensing, Surveying and Mapping Collaborative Application, Ministry of Natural Resources
[3]South China Sea Development Research Institute, Ministry of Natural Resources
{hanhongyong, weiwang, zhanggaowei, yiwang}@bupt.edu.cn, lmj_21@163.com

Abstract

Coral reefs play a crucial role in marine ecosystems, offering a nutrient-rich environment and safe shelter for numerous marine species. Automated coral image recognition aids in monitoring ocean health at a scale without experts' manual effort. Recently, large vision-language models like CLIP have greatly enhanced zero-shot and low-shot classification capabilities for various visual tasks. However, these models struggle with fine-grained coral-related tasks due to a lack of specific knowledge. To bridge this gap, we compile a fine-grained coral image dataset consisting of 16,659 images with taxonomy labels (from *Kingdom* to *Species*), accompanied by morphology-specific text descriptions for each species. Based on the dataset, we propose CORAL-Adapter, integrating two complementary kinds of coral-specific knowledge (biological taxonomy and coral morphology) with general knowledge learned by CLIP. CORAL-Adapter is a simple yet powerful extension of CLIP with only a few parameter updates and can be used as a plug-and-play module with various CLIP-based methods. We show improvements in accuracy across diverse coral recognition tasks, *e.g.,* recognizing corals unseen during training those are prone to bleaching or from different oceans.

Introduction

Coral reefs are among the most biodiverse ecosystems on the earth, providing habitats for over a quarter of marine species, including fish, mollusks, and crustaceans. Unfortunately, coral reefs face significant threats from human activities as well as natural challenges like global warming and ocean acidification. According to studies (ICRAN Online; accessed 25-May-2014; Elawady 2015), more than 70% of coral reefs globally are unhealthy. If current trends continue, all coral reefs worldwide could be bleached by the end of the century. To offer early warnings, assess harmful factors, and ensure the resilience of coral reef ecosystems, continuous monitoring of their benthic communities is essential. During this monitoring process, accurate coral recognition is crucial, as scientists use annotated coral data for statistical analyses of coral habitat coverage, spatial relationships, and health status (Beijbom et al. 2012).

With advancements in deep learning, computer vision, and AUV, automated coral recognition has been increasingly applied in underwater coral monitoring (Williams et al. 2019; Teague et al. 2022). While state-of-the-art methods have shown commendable accuracy, they are usually designed to predict a specific, predetermined set of coral taxa. Therefore, generalization remains one of the most significant barriers to applying AI in coral reef protection. Complex marine environments, characterized by factors such as temperature, light, and salinity, vary significantly across different locations and consequently support diverse coral species. The substantial morphological differences among these coral species primarily rely on expert knowledge and make existing classification methods difficult to generalize to new, unseen coral species in unexplored oceans.

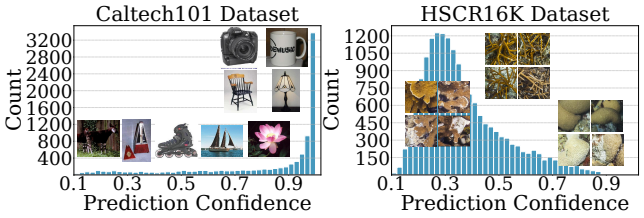

Figure 1: Histograms of confidence scores from zero-shot predictions of CLIP on datasets Caltech101 and HSCR16K. While CLIP is confident for online images in Caltech101 (left), its performances are far from satisfying for classifying coral images (right).

Recently, vision-language models (VLMs) pre-trained on hundreds of millions of paired image-text data, such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), have demonstrated extraordinary generalization capabilities on aligning image and text modalities in new domains. They enable zero-shot or few-shot learning for a wide range of image classification tasks (Pourpanah et al. 2022). However, while performing superb in general purpose online images (see the distribution of the confidence scores in Fig. 1.*left*), they are not directly applicable for classifying coral images due to unique coral features (see Fig. 1.*right*). Due to two major challenges, the first is the **Lack of a coral-specific recognition dataset**. Current coral-related datasets neither follow biological taxonomy (from *Kingdom* and *Phylum* to *Genus* and *Species*) nor contain rich stony coral species which are more vital for ocean health for serving as habitats,

*Corresponding author

breeding sites, and feeding grounds for thousands of marine species (Shihavuddin et al. 2013a). In addition, coral-specific image-text prompts are lacked. Generic text prompt templates in CLIP, such as "a photo of a class name" cannot introduce necessary domain knowledge to highlight detailed differences among coral species (*Case 1* in Fig. 2). The second is the **Absence of Suitable Fine-Tuning Strategies**. Conventional methods for adapting pre-trained models to downstream tasks generally involve fine-tuning the entire model, which is computationally demanding. Emerging PEFT-based CLIP-adapter models (Zhou et al. 2022; Yu et al. 2023; Gao et al. 2024; Huang et al. 2024) produce general visual features and often fail to deal with coral reefs' morphological features and rich hierarchies in their taxonomy (*Case 2* in Fig. 2).

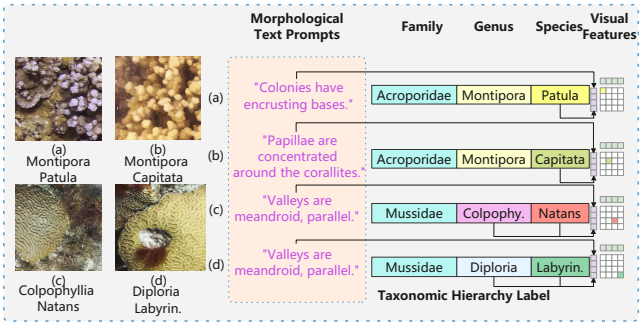

Figure 2: Case 1: *Montipora Patula* (a) and *Montipora Capitata* (b) exhibit noticeable differences in texture details. But CLIP's text prompt templates fail to describe them. Case 2: *Colpophyllia Natans* (c) and *Diploria Labyrin.* (d) have similar morphological features, the taxonomic information (from *Genus* to *Family*) can help distinguish them. The "Diploria Labyrin." indicates species *Diploria Labyrinthiformis*. A detailed figure can be found in the appendix.

