SeafloorAI: A Large-scale Vision-Language Dataset for Seafloor Geological Survey

Kien X. Nguyen[1], **Fengchun Qiao**[1], **Arthur Trembanis**[2], **Xi Peng**[1]
[1]Deep-REAL Lab, Department of Computer and Information Sciences, University of Delaware
[2]School of Marine Science and Policy, University of Delaware
{kxnguyen,fengchun,art,xipeng}@udel.edu

Abstract

A major obstacle to the advancements of machine learning models in marine science, particularly in sonar imagery analysis, is the scarcity of AI-ready datasets. While there have been efforts to make AI-ready sonar image dataset publicly available, they suffer from limitations in terms of environment setting and scale. To bridge this gap, we introduce `SeafloorAI`, the first extensive AI-ready datasets for seafloor mapping across 5 geological layers that is curated in collaboration with marine scientists. We further extend the dataset to `SeafloorGenAI` by incorporating the language component in order to facilitate the development of both *vision*- and *language*-capable machine learning models for sonar imagery. The dataset consists of 62 geo-distributed data surveys spanning 17,300 square kilometers, with 696K sonar images, 827K annotated segmentation masks, 696K detailed language descriptions and approximately 7M question-answer pairs. By making our data processing source code publicly available, we aim to engage the marine science community to enrich the data pool and inspire the machine learning community to develop more robust models. This collaborative approach will enhance the capabilities and applications of our datasets within both fields. Our code repository are available [1] under the CC-BY-4.0 license.

1 Introduction

Seafloor mapping stands at the forefront of marine science, utilizing cutting-edge technologies like multibeam echosounders and side-scan sonar to unveil the hidden complexities of the ocean floor [67, 68]. Beyond scientific research, seafloor mapping is instrumental in identifying potential resources, assessing environmental impacts, and supporting sustainable ocean management practices in the context of the blue economy [42]. However, the current analysis techniques in seafloor mapping are predominantly labor-intensive and reliant on manual interpretation by marine scientists, necessitating hundreds of hours spent meticulously examining data surveys to analyze seabed imagery [66]. This hands-on approach is not only time-consuming but also susceptible to user *subjectivity* and the limitations of individual expertise, thus introducing potential *inconsistencies* in analysis [56].

The integration of machine learning (ML) holds the promise of enhancing efficiency and reliability in seafloor mapping by automating the segmentation and classification tasks [3, 54, 38]. However, the lack of public AI-ready datasets poses a significant challenge in leveraging the full potential of AI technologies for this purpose. While there have been efforts to make AI-ready sonar image datasets publicly available, they suffer from limitations in terms of environment setting and scale. For example, the dataset in [63] was captured in a water tank, which does not accurately represent the ocean's complex conditions. Additionally, other work have only produced small-scale datasets

[1] `https://github.com/deep-real/SeafloorAI`

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

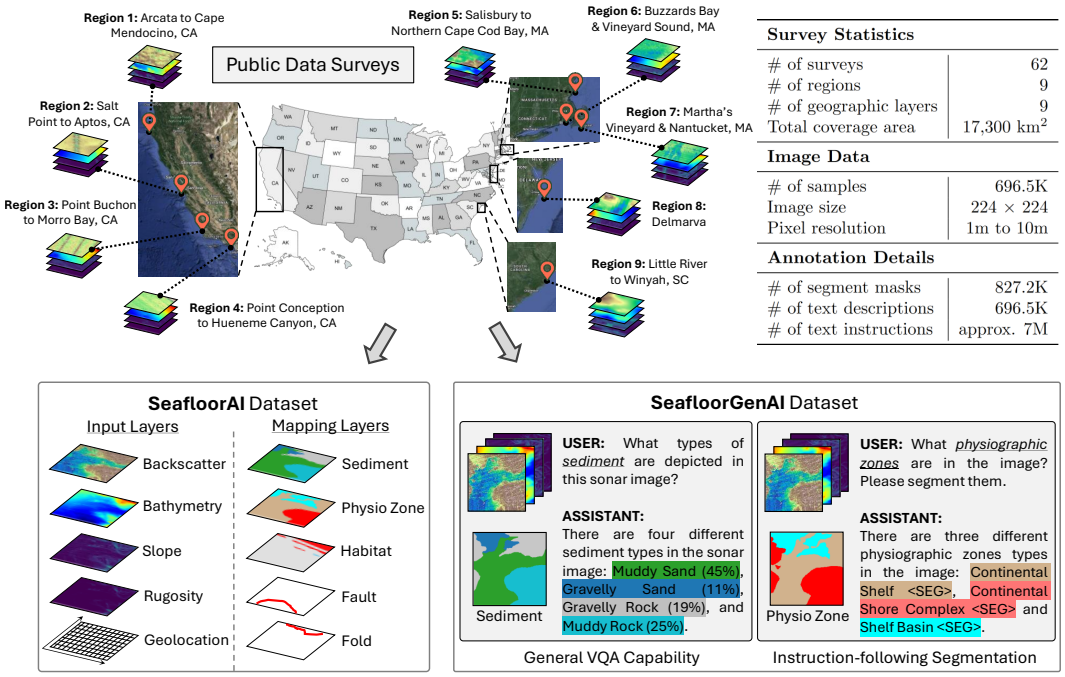


Figure 1: Overview of the spatially distributed seafloor mapping datasets. The table highlights key dataset statistics. We incorporate 62 public data surveys published by USGS and NOAA from 9 major regions to construct `SeafloorAI` and `SeafloorGenAI` datasets. Our dataset contains 9 geological layers, 4 of which are raw signals, *i.e.*, Backscatter, Bathymetry, Slope and Rugosity, and 5 annotated by human experts, *i.e.* Sediment, Physiographic Zone, Habitat, Fault and Fold. `SeafloorAI` serves as a dataset for standard computer vision tasks, *i.e.* semantic segmentation, whereas `SeafloorGenAI` constitutes a dataset for generative vision-language tasks, *i.e.*, general visual question answering and instruction-following mapping. `<SEG>` denotes the segmentation mask output by the model.