To address the above challenges, we first introduce the large-scale **H**ierarchical **S**tony **C**oral **R**ecognition dataset (HSCR16K) which contains 16,659 stony coral images from 16 species organized from the *Kingdom* to the *Species* according to coral biological taxonomy. These stony corals were from 13 marine regions across the Pacific Ocean and the Atlantic Ocean. The rich species/region diversity make it useful in global coral protection research. Our coral dataset also features the development of coral-specific text prompts. Marine scientists in our team design multiple sets of morphological prompts for each species, for example, "Structure of *Acropora palmata*: has parallel, obliquely inclined, very thick tapered branches. Branches are horizontally flattened towards their extremities. Corallites are tubular and irregular in length.". We thus propose CORAL-Adapter, an adaptive fine-tuning method for CLIP which comprises two types of models: the morphological adapter and the taxonomic adapter. Specifically, by linking the morphological prompts with the captured visual features, the morphological adapter differentiates fine-grained visual features of different species. However, we found that in some challenging cases, morphological features alone are insufficient for distinguishing between certain species, such as species *Colpophyllia Natans* and species *Diploria Labyrinthiformis* (see Fig. 2). These two coral species have similar morphological features but belong to different genera. In addition to species labels, corals are interconnected within a comprehensive taxonomy. Even if morphological features are similar, distinguishable representations for classification can be learned from taxonomic hierarchy (from *Genus* to *Family*). To address this, the taxonomic adapter is designed to further align visual representations with hierarchical taxonomic structures. Our contributions are as follows:

- We release HSCR16K, the first hierarchical coral dataset labeled to the finest species level, for the stony coral recognition. It contains 16,659 images of 16 species across 13 marine regions with rich knowledge, establishing a benchmark for subsequent work and future studies.
- We first attempt to apply VLMs to coral recognition tasks and present CORAL-Adapter, a novel framework to integrate two complementary kinds of coral-specific knowledge (biological taxonomy and coral morphology) with general knowledge learned by CLIP. Furthermore, CORAL-Adapter is simple with only a few parameter updates and can be applied as a plug-and-play module to various CLIP-based methods.
- We conduct a series of coral recognition tasks on the dataset. CORAL-Adapter achieves strong performances in both zero-shot and few-shot settings, indicating the effectiveness and potential of our method.

Related Work

Related Datasets

Coral datasets are fundamental for studying coral conservation. Several coral datasets and benchmarks (*e.g.,* TasCPC (Meyer et al. 2011), MLC (Beijbom et al. 2012), RSMAS (Shihavuddin et al. 2013a), EILAT (Shihavuddin et al. 2013b), Benthoz15 (Bewley et al. 2015) and ATCRC (Rashid and Chennu 2020), Mask3K (Li et al. 2020), WaterMask (Lian et al. 2023), LaRS (Žust, Perš, and Kristan 2023), USOD10K (Hong et al. 2023), CoralSCOP (Zheng et al. 2024), Marineinst (Zheng et al. 2025)) have been proposed successively. Most of these datasets (WaterMask, USOD10K, CoralSCOP, Marineinst, EILAT, MLC) are not labeled to the finest species level and tend to categorize the same coral family or genus as a coral class. RSMAS, Benthoz15 and ATCRC contain very few stony coral species, but include a large number of non-corals like algae, fish, and other invertebrates. By contrast, our dataset provides 16 stony coral species over 16,659 images across 13 diverse marine regions, organized by taxonomic rank. In addition, rich language knowledge about morphology are provided. Please refer to Tab. 1 for comparisons.

Coral Recognition and Protection

Automated recognition of specific coral species enables scientists to continuously monitor and analyze their health, which is crucial for effectively preserving coral reef ecosystems. Earlier recognition methods depended on hand-crafted

Datasets	Img.	Spec.	Reg.	Tax.	Pmt.
TasCPC	1,258	*	1	✗	✗
MLC	2,055	*	1	✗	✗
EILAT	1,123	*	1	✗	✗
RSMAS	766	7	Not Spec.	✗	✗
Benthoz15	Not Ava.	1	1	✗	✗
ATCRC	147	8	1	✗	✗
CoralSCOP	41,297	*	Not Spec.	✗	✗
HSCR16K	**16,659**	**16**	**13**	✓	✓

[1] * : Species are not following standard Linnaean system.
[2] Tax.: the dataset is organized by coral taxonomy.
[3] Pmt.: the dataset includes the text prompt.

Table 1: The comparisons of existing coral datasets. The "Img.", "Spec.", and "Reg." indicate the number of images, the number of coral species, and the number of marine regions, respectively.

features and traditional machine-learning techniques (Stokes and Deane 2009; Beijbom et al. 2012). The rapid proliferation of deep learning brings methodological shifts. Mahmood et al. (2016) extract CNN-based features at multiple scale for coral recognition. Firdous and Sabena (2023) integrate two different CNN models in coral reef recognition. However, these deep models are typically trained to predict a fixed set of predetermined coral taxa, which limits their generalization and usability, as different coral species are distributed across various regions of the oceans.

Vision-Language Model Adaption

Visual-language models, such as CLIP (Radford et al. 2021) and BioCLIP (Stevens et al. 2024), have demonstrated remarkable generalization ability. Instead of fine-tuning the entire large model, some adaption-style methods have been proposed for VLMs on downstream tasks, such as scene classification, action recognition and land cover classification with satellite images. They introduce lightweight modules with a small number of trainable parameters. For instance, CLIP-Adapter (Gao et al. 2024) adopts an additional bottleneck layer to learn new features integrated with CLIP's zero-shot prior knowledge. TaskRes (Yu et al. 2023) keeps the original text-based classifier and learns a new classifier for the target task by fine-tuning a set of prior-independent parameters as a residual to the original one. LP++ (Huang et al. 2024) generalizes the standard linear-probe classifier by integrating text knowledge from the learnable text embedding with class-wise multipliers blending image and text features. However, the unique coral characteristics of corals (*e.g.,* texture, color, and structure), influenced by complex marine environments, pose significant challenges in leveraging VLMs for coral species recognition.

Dataset

Data Collection

Our coral images are sourced from the XL Catlin Seaview Survey Project (González-Rivero et al. 2019). The project has been dedicated to assessing potential ocean risks since 2009. This project provides a vast collection of raw underwater images from different oceans, including corals, algae, fish, and invertebrates. Among them, 2,560 raw images are from Pacific Ocean and Atlantic Ocean, which are the only two oceans providing detailed species information. We segment these 2,560 images into patches based on the coordinates of each object, resulting in a total of 174,971 patch images. Subsequently, we filter out images using the following criteria: 1) non-coral images like algae, fish, and rubbish; 2) soft coral images; 3) stony coral images without species-level label (species labels mixes with higher-order taxa); 4) the low-quality coral images according to Underwater Color Image Quality Evaluation (UCIQE) (Yang and Sowmya 2015) and Underwater Image Quality Measurement (UIQM) (Panetta, Gao, and Agaian 2015). The processed dataset contains the remaining data after removing entries through steps (1)-(4). Then, marine biologists reorganize and relabel the remaining coral images following biological taxonomy (*Kingdom* down to *Species*). We hope that the inclusion of biological taxonomy data will aid in the creation of a precise and robust recognition system. Finally, marine biologists perform the manual double-check, *e.g.,* eliminating the repeated images or mislabeled images. In this way, we build HSCR16K dataset with 16,659 images including 16 stony coral species, 10 genera, and 8 families across 13 marine regions (Nordquist 2011). Fig. 3 shows treemap of HSCR16K (*Order* through *Family*). Further statistical details of the dataset, including class distributions, can be found in the appendix A.

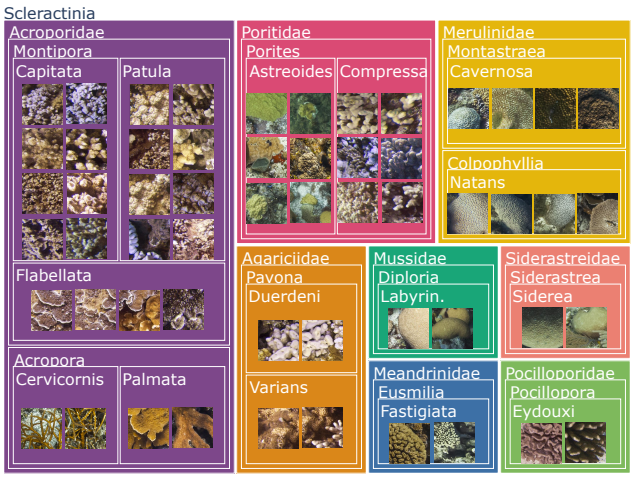

Figure 3: Treemap of HSCR16K. Different colors show the different stony coral families. Nested boxes of the same color represent family, genus, and species, respectively. The size of the boxes reflects the relative number of images.

Coral-Specific Prompts

For coral recognition, a meaningful text description that integrates coral-specific knowledge is necessary. However, the general text prompt templates can not introduce the domain knowledge necessary, such as "a photo of a class name", "a