with limited area coverage [39, 61], not accounting for the generalizability of the ML models in a spatially distributed setting. On the other end, abundant public hydrographic surveys conducted by the U.S. Geological Survey (USGS) and the National Oceanographic and Atmospheric Administration (NOAA)[2] [17, 48, 49, 47, 50, 2, 5] have yet to be extensively utilized by the ML community.

To bridge this gap, we introduce `SeafloorAI`, the first extensive AI-ready sonar imagery dataset for seafloor mapping. We compiled 62 public hydrographic surveys to construct a large-scale, geo-distributed and multi-purpose dataset, with the effort to map various geological layers. Furthermore, inconsistencies in the nomenclature of geological attributes across data surveys pose a challenge on the unification and development of an extensive dataset. In collaboration with marine scientists, we have developed a framework that standardizes such nomenclature by adopting the Barnhardt classification [6] and the Coastal and Marine Ecological Classification Standard (CMECS) [1]. It guarantees uniformity throughout the dataset, enabling the evaluation of robust ML models in a spatially distributed setting. The data pool currently consists of 696K sonar images, 827K segmentation masks for 5 geological layers: Sediment, Physiographic Zone, Habitat, Fault, and Fold.

Finally, we incorporate the language component into our dataset for the development of generative vision-language models (VLMs) in marine science research. VLMs facilitate seamless interactions through textual queries and provide clear, understandable explanations throughout the analysis process [33, 34]. In addition, the ability to automate a report of the survey's findings, such as sediment composition, habitats, *etc.*, would reduce the time and effort required for manual preparation. To this end, we present a data curation pipeline that leverages both domain knowledge from marine scientists and language generation capability of `GPT-4` [46]. Specifically, we employ in-context learning [8] to generate analysis-driven question-answer pairs for each image, resulting in 7M samples and 696K language descriptions. We name the vision-language dataset `SeafloorGenAI`.

[2]Provides public domain data license.

Location	Region Index	Image Resolution	Input Layers	Mapping Layers					Area (km^2)
				Sediment	Physio Zone	Habitat	Fault	Fold	
California	Region 1	2m/pixel	25,817	25,672	25,823				672
	Region 2	2m/pixel	123,774			123,480	123,774	123,774	3,148
	Region 3	2m/pixel	21,270	20,861	21,253				564
	Region 4	2m/pixel	42,771			25,579	42,771	42,771	1,419
Massachussetts	Region 5	10m/pixel	15,827	4,647	3,387				5,496
	Region 6	1m/pixel	122,441	122,236	118,175				228
	Region 7	1m/pixel	1,593	1,507	1,510				454
Delmarva	Region 8	2m/pixel	329,881						4,525
South Carolina	Region 9	4m/pixel	13,141						808
		Total	696,515	174,923	170,148	149,059	166,545	166,545	17,314

Table 1: Summary of the seafloor mapping data available for each region. The input layers for sonar images include Backscatter, Bathymetry, Slope and Rugosity. Due to different mapping objectives of the original data surveys, the availability of segmentation masks is not uniform across mapping layers. Regions with unlabeled data can be utilized to pre-train the model via self-supervised learning [19].

Our contributions are summarized as follows:

1. We compile 62 public hydrographic data surveys from USGS and NOAA into a large, geo-distributed, multi-purpose and multi-modal dataset for seafloor mapping research.
2. We provide a standardization of naming convention across these surveys, under the *rigorous supervision of marine scientists*, to unify an extensive AI-ready dataset.
3. We present a data curation pipeline that produces detailed descriptions and question-answer pairs for the development of large generative vision-language models in marine science.
4. Our geo-distributed dataset contains 696K sonar images, 827K segmentation masks, 696K language descriptions and 7M question-answer pairs, covering a total area of 17,300 square kilometers.
5. We open-source our data processing code so that marine scientists could efficiently contribute their data surveys to expand the data pool.

2 Related Work

Underwater Imagery Datasets. Over the years, researchers at USGS and NOAA have carried out frequent hydrographic surveys [17, 48, 49, 47, 50, 2, 5] to collect and provide accurate and reliable information about the physical features of the water bodies and the seafloor. They are instrumental in creating accurate nautical charts to identify underwater hazards, aiding in the planning of marine infrastructure, and providing essential data for scientific research and environmental conservation. Furthermore, the data supports various economic activities, such as fishing, aquaculture, and energy production, by enabling sustainable and efficient operations.

In recent years, substantial efforts have been made to create public AI-ready underwater datasets, including forward-looking sonar (FLS), side-scan sonar (SSS), and RGB imagery. These datasets are utilized to develop machine learning models tailored for domain applications, focusing on classification or detection of geological features [3, 11, 7, 39, 54, 38] and man-made objects [77, 70, 26, 75, 32, 44, 13, 73, 72, 43, 84, 71]. Singh and Valdenegro-Toro [63] were pioneers with their FLS image dataset aimed at object detection, but their use of a controlled water tank setting may not fully reflect the complex oceanic conditions, limiting the generalizability of their results. Xie et al. [74] addressed this by extending object detection to data collected in natural water bodies, enhancing its real-world applicability. Sethuraman et al. [61] developed an SSS dataset for shipwreck detection, though its small sample size could limit model robustness. Others have also explored RGB underwater imagery for trash detection [69] and semantic segmentation [27].

Our research focuses on transforming the USGS and NOAA hydrographic surveys into a comprehensive, multi-scale, multi-purpose and multi-modal SSS imagery dataset. This initiative aims to propel advancements in both marine science and machine learning research, creating a bridge between extensive marine data resources and innovative computational techniques.

Why side-scan sonar? Compared to FLS and RGB imagery, SSS offers distinct advantages for underwater imagery analysis. Side-scan sonar provides a wider coverage area, and creates high-resolution images that clearly delineate the seabed texture, which is essential for geological surveys, shipwreck location, and habitat mapping. Unlike FLS, which is primarily used for obstacle avoidance, SSS offers a broad, fan-shaped beam that scans the ocean floor to either side of the towfish or autonomous underwater vehicle, capturing detailed images of the seafloor texture. Moreover, SSS is less affected by water turbidity compared to RGB cameras, which struggle with visibility in murky waters and suffer from significant color loss at depth due to light absorption. This allows SSS to produce consistent and reliable imagery under a variety of underwater conditions, where optical methods would fail. Still, SSS is only a 2D representation of the seabed. We also incorporate 3D information such as water depth to describe the underwater topography. This allows for a broad scope of underwater imagery analysis, providing robust data suitable for in-depth assessments.

Comparison with Existing Datasets. Our dataset is a comprehensive and expansive dataset that serves two primary purposes: (1) to act as a benchmark for various tasks and (2) to train foundation vision or vision-language models with a focus on seafloor morphodynamic analysis. In contrast to existing datasets [63, 74, 61, 69], which may specialize in single machine learning tasks or offer limited data samples, our dataset provides a diverse array of seafloor mapping tasks sourced from geographically diverse regions. Additionally, we make our data processing source code publicly available, encouraging further expansion of the dataset towards the magnitude of large-scale natural imagery datasets [62, 9, 28, 60, 59, 35].

Datasets in other Scientific Domains. Following the success of large foundation models in natural imagery [55, 14, 83, 82, 35, 12, 4, 79, 57, 31, 76, 81], there has been a significant push to develop expansive datasets tailored for training large foundation models for specific domain applications. In remote sensing, initiatives such as RSVQA [37], RSVQA-BEN [36], and RSGPT [23] have been developed to enhance general VQA capabilities, while MUSE [30] targets more complex reasoning tasks. Similarly, in medical imaging, datasets such as PathVQA [22], PMC-VQA [80], XrayGPT [65], LLaVA-Med [33], and OmniMedVQA [24] aim to improve the visual and textual understanding of various body parts through the analysis of MRI, X-rays, *etc.* These datasets comprise hundreds of millions of samples, posing significant acquisition challenges, particularly in marine science where data annotation is notably expensive. To address this, our initiative seeks to develop a large-scale dataset, aiming to significantly expand the resources available for marine science.

3 The `SeafloorAI` Dataset

3.1 Dataset Overview

`SeafloorAI` is a large, geo-distributed, multi-purpose dataset designed to map various geological layers of the seafloor. It is catered for training computer vision models, *i.e.* CNNs and Vision Transformers that produce semantic segmentation masks. Furthermore, it facilitates the studies of fundamental ML problems such as robust optimization [53, 51, 52, 45]. The dataset also serves as a basis for constructing the generative vision-language variant, `SeafloorGenAI`, discussed in Sec. 4.

Our dataset is compiled from 62 geological data surveys published on USGS and NOAA repositories, spanning an area of 17,300 square kilometers. This dataset features a broad geographical distribution, covering the nearshore zones of several states, including California [18], Massachusetts [49, 47, 50, 2], Delmarva [48], and South Carolina [5]. These areas are further divided into 9 regions. The data for this dataset were collected over a period spanning from 2004 to 2024, using a variety of single side-scan sonars and multibeam echosounders with different frequencies. These instruments were employed to record the texture (Backscatter) and depth (Bathymetry) of the seafloor.

The surveys have been meticulously annotated by domain experts, focusing on five key geological layers: Sediment, Physiographic Zone, Habitat, Fault, and Fold as detailed in Tab. 1. This expansive and detailed dataset provides a comprehensive view of geological and environmental features across a wide range of coastal environments. In summary, we convert the raw raster data into a large-scale machine learning-ready dataset containing 696,515 input samples, and 827,220 annotated segmentation masks across various layers.

3.2 Data Processing

The input layers, consisting of Backscatter and Bathymetry signals, are provided as raster data in `GeoTIFF` format. The five mapping layers serve as the ground-truth annotations, defining five tasks for the model training and evaluation. These layers come in `shapefile` format that stores the location (*i.e.*, longitude and latitude), shape (*i.e.*, polygons) and attributes of geological features (*i.e.*, sediment type). These polygons define the regions of interest on raster images, effectively delineating the boundaries of different categories that we want to segment.