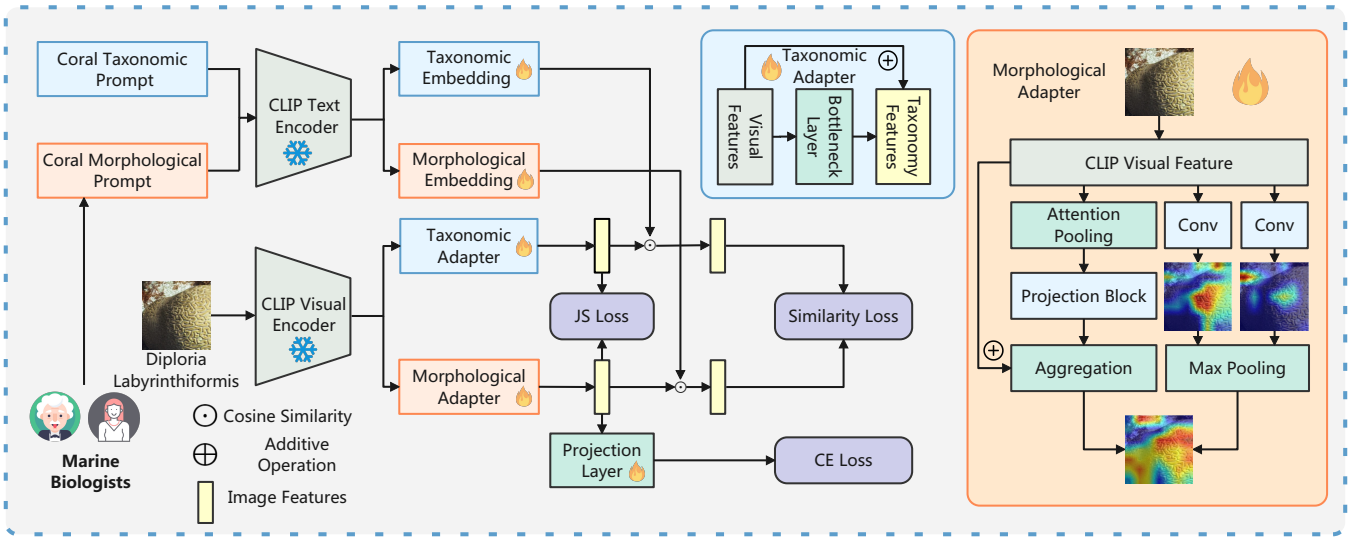

Figure 4: The architecture of CORAL-Adapter, which integrates two complementary kinds of coral-specific knowledge (biological taxonomy and coral morphology) with general knowledge learned by CLIP. Tab. 2 presents the detailed taxonomic prompt and morphological prompt for the coral species *Diploria Labyrinthiformis*.

centered satellite photo of class name", "a photo of a person doing action name" (Zhang et al. 2022; Gao et al. 2024; Yu et al. 2023; Yang et al. 2024). We collaborate with marine biologists to construct various sets of prompts from multiple aspects, including coral surface texture, color, structure, and coral taxonomic description. We take the standard seven-level biology taxonomy (*Kingdom*, *Phylum*, *Class*, *Order*, *Family*, *Genus*, and *Species*) into account, and combine the biology taxonomy label to build the taxonomic prompt. The coral taxonomic prompt and coral morphological prompt are shown in Tab. 2.

Text Type	Example
Taxonomy	*Animalia, Cnidaria, Anthozoa, Scleractinia, Faviidae, Diploria, labyrinthiformis*
Morphology	**Structure**: Colonies are massive and usually hemispherical. Columellae are fine and do not form distinct centres. **Color**: Tan to yellowish- or grey-brown. **Texture**: Valleys are meandroid, parallel or sinuous, deep. Ambulacral grooves vary greatly in width within the same colony but may be wider than the valleys giving the superficial appearance of alternating valleys of two different sorts.

Table 2: Text types of coral-specific prompts.

Method

Model Architecture

CLIP CLIP (Radford et al. 2021) consists of a text encoder and an image encoder. In the training process, the visual representation from image encoder and the text embedding from text encoder are aligned in a new embedding space using the cosine distance. In detail, given a query image y and a hand-crafted template "a photo of [CLASS]", let $f \in \mathbb{R}^{D\times 1}$ be the visual representation and $\{w_i\}_{i=1}^{N}, w_i \in \mathbb{R}^{D\times 1}$ be the text embedding. The predicted probability of query image y for class i is formulated as

$$\text{logits}\,(y_c = i) = \frac{w_i^\top f}{\|w_i\|\,\|f\|}. \tag{1}$$

CORAL-Adapter As presented, CLIP cannot introduce additional coral-specific knowledge necessary. To this end, we propose CORAL-Adapter comprising two key components: the taxonomic adapter and the morphological adapter (see Fig. 4). These two components are optimized by a combination of three different kinds of loss, and the overall loss $\mathcal{L}$ can be written as

$$\mathcal{L} = \mathcal{L}_{\text{sim}} + \lambda_1 \mathcal{L}_{\text{js}} + \lambda_2 \mathcal{L}_{\text{ce}}, \tag{2}$$

where $\mathcal{L}_{\text{sim}}$ is the similarity loss between predicted logits $logits_m$ from morphological adapter and predicted logits $logits_t$ from the taxonomic adapter; $\mathcal{L}_{\text{js}}$ is the reversed Jensen–Shannon (JS) divergence loss and $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss between label and predicted logits from projection layer. λ_1 and λ_2 are the trade-off parameters.

The predicted logits $logits_m$ with the proposed morphological adapter can be formulated as

$$\begin{aligned} \text{logits}_{\text{m}}\,(y_c = i) &= \frac{\hat{w_{mi}}^\top \hat{f_m}}{\|\hat{w_{mi}}\|\,\|\hat{f_m}\|}, \\ \text{where } \hat{f_m} &= \text{MorpAdapter}(f, f_1) \\ \hat{w_m} &= \alpha w + (1-\alpha)\theta, \end{aligned} \tag{3}$$

where f and f_1 be visual features of the last two layers of CLIP visual encoder. θ is the learnable text embedding and α is the trade-off parameter between prior knowledge and

newly learned knowledge. $\hat{w_m}$ is the learnable morphological text embedding, and $\hat{f_m}$ is the morphological visual representation. Similar to the morphological adapter, the predicted logits $logits_t$ from the taxonomic adapter can also be derived. The output of the morphological adapter is further fed into a projection layer. The morphological adapter and the projection layer are jointly optimized through the cross-entropy loss $\mathcal{L}_{\text{ce}}$, which will further reduce the bias between the predicted logits $logits_m$ and the true label y_t.

To avoid over-reliance on prior knowledge from pre-trained CLIP model (Yu et al. 2023), we design the reversed divergence loss $\mathcal{L}_{\text{js}}$. It can amplify the differences between taxonomic adapter and morphological adapter, and guide the model to learn new knowledge. $\mathcal{L}_{\text{js}}$ can be defined as

$$\begin{aligned}\mathcal{L}_{\text{js}} = &-\frac{1}{2}\sum \hat{f}_t \log(\frac{\hat{f}_t}{\hat{f}_t - \hat{f_m}}) \\ &+ \frac{1}{2}\sum \hat{f_m} \log(\frac{\hat{f_m}}{\hat{f}_t + \hat{f_m}}),\end{aligned} \tag{4}$$

where $\hat{f}_t$ and $\hat{f_m}$ are visual representations from the taxonomic adapter and morphological adapter, respectively.