Next, we present the steps for data processing at a high level, and then go further into details with each geological layer. First of all, we reproject all layers from all surveys to the WGS84 (EPSG:4326)[3] coordinate reference system. Then, we rasterize the `shapefile` to `GeoTIFF` format, effectively converting all the annotations into 2D arrays occupying the same geo-location. Finally, we use a sliding window to split the 2D raster layers into 224×224 patches with a step size of 56 to avoid information loss at the edges. These patches serve as the inputs and outputs for the machine learning algorithms. This process is also referred to as "patchifying".

Input Layers: Backscatter & Bathymetry. Backscatter in marine science refers to the amplitude of the echoes of sound waves emitted/received by a transducer that bounce off objects or the seafloor and return to the receiver. By analyzing the time it takes for the sound waves to return and their acoustic intensity, scientists and researchers can create underwater maps of the submerged terrain and identify the composition and characteristics of the seafloor, as well as the presence of underwater objects or marine life. In our dataset, we normalize the backscatter signals to the [0, 255] range, with 255 representing the nodata value. Regarding Bathymetry, we set the nodata value to be a negative number of significant magnitude, *i.e.*, -9999. Additionally, we convert Bathymetry measurements from meters to kilometers, compressing these values into a [0,1] range for normalization purposes.

We further calculate two morphologic derivatives from Bathymetry, namely Slope and Rugosity, to more comprehensively represent the topographical features of the seafloor in the input space. Slope refers to the *steepness* of the seabed, calculated as the rate of change in elevation over a given distance. It is crucial for understanding sediment transport, habitat diversity, and the stability of underwater structures. We use `GDAL` [16] implementation of the Zevenbergen & Thorne formula [78] to estimate the slope. In brevity, the formula computes the differences in elevation between a central pixel and its eight surrounding pixels for a more smoothed and stable slope estimation. Rugosity, on the other hand, measures the *roughness* or irregularity of the ocean floor. It quantifies the amount of surface area relative to a flat plane, offering vital clues about the complexity of habitats, which affects biodiversity and ecological interactions.

For each region, we resample Bathymetry, Slope and Rugosity to the Backscatter's resolution. As a result, our dataset contains a range of resolutions across regions, from 1m to 10m per pixel, enabling both coarse and fine-grained understanding of seafloor morphodynamic analysis. After patchifying the rastered map, we only keep patches where the number of nodata pixels is below 10% the number of total pixels. In the final step, we apply interpolation to fill in the missing pixels, and median filtering to reduce speckle noise. The input contains 6 channels, including these 4 layers and 2 geo-location channels (pixel-wise longitude and latitude), resulting in a dimension of $224 \times 224 \times 6$.

R	Rg	Gr	**G**
Rs	Rm	Gs	Gm
Sr	Sg	Mr	Mg
S	Sm	Ms	**M**

Figure 2: The Barnhardt classification scheme [6] is based on four end-member units: (**R**)ock, (**G**)ravel, (**S**)and, and (**M**)ud. The other twelve composite categories represent the combinations of the four units, where the dominant texture ($> 50\%$) is in upper case, and the subordinate ($< 50\%$) in lower.

Mapping Layers: Sediment, Physiographic Zone & Habitat. Our dataset is derived from 62 different surveys spanning both the East and West Coasts of the United States. Given the diverse origins of the data, there are inherent inconsistencies in the annotations, such as varying standards or differing vocabularies used to label the same categories. To address this, we have developed a unification process for ground-truth labels, leading to the creation of multi-class segmentation masks for Sediment, Physiographic Zone, and Habitat. This standardization process is meticulously overseen by domain experts to ensure the accuracy and quality of the annotations.

[3] More information at `https://docs.up42.com/data/reference/utm`.

(1) Sediment. Sediments on the seafloor, composed of varied particles from multiple sources, are crucial for creating habitats, indicating geological processes, and aiding in environmental and ecological research. They play a key role in resource exploration by helping to identify potential sites for natural resource extraction and in climate change studies by preserving historical climate data. Detailed seafloor mapping using sediment analysis is vital for accurate marine navigation, scientific research, and effective marine resource management. We define a unified annotation standard for the Sediment layer, following the Barnhardt classification table [6], which is a classification scheme based on four end-member units: (**R**)ock, (**G**)ravel, (**S**)and, and (**M**)ud. The other twelve composite units represent the combinations of the four units, where the dominant texture ($> 50\%$ of the area) is in upper case, and the subordinate ($< 50\%$ of the area) is in lower case, illustrated in Fig. 2. Finally, we construct semantic segmentation masks for each input patch where each pixel contains an integer value from 0 to 16, with 0 denoting the pixels without annotations.

(2) Physiographic Zone. By definition, a physiographic zone refers to a distinct geographical region characterized by a uniformity in topography and underlying geological structure that sets it apart from adjacent areas. These zones are typically defined based on natural landscape features, such as the configuration of the terrain, rock formations, and soil types. Classifying these zones requires the holistic understanding of multiple geological features, hence the necessity to include the bathymetric derivatives, such as Slope and Rugosity, as input. Similar to Sediment, we also define a standard for the Physiographic Zone layer. We follow the CMECS unit code for Physiographic Province which belongs in the Geoform Component [1]. There are 21 different categories for Physiographic Zone, as shown in Fig. 3.

Figure 3: Twenty-one physiographic zone categories from CMECS.

(3) Habitat. One of the aims of seafloor mapping efforts is to delineate benthic habitats as a high-level outcome. Hall et al. [20] defined Habitat as "the resources and conditions present in an area that produce occupancy ... by a given organism." According to CMECS, a benthic habitat refers to the ecological regions at the lowest level of a body of water, including the sediment surface and sub-surface layers [1]. Benthic habitats are critical areas because they provide living space for a wide range of organisms, both flora and fauna, which are integral to the marine ecosystem. Specifically focusing on abiotic benthic habitats, these are characterized by non-living physical and chemical aspects of the environment that influence the type and abundance of organisms living there. To unify the annotations across surveys, we first gather all 144 descriptions of the polygons from the public data surveys. We then categorize these descriptions into broader groups, ultimately consolidating them into 9 distinct categories for Habitat, depicted in Fig. 4.

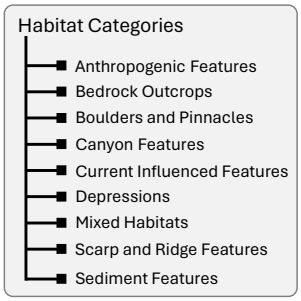


Figure 4: Nine major categories for abiotic habitat defined in `SeafloorAI`.

Mapping Layers: Fault & Fold. Faults and folds are significant geological features on the seafloor that are formed by tectonic movements within the Earth's crust. Faults occur when rock layers break and slide past each other due to tectonic forces, creating distinct disruptions in the seabed. Folds are bends in rock layers that occur when these layers are compressed and folded, resulting in curved or wavy stratifications. Detecting these features is crucial for understanding seismic activity and geological history of the marine environment. In our study, we formulate the binary segmentation task to identify the presence of these geological features within specific image patches, assigning the pixels containing the features a value of 1, and 0 otherwise.

4 The SeafloorGenAI Dataset

SeafloorGenAI incorporates vision and language understanding via visual question answering (VQA), facilitating the advancement of large vision-language models in the marine science field and the conventional studies on multi-modal learning [55, 31, 33, 41, 40]. This integration enables smooth interactions between domain experts and AI, providing clear explanations and streamlining the process of data analysis and discovery. Our dataset, consisting of 7M QA pairs and 696K language descriptions, is designed to support *general VQA capability* and *instruction-following mapping*.

General descriptions and VQA. Following previous work from other domains [33, 30], we utilize large language models (LLMs), specifically GPT-4, to generate the language descriptions and question-answer pairs for each sonar imagery sample. We employ in-context learning (ICL) [8], providing few-shot input-output pairs for the LLM. In this case, the input contains the *key analytical indicators* and the output is the description written by the marine scientists for the same image. To construct the ICL input, we, in collaboration with marine scientists, **identify** the essential information required for analysis. Subsequently, we use standard statistical and computer vision tools to **extract** three categories of information: (1) *geophysical parameters*, (2) *spatial distribution* and (3) *geological composition*. The objective is to help the model "see" the sonar image through as much detailed language descriptions as possible. For the ICL output, we ask marine scientists to manually **describe** in domain language 50 randomly selected samples from the SeafloorAI dataset. ICL ensures GPT-4 can accurately mimic the domain-specific language, enhancing the quality and relevance of the generated answers. Next, we design a **prompt** to GPT-4, comprised of the input-output pairs and the extracted analytical indicators, to generate general descriptions and question-answer pairs for the remaining images. Finally, the domain experts carefully **evaluate** the generated language annotations to ensure quality and consistency. The last two steps form a feedback loop, creating an iterative prompt refinement process. Fig. 5 illustrates the described pipeline.

In Fig. 6, we show a sample selected from the SeafloorGenAI dataset. We can see that GPT-4 is able to generate QA pairs that relate different geological layers at the same location. This helps unravel complex ecological dynamics, which is beneficial to many domain applications. We now discuss how each type of information (*i.e.* geophysical parameters, spatial distribution and geological composition) is extracted from the image.

(1) Geophysical parameters. These parameters are important, serving as the base for further analysis of the area. In our data processing pipeline, we employ classical analysis techniques to extract key geophysical parameters from processed data, such as water depth, mean and standard deviation of backscatter intensity, and ranges of slope, *etc.* These parameters are then systematically converted into textual format. This transformation facilitates a structured representation of complex numerical data, making it more accessible and interpretable for further analysis and reporting.

An example of Geophysical Parameters in the Input layers

```
Geolocation: (42.55°, -70.67°) to (42.53°, -70.64°)
Depth range: -36.4 to -54.2 meters
Backscatter mean and standard deviation: 119.7 and 72.2
Slope range: 1.7 to 9.2 degrees
Rugosity range: 0.01 to 0.02
```

(2) Geological composition. Understanding geological composition allows marine scientists to gain a holistic view of seafloor characteristics by examining how geological features are proportionally distributed within a specific area. Technically, this involves calculating the ratio of total pixels for each geological category relative to the overall pixels in the segmentation mask. As a result, we achieve the following:

An example of Geological Composition in the Sediment layer

```
Muddy Sand (Sm) accounts for 45% of the image.
Muddy Rock (Rm) accounts for 25% of the image.
Gravel Rock (Rg) accounts for 11% of the image.
```