Morphological Adapter

We design the morphological adapter to capture morphological representation including texture, structure and color from coral species. Let f and f_1 be visual features of the last two layers of CLIP visual encoder. By computing the outer product of feature vectors (Kim et al. 2016) between two different layers, the morphological adapter can capture second-order relationships between different texture features within local regions, enable it to learn complex patterns and subtle differences in coral texture. Max pooling of the outer product of feature vectors can further preserve key texture information of corals and highlight texture differences among various coral species. The aforementioned procedure can be written as

$$\begin{aligned}z(f, f_1) &= f^T W_m f_1 \\ t(z) &= maxpool(z),\end{aligned} \tag{5}$$

where W_m is the learnable projection matrix, z is the output of the outer product of feature vectors and t is the result of the *maxpool* operation. Additionally, global structural information and the correlation between structures are considered crucial for distinguishing coral species. To this end, we propose a learnable structure block that integrates non-local networks (Wang et al. 2018). Given f the visual features of last layer from CLIP visual encoder, the global structure information can be obtained by

$$s(f) = F(f_i, \delta(\sum_{j=1}^{N} \beta_j f_j)), \tag{6}$$

where N is the number of groupings of the coral image based on the structure position, $\sum_{j=1}^{N} \beta_j f_j$ represents the coral structure modeling block. This block groups the features of all coral's structure using the global attention pooling with weight β to extract the global structure representations. In Eq. (6), δ is the projection block used to get channel-wise dependencies for capturing the subtle variations in coral color, and F integrates the global structure representations with the features of local coral structure. Finally, the morphological adapter can be formulated as

$$MorpAdapter(f, f_1) = maxpool(z(f, f_1)) + s(f). \tag{7}$$

Morphological adapter links the textual prompts of coral morphology with the visual representations of the coral's texture, structure and color, enabling effective adaptation to coral recognition tasks.

Taxonomic Adapter

To integrate complementary knowledge from the taxonomic hierarchy, we design the taxonomic adapter. Inspired from the CLIP-Adapter (Gao et al. 2024), the taxonomic adapter consists of a bottleneck block $z_t(f)$ and the residual connection, where $z_t(f)$ can be formulated as

$$z_t(f) = \text{ReLU}\left(f^T W_t^1\right) W_t^2, \tag{8}$$

where W_t^1 and W_t^2 are the weights of the bottleneck block, f is the visual representation from CLIP visual encoder. The newly learned taxonomic hierarchy knowledge is added with the CLIP's prior knowledge visa the residual connection. Finally, the taxonomic adapter with the residual connection can be written as

$$TaxoAdapter(f) = \alpha z_t(f)^T + (1 - \alpha) f, \tag{9}$$

where α is the residual ratio for balancing the newly learned taxonomic knowledge and prior knowledge from CLIP.

Experiments

Implementation Details All comparison models are built upon the pre-trained CLIP model. We choose the ResNet50 (He et al. 2016) as CLIP visual encoder, and the transformer (Dosovitskiy et al. 2020) as CLIP text encoder. CORAL-Adapter is trained for 100 epochs on all experiments with a batch size of 64, using the AdamW (Kingma and Ba 2014) optimizer and a learning rate of 0.001. The average results over three runs are reported for comparison. Our dataset, code, and appendix are publicly available at https://doi.org/10.6084/m9.figshare.26702314.v1.

Baselines We compare our CORAL-Adapter with Zero-shot CLIP (ZS CLIP) (Radford et al. 2021), and other state-of-the-art PEFT methods including CoOp (Zhou et al. 2022), CLIP-Adapter (Gao et al. 2024), TaskRes (Yu et al. 2023), LP++ (Huang et al. 2024). Further baseline methods (He et al. 2016; Zhang et al. 2022; Wu et al. 2024) and evaluation metrics (Micro Accuracy, Macro Accuracy and Macro F1 Score) are provided in the appendix B.

Few-Shot Learning

Coral Species Recognition We randomly split the HSCR16K dataset in a 6:2:2 ratio into training, validation, and testing sets. Following the few-shot setting of CoOp (Zhou et al. 2022), we train our CORAL-Adapter under the few-shot setting of 1,2,4,8,16 shots (*i.e.,* number of "shots" per class) and then test the model on full

Methods	×0	×1	×2	×4	×8	×16
ZS CLIP	13.5	-	-	-	-	-
CoOp	-	18.8	18.4	48.9	56.6	63.8
CLIP-Adapter	-	16.1	19.5	20.6	25.9	42.4
TaskRes	-	19.2	26.6	37.4	40.8	46.8
LP++	-	29.5	31.6	55.4	50.7	60.7
Ours	-	**41.4**	**50.1**	**64.8**	**66.2**	**77.3**

Table 3: Performance comparison of our CORAL-Adapter with the SOTA methods on coral species recognition, including zero-shot, 1-/2-/4-/8-/16-shots. The "ZS CLIP" indicates the Zero-shot CLIP.

test splits. In coral species recognition, samples for some coral species may be scarce or difficult to obtain. Few-shot learning enables effective recognition with just a small numberisualization results and analysis. In of labeled samples (*e.g.,* from 1 to 16 images for each coral species), reducing the need for extensive labeled data. The detailed performances are presented in Tab. 3. CORAL-Adapter clearly shows remarkable performance over all other methods on all shot settings. Specifically, Zero-shot CLIP is not suitable for this task (only 13.5% accuracy), while CORAL-Adapter achieves a 27.9% improvement even under extreme condition of 1-shot. Compared with other adaption methods, CORAL-Adapter achieves huge performance boosts, *e.g.*, ranging from 10% to 20% in the 1-shot setting and from 10% to 40% in the 8-shot setting. In summary, thanks to the learned taxonomic relationships and morphological representation, CORAL-Adapter exhibits strong coral recognition capabilities.