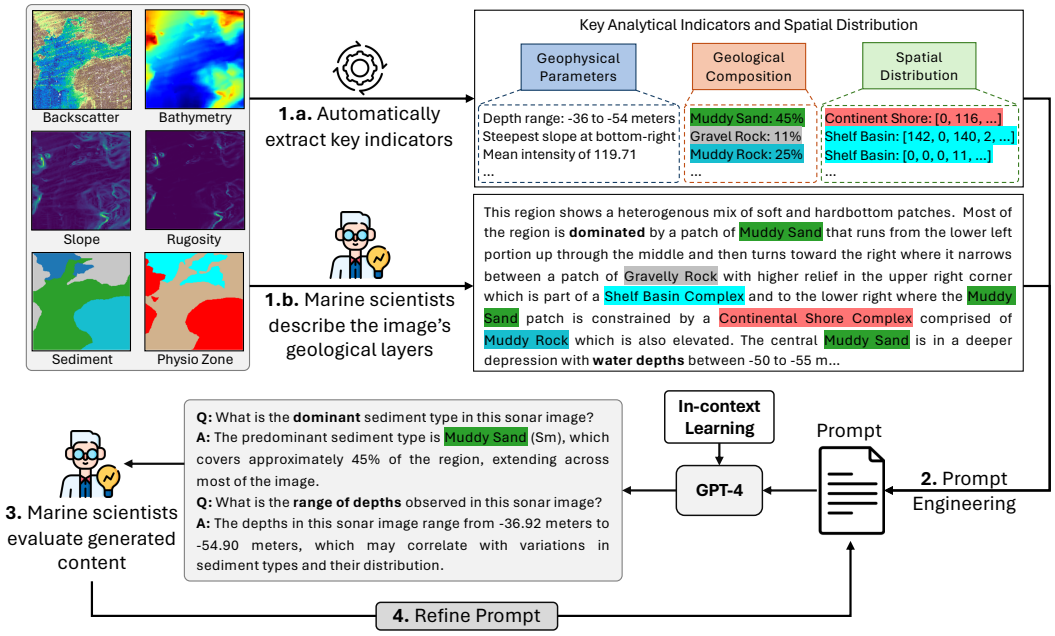


Figure 5: Pipeline for generating question-answer pairs for sonar imagery samples using `GPT-4`: Marine scientists first identify necessary information, followed by the extraction of *geophysical parameters*, *geological composition*, and *spatial distribution*. They then provide descriptions for a handful of samples from the `SeafloorAI` dataset. These description are used to design a prompt for `GPT-4` to generate high-quality, domain-specific question-answer pairs, via in-context learning [8].

(3) Spatial distribution. Spatial distribution complements geological composition, thus giving a more comprehensive description of the image. We convert the segmentation mask of each category to polygons, which can then be fed as language into `GPT-4`. We first find the contours of the masks using conventional computer vision techniques, then transform them into polygon representation with the format $[x_1, y_1, ..., x_n, y_n]$, where x_i and y_i are the coordinates of the i^{th} point in n points.

An example of Spatial Distribution in the Physiographic Zone layer

```
Continential Shore Complex polygon at [0, 116, 0, 186, ..., 1, 117]
Shelf Basin polygon at [142, 0, 140, 2, ..., 156, 0]
```

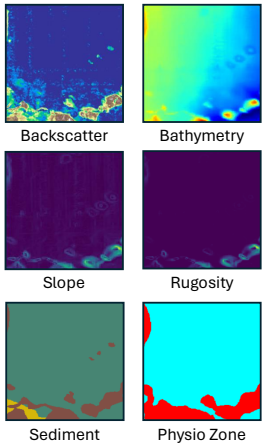
Instruction-following Mapping. Besides VQA, we aim to equip the AI assistant with the capability to map various seafloor features across different layers in response to specific instructions. This facilitates a seamless and intuitive interaction between the AI and marine scientists, allowing for easy querying and efficient analysis. We design our dataset to be compatible with state-of-the-art VLM models, such as PixelLM [57] and LISA [31] for both single and multi-instance segmentation tasks.

Examples of single and multi-instance instruction-following mapping in SeafloorGenAI

```
(1) Q: Please segment [CATEGORY] in [LAYER].
    A: Sure, <SEG>.

(2) Q: What are present in the image for [LAYER]? Please segment them.
    A: [CATEGORY_1] <SEG_1>, [CATEGORY_2] <SEG_2>, ..., [CATEGORY_N] <SEG_N>.

(3) Q: Identify the areas of [CATEGORY_1] from [LAYER_1] and [CATEGORY_2]
       from [LAYER_2].
    A: Sure, [CATEGORY_1] from [LAYER_1] <SEG_1> and [CATEGORY_2]
       from [LAYER_2] <SEG_2>.
```



Description

This sonar image captures a marine area characterized by varied sediment types and physiographic zones. The area spans from coordinates **42.4753**, **-70.6576** to **42.4518**, **-70.6340**, revealing a depth range from **-60.43 meters** to **-83.66 meters**. Dominated by Sandy Mud covering **87.48%** of the area, the image presents a predominantly low-energy. Gravelly Rock and Sand patches are scattered, indicating regions of slightly higher relief, which hint at more dynamic, possibly higher-energy conditions. The image is segmented into two primary physio zones: a Shelf Basin comprising **83.79%** and a Continental/Island Shore Complex **16.21%**, suggesting a diverse underwater landscape with significant transitions from deeper basin areas to more complex nearshore environments.

Question-Answer Pairs

Q. What percentage of the sonar image is covered by Sandy Mud, and how is it represented spatially?
A. Sandy Mud dominates the sonar image, covering **87.48%** of the area. It is spatially represented by a large, extensive polygon that stretches across most of the image.

Q. How is rugosity distributed across the different sediment types in the sonar image?
A. Rugosity has a low mean value of **0.01**, suggesting overall flatness across the image.

Q. How does the physiographic zone of Shelf Basin relate to the observed sediment types?
A. The Shelf Basin, covering **83.79%** of the area, predominantly contains Sandy Mud, suggesting a deeper, more uniform basin environment conducive to the deposition of fine sediments.

Figure 6: An example in the SeafloorGenAI dataset, originated from Region 5. It features a GPT-4 generated description and question-answer pairs designed to efficiently assist marine scientists in data analysis. The generated description covers all three key analytical indicators. Noticeably, the last QA pairs focuses on cross-layer understanding (*i.e.*, Sediment and Physiographic Zone), which is helpful for unraveling complex ecological dynamics on the seabed.

5 Experiments

We report some baseline experiment runs on SeafloorAI for multi-class segmentation. Due to space limit, we move the experiments for binary segmentation to the Supplementary Material.

Evaluation Metrics. We use pixel-wise accuracy (Acc), Dice coefficient (Dice) and Jaccard coefficient (mIoU) to evaluate the baseline models.

Data Split. We present the data splits for Sediment, Physiographic Zone and Habitat, as well as the motivation for such splits. Due to the availability of the categories in each region, we make sure that the training regions possess the set of categories that cover the testing region(s). We present our data splits for the layers in Tab. 2. For the source data, we randomly split them into 90% for training and 10% for validation. The validation set is used to select the best model for testing on the target data.

Task	Layer	Source	Target
Multi-class Segmentation	Sediment	Region 1, Region 5, Region 6, Region 7	Region 3
	Physio Zone	Region 1, Region 3, Region 5, Region 6	Region 7
	Habitat	Region 2	Region 4

Table 2: Geo-distributed data splits for the SeafloorAI dataset for multi-class segmentation.

Training Details. We employ the UNet architecture with different backbones as baselines. The UNet architecture [58] consists of a contracting path (encoder) and an expanding path (decoder), forming a U-shape. We use UNet-Base [58], UNet-ResNet18 [21] and TransUNet-ViT-B/32 [10, 15] as our baseline models for the multi-class segmentation tasks. We adopt cross-entropy as the loss function. The model was trained using the Adam optimizer [29]. The learning rate was initially set to 0.001 with a cosine annealing schedule. We use a batch size of 64 for 100 epochs, setting the patience to 5 epochs for early stopping. We perform 3 runs with different random seeds and report the model performance in Tab. 3. All runs are conducted on a single NVIDIA RTX A6000 GPU.

Results. Tab. 3 reports the results on the geo-distributed setting, which is similar to out-of-distribution generalization [53, 51]. We report the in-distribution (ID; on source data) and out-of-distribution (OOD; on target data) pixel-wise accuracy, Dice coefficient and Jaccard coefficient. Overall, we can see that all baseline models suffer from a signification performance degradation under distribution shift. This might be due to covariate shift (sensor types and configurations) and subpopulation shift (class imbalance). Therefore, ensuring that a model generalizes well to new, unseen distributions is a fundamental challenge. Standard training methods often assume that the training and testing data come from the same distribution, which is rarely the case in real-world applications.

	Sediment								
	Acc ID	Acc OOD	Δ Acc	Dice ID	Dice OOD	Δ Dice	mIoU ID	mIoU OOD	Δ mIoU
UNet-Base	77.45 ± 0.81	21.49 ± 0.91	-55.96	79.73 ± 0.83	21.59 ± 0.97	-58.14	66.46 ± 1.15	12.29 ± 0.61	-54.17
UNet-ResNet18	78.45 ± 0.67	34.71 ± 6.79	-43.74	80.78 ± 0.71	35.01 ± 6.86	-45.77	67.90 ± 1.00	22.08 ± 5.73	-45.82
TransUNet	67.90 ± 2.18	28.32 ± 1.04	-39.58	69.94 ± 2.27	29.16 ± 1.05	-40.16	53.98 ± 2.65	17.41 ± 0.71	-36.57
	Physio Zone								
	Acc ID	Acc OOD	Δ Acc	Dice ID	Dice OOD	Δ Dice	mIoU ID	mIoU OOD	Δ mIoU
UNet-Base	93.05 ± 0.16	56.56 ± 0.87	-36.49	95.81 ± 0.18	57.09 ± 0.69	-38.72	91.98 ± 0.32	43.22 ± 0.84	-48.76
UNet-ResNet18	92.87 ± 0.10	56.74 ± 2.53	-36.13	95.63 ± 0.09	59.86 ± 2.54	-35.77	91.66 ± 0.17	42.97 ± 3.00	-48.69
TransUNet	90.63 ± 0.20	56.24 ± 1.66	-34.39	93.28 ± 0.27	57.51 ± 1.84	-35.77	87.49 ± 0.47	43.86 ± 2.04	-43.63
	Habitat								
	Acc ID	Acc OOD	Δ Acc	Dice ID	Dice OOD	Δ Dice	mIoU ID	mIoU OOD	Δ mIoU
UNet-Base	92.02 ± 0.18	70.54 ± 1.72	-21.48	94.82 ± 0.20	71.04 ± 1.54	-23.78	90.19 ± 0.37	56.75 ± 2.01	-33.44
UNet-ResNet18	92.70 ± 0.12	76.40 ± 1.33	-16.30	95.50 ± 0.11	76.59 ± 1.28	-18.91	91.43 ± 0.20	65.17 ± 1.80	-26.26
TransUNet	88.67 ± 0.56	70.56 ± 0.72	-18.11	91.34 ± 0.59	72.76 ± 0.83	-18.58	84.15 ± 0.99	59.38 ± 1.24	-24.77

Table 3: Performance of the baselines in the geo-distributed setting for multi-class segmentation.