Coral Genera Recognition Our hierarchical dataset provides detailed taxonomic information for each species, and thus supports genera or family recognition to offer a macro perspective on the overall coral community (Madin et al. 2016). In this section, we aim to validate whether CORAL-Adapter can effectively achieve high-level recognition and meet various real-world needs. We re-organize the HSCR16K dataset according to the coral genera and re-split the dataset in a 6:2:2 ratio into training, validation, and testing sets. Following the few-shot setting from the coral species recognition, we present the genera recognition results in Tab. 4. Notably, just one coral sample is sufficient for CORAL-Adapter to achieve impressive performance for coral genera recognition, and it exhibits a remarkable 25.6% improvement over CoOp and an 11.9% improvement compared to LP++, a latest adaptation method. In the 16-shot setting, CORAL-Adapter achieves the improvements of +9.6%, +22.8%, +24.8%, and +14.9% over other state-of-the-art methods, respectively. The experiments on coral genera recognition demonstrate that the strategy used in CORAL-Adapter, which integrates morphological information with coral taxonomic relationships, is also effective for recognizing different levels of coral taxonomy and adapting to various coral protection tasks. Appendix B has more qualitative results about family-level recognition.

Methods	×0	×1	×2	×4	×8	×16
ZS CLIP	15.1	-	-	-	-	-
CoOp	-	11.5	46.9	18.3	57.9	66.1
CLIP-Adapter	-	20.2	21.5	22.9	30.9	52.9
TaskRes	-	20.9	33.1	35.4	45.6	50.9
LP++	-	25.2	40.0	46.4	51.5	60.8
Ours	-	**37.1**	**58.9**	**70.0**	**70.0**	**75.7**

Table 4: Performance comparison of our CORAL-Adapter with the SOTA methods on coral genera recognition, including zero-shot, 1-/2-/4-/8-/16-shots. The "ZS CLIP" indicates the Zero-shot CLIP.

Zero-Shot Learning

Generalization to Unseen Bleaching-Prone Coral Species Coral species that are prone to bleaching, such as *Acropora palmata* and *Acropora cervicornis*, are typically very sensitive to environmental changes like high temperatures and pollution. Widespread bleaching and subsequent death of these coral species could result in the degradation of coral reefs, which may ultimately lead to the disruption of the marine ecosystem. Therefore, the generalization capability to unseen bleaching-prone coral species has crucial practical significance for coral preservation. According to the setting of the existing work (Marshall and Baird 2000; Pisapia, Burn, and Pratchett 2019; Burn et al. 2023), we split our proposed dataset into two groups: other coral species (base classes) and bleaching-prone coral species (novel classes). Following the previous dataset splitting method and the evaluation approach for generalization performance used by CoOp (Zhou et al. 2022), we train our model on the 16-shot training set from other coral species and test the generalization performance of trained models on testing set from bleaching-prone coral species. Tab. 5 shows the generalization experiment results. CORAL-Adapter achieves the improvements of +50.8%, +14.2%, +19.8%, +10.8% compared to other state-of-the-art adaptation methods, respectively.

Generalization to Unexplored Ocean In addition, the ability to generalize coral recognition across different regions or oceans allows scientists to monitor corals globally, enabling the early detection of signs of coral bleaching, disease, or other changes. Therefore, we divide our proposed dataset into two subsets: the Pacific dataset (base classes) and the Atlantic dataset (novel classes). There is no overlap in coral species between the Pacific dataset and the Atlantic dataset. Following the previous generalization to unseen bleaching-prone coral species setting, we train our model on coral species from the Pacific dataset and test the generalization performance of trained models on the testing set from the Atlantic dataset. As shown in the Tab. 5, the experimental results demonstrate outstanding generalization performance of CORAL-Adapter, which surpasses the other methods (Zhou et al. 2022; Gao et al. 2024; Yu et al. 2023; Huang et al. 2024) with gains of up to +40.3%, +38.6%, +34.6%, +19.8%, respectively.

Methods	Unseen species		Unexplored ocean	
	Base	Novel	Base	Novel
CoOp	89.0	9.7	60.3	10.1
CLIP-Adapter	59.6	46.3	35.1	11.8
TaskRes	75.7	40.7	52.3	15.8
LP++	79.6	49.7	57.8	30.6
Ours	**85.2**	**60.5**	**76.7**	**50.4**

Table 5: Performance comparison of our CORAL-Adapter with the SOTA methods in terms of generalization to unseen bleaching-prone coral species and unexplored ocean. The models are trained on the set of base classes and evaluated on the novel classes.

Method Analysis

The Effectiveness of Coral-Customized Prompts Our coral-customized prompts bring domain knowledge into the model. Such knowledge may be useful to models other than CORAL-Adapter. As shown in Tab. 6, without any training sample, Zero-shot CLIP with coral-customized prompts gains up to +5.68% and +26.58% on coral species recognition and coral genera recognition over the original Zero-shot CLIP. Grad-CAM (Selvaraju et al. 2017) can visualize and highlight which parts of an image are most influential in a model's decision-making process. Fig. 5 presents Grad-CAM visualizations showing the attention regions of different methods. Compared to Zero-shot CLIP, Zero-shot CLIP with our coral text prompts emphasizes the texture and structure of corals more. While it may not capture texture and structure features as comprehensively as CORAL-Adapter, Grad-CAM still indicates that our prompts are effective.

Method	Species	Genera
Zero-shot CLIP	13.46	15.14
Zero-shot CLIP w/prompt	19.14	41.72
CORAL-Adapter w/prompt	**77.27**	**75.74**

Table 6: Analysis on our designed prompt. The "prompt", "Species" and "Genera" indicate the coral-customized prompts, coral species recognition and coral genera recognition, respectively.

The Effectiveness of Different Components We further conduct ablation studies to analyze the benefits of CORAL-Adapter. On the 16-shot setting, we explore the effectiveness of different components on few-shot (coral species recognition) and zero-shot (unseen bleaching-prone coral species) tasks. As shown in Tab. 7, compared to the taxonomic adapter, the morphological adapter contributes more significantly to enhancing accuracy. CORAL-Adapter with taxonomic adapter and morphological adapter can achieve the best performance. In addition, our proposed architecture demonstrates successful generalization to other domains, including flower recognition (Nilsback and Zisserman 2008) and texture classification (Sharan, Rosenholtz, and Adelson 2014). Detailed results are presented in the appendix B.

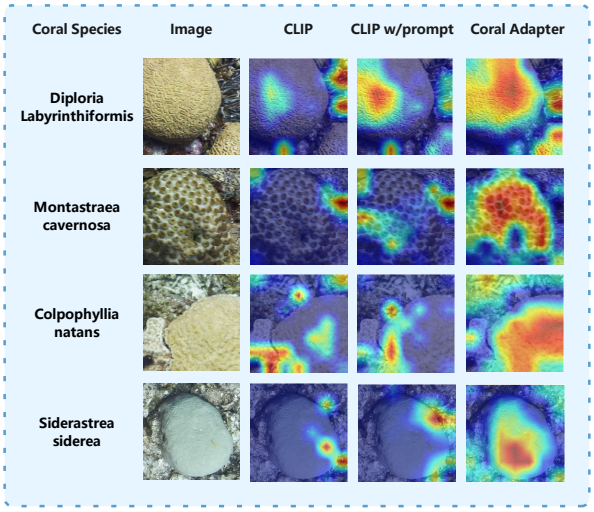

Figure 5: Grad-CAM visualizations on HSCR16K dataset for Zero-shot CLIP, Zero-shot CLIP with coral-customized prompt, and CORAL-Adapter.

Method	Few-shot	Zero-shot
CORAL-Adapter w/o morp	28.88	48.55
CORAL-Adapter w/o taxo	46.50	54.45
CORAL-Adapter	**77.27**	**60.50**

Table 7: Ablation study on different components. The "taxo" and "morp" indicate the taxonomic adapter, and morphological adapter, respectively. The "Few-shot" and "Zero-shot" indicate the coral species recognition and generalization to unseen bleaching-prone coral species, respectively.

Conclusion

Automated coral species recognition based on AI techniques such as pre-trained large vision-language models would significantly accelerate the coral research and protection. This paper explores the challenges involved in adapting the pre-trained CLIP model for coral species recognition. To address these challenges, we compiled HSCR16K–a diverse hierarchical coral dataset, and designed CORAL-Adapter–a parameter-efficient fine-tuning model. CORAL-Adapter integrates a morphological adapter and a taxonomic adapter into CLIP, infusing the model with coral-specific knowledge. Extensive experiments across coral-related tasks show that CORAL-Adapter outperforms the competitive baselines, with updating only a few additional parameters in the total CLIP parameters. Our work thus provides valuable insights into leveraging large vision-language models to advance sustainable coral research and protection.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under grants 62076232 and 62172049, and Science and Technology Development Foundation of South China Sea Bureau, Ministry of Natural Resources under grant 230208. We sincerely thank the marine biologists in the fourth author's research team (Nansha Islands Coral Reef Ecosystem National Observation and Research Station) for their help on data collection and processing.

References

Beijbom, O.; Edmunds, P. J.; Kline, D. I.; Mitchell, B. G.; and Kriegman, D. 2012. Automated annotation of coral reef survey images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1170–1177. IEEE.

Bewley, M.; Friedman, A.; Ferrari, R.; Hill, N.; Hovey, R.; Barrett, N.; Marzinelli, E. M.; Pizarro, O.; Figueira, W.; Meyer, L.; et al. 2015. Australian sea-floor survey data, with images and expert annotations. *Scientific Data*, 2(1): 1–13.

Burn, D.; Hoey, A.; Matthews, S.; Harrison, H.; and Pratchett, M. 2023. Differential bleaching susceptibility among coral taxa and colony sizes, relative to bleaching severity across Australia's Great Barrier Reef and Coral Sea Marine Parks. *Marine Pollution Bulletin*, 191: 114907.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elawady, M. 2015. Sparse coral classification using deep convolutional neural networks. *arXiv preprint arXiv:1511.09067*.

Firdous, R. J.; and Sabena, S. 2023. Collaborative CNN with Multiple Tuning for Automated Coral Reef Classification. In *Proceedings of the IEEE International Conference on Computer, Communication, and Signal Processing*, 81–93. IEEE.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

González-Rivero, M.; Rodriguez-Ramirez, A.; Beijbom, O.; Dalton, P.; Kennedy, E. V.; Neal, B. P.; Vercelloni, J.; Bongaerts, P.; Ganase, A.; Bryant, D. E.; et al. 2019. Seaview Survey Photo-quadrat and Image Classification Dataset. https://espace.library.uq.edu.au/data/UQ_734799/. Online; accessed 10 December 2024.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. IEEE.

Hong, L.; Wang, X.; Zhang, G.; and Zhao, M. 2023. USOD10K: a new benchmark dataset for underwater salient object detection. *IEEE Transactions on Image Processing*.

Huang, Y.; Shakeri, F.; Dolz, J.; Boudiaf, M.; Bahig, H.; and Ben Ayed, I. 2024. LP++: A Surprisingly Strong Linear Probe for Few-Shot CLIP. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 23773–23782. IEEE.

ICRAN. Online; accessed 25-May-2014. Coral Reefs: Ten Questions - Ten Answers. http://www.icran.org/peoplereefs-tenquestions.html.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, 4904–4916. PMLR.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, L.; Rigall, E.; Dong, J.; and Chen, G. 2020. MAS3K: An open dataset for marine animal segmentation. In *Proceedings of the International Symposium on Benchmarking, Measuring and Optimization*, 194–212. Springer.

Lian, S.; Li, H.; Cong, R.; Li, S.; Zhang, W.; and Kwong, S. 2023. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, 1305–1315. IEEE.

Madin, J. S.; Hoogenboom, M. O.; Connolly, S. R.; Darling, E. S.; Falster, D. S.; Huang, D.; Keith, S. A.; Mizerek, T.; Pandolfi, J. M.; Putnam, H. M.; et al. 2016. A trait-based approach to advance coral reef science. *Trends in Ecology & Evolution*, 31(6): 419–428.

Mahmood, A.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F.; Hovey, R.; Kendrick, G.; and Fisher, R. B. 2016. Coral classification with hybrid feature representations. In *Proceedings of the IEEE International Conference on Image Processing*, 519–523.

Marshall, P.; and Baird, A. 2000. Bleaching of corals on the Great Barrier Reef: differential susceptibilities among taxa. *Coral reefs*, 19: 155–163.

Meyer, L.; Hill, N.; Walsh, P.; and Barrett, N. 2011. Methods for the processing and scoring of auv digital imagery from south eastern tasmania. *Institute for Marine and Antarctic Studies Internal Report*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Nordquist, M. 2011. *United Nations Convention on the law of the sea 1982, Volume VII: a commentary*, volume 7. Brill.

Panetta, K.; Gao, C.; and Agaian, S. 2015. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3): 541–551.

Pisapia, C.; Burn, D.; and Pratchett, M. 2019. Changes in the population and community structure of corals during recent disturbances (February 2016-October 2017) on Maldivian coral reefs. *Scientific Reports*, 9(1): 8402.

Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C. P.; Wang, X.-Z.; and Wu, Q. J. 2022. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4051–4070.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the IEEE International Conference on Machine Learning*, 8748–8763. IEEE.

Rashid, A. R.; and Chennu, A. 2020. A trillion coral reef colors: Deeply annotated underwater hyperspectral images for automated classification and habitat mapping. *Data*, 5(1): 19.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. IEEE.

Sharan, L.; Rosenholtz, R.; and Adelson, E. H. 2014. Accuracy and speed of material categorization in real-world images. *Journal of vision*, 14(9): 12–12.

Shihavuddin, A.; Gracias, N.; Garcia, R.; Gleason, A. C.; and Gintert, B. 2013a. Image-based coral reef classification and thematic mapping. *Remote Sensing*, 5(4): 1809–1841.

Shihavuddin, A.; Gracias, N.; Garcia, R.; Gleason, A. C.; and Gintert, B. 2013b. Image-based coral reef classification and thematic mapping. *Remote Sensing*, 5(4): 1809–1841.

Stevens, S.; Wu, J.; Thompson, M. J.; Campolongo, E. G.; Song, C. H.; Carlyn, D. E.; Dong, L.; Dahdul, W. M.; Stewart, C.; Berger-Wolf, T.; et al. 2024. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 19412–19424. IEEE.

Stokes, M. D.; and Deane, G. B. 2009. Automated processing of coral reef benthic images. *Limnology and Oceanography: Methods*, 7(2): 157–168.

Teague, J.; Megson-Smith, D. A.; Allen, M. J.; Day, J. C.; and Scott, T. B. 2022. A review of current and new optical techniques for coral monitoring. *Oceans*, 3(1): 30–45.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803. IEEE.

Williams, I. D.; Couch, C. S.; Beijbom, O.; Oliver, T. A.; Vargas-Angel, B.; Schumacher, B. D.; and Brainard, R. E. 2019. Leveraging automated image analysis tools to transform our capacity to assess status and trends of coral reefs. *Frontiers in Marine Science*, 6: 222.

Wu, G.; Chen, J.; Zhang, W.; and Wang, R. 2024. Feature Adaptation with CLIP for Few-shot Classification. In *Proceedings of the ACM International Conference on Multimedia in Asia*, MMAsia '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702051.

Yang, L.; Zhang, R.-Y.; Wang, Y.; and Xie, X. 2024. MMA: Multi-Modal Adapter for Vision-Language Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 23826–23837. IEEE.

Yang, M.; and Sowmya, A. 2015. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12): 6062–6071.

Yu, T.; Lu, Z.; Jin, X.; Chen, Z.; and Wang, X. 2023. Task residual for tuning vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10899–10909. IEEE.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Proceedings of the IEEE European Conference on Computer Vision*, 493–510. Springer.

Zheng, Z.; Chen, Y.; Zeng, H.; Vu, T.-A.; Hua, B.-S.; and Yeung, S.-K. 2025. MarineInst: A Foundation Model for Marine Image Analysis with Instance Visual Description. In *Proceedings of the IEEE International Conference on Computer Vision*, 239–257. Springer.

Zheng, Z.; Liang, H.; Hua, B.-S.; Wong, Y. H.; Ang, P.; Chui, A. P. Y.; and Yeung, S.-K. 2024. CoralSCOP: Segment any COral Image on this Planet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 28170–28180. IEEE.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Žust, L.; Perš, J.; and Kristan, M. 2023. Lars: A diverse panoptic maritime obstacle detection dataset and benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 20304–20314. IEEE.