6 Human Evaluation for Language Annotations

Although `GPT-4` has shown strong capabilities in data annotations [64], hallucinations in LLMs are inevitable [25]. To ensure the quality of the language annotations generated by `GPT-4`, we describe an iterative prompt refinement process that involves human expert evaluation.

To maximize budget efficiency, we designed our procedure with several iterations of feedback and refinement. The idea is to engineer and refine our prompt to `GPT-4` on a small subset of data before applying it to the whole dataset. For each iteration, (1) we annotated 1,000 random samples with `GPT-4`; (2) the marine scientists reviewed the quality of the generated annotations and gave feedback based on the three criteria: (i) *factual consistency* to the original annotations, (ii) *factual completeness* with respect to the analytical indicators and (iii) *coherence* to domain language; (3) we refined our prompts to `GPT-4`, a.k.a prompt engineering, to achieve higher quality language annotations, (4) we repeated the steps for the next iteration. Finally, when the quality is met, we will populate the entire dataset with language annotations. Due to space limitation, we include more details in the Supplementary Material.

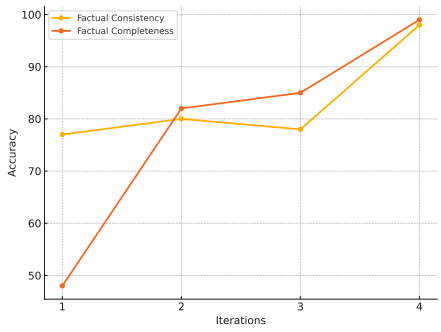


Figure 7: Accuracy for factual consistency and completeness increases over the iterations thanks to rigorous the prompt refinement procedure. `GPT-4` performs worse on factual completeness potentially due to hallucinations.

7 Limitations and Future Work

Despite the extensiveness of our dataset, there are notable limitations to discuss. Firstly, the availability of layers is not uniform across all regions; for instance, Region 5 is missing Habitat, Fault, and Fold layers. This is due to the different mapping objectives when the data surveys were first collected. Additionally, the existing nine Habitat categories are somewhat coarse and exclude biotic classifications. We are actively collaborating with marine scientists to refine and expand the Habitat layer, making it more detailed and comprehensive. The current version of the `SeafloorGenAI` dataset provides annotations suitable for straightforward analytical queries and lacks the data for deeper reasoning abilities. Moving forward, we plan to enhance the dataset to support the development of reasoning-capable models similar to referring and reasoning segmentation as in [57, 31, 76], offering more profound insights into marine science questions and paving the way for data discovery. Developing this enhanced version of the dataset will require a structured and systematic approach to understanding domain-specific knowledge to accurately annotate the data. In terms of modeling, our plan for future work involves training a generative vision-language model on the `SeafloorGenAI` dataset, serving as a foundation ML model in marine science research.

Acknowledgement

This work is supported by the DoD DEPSCoR Award AFOSR FA9550-23-1-0494, the NSF CAREER Award No. 2340074, the NSF SAFE Award No. 2416937, and the NSF III CORE Award No. 2412675. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the supporting entities.

References

[1] Coastal and Marine Ecological Classification Standard (CMECS) — repository.library.noaa.gov. `https://repository.library.noaa.gov/view/noaa/27552`. [Accessed 24-05-2024].

[2] Seth D Ackerman, Adrienne L Pappal, Emily C Huntley, Dann S Blackwood, and William C Schwab. Geological sampling data and benthic biota classification: Buzzards bay and vineyard sound, massachusetts, 2015.

[3] Riccardo Arosio, Brandon Hobley, Andrew J Wheeler, Fabio Sacchetti, Luis A Conti, Thomas Furey, and Aaron Lim. Fully convolutional neural networks applied to large-scale marine morphology mapping. *Frontiers in Marine Science*, 10:1228867, 2023.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[5] Wayne E. Baldwin, Robert A. Morton, Jane F. Denny, William C. Schwab Shawn V. Dadisman, Paul T. Gayes, and Neal W. Driscoll. Maps showing the stratigraphic framework of south carolina's long bay from little river to winyah bay, 2004.

[6] Walter A. Barnhardt, Joseph T. Kelley, Stephen M. Dickson, and Daniel F. Belknap. Mapping the gulf of maine with side-scan sonar: A new bottom-type classification for complex seafloors. *Journal of Coastal Research*, 14(2):646–659, 1998.

[7] Tim Berthold, Artem Leichter, Bodo Rosenhahn, Volker Berkhahn, and Jennifer Valerius. Seabed sediment classification of side-scan sonar data using convolutional neural networks. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8, 2017.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.

[10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

[11] Johnny L Chen and Jason E Summers. Deep neural networks for learning classification features and generative models from synthetic aperture sonar big data. In *Proceedings of Meetings on Acoustics*, volume 29. AIP Publishing, 2016.

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024.

[13] Zhen Cheng, Guanying Huo, and Haisen Li. A multi-domain collaborative transfer learning method with multi-scale repeated attention mechanism for underwater side-scan sonar image classification. *Remote Sensing*, 14(2):355, 2022.

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[16] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2024.

[17] Nadine E Golden. California state waters map series data catalog, 2013.

[18] Nadine E Golden. California state waters map series data catalog, 2013.

[19] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends, 2024.

[20] Linnea S. Hall, Paul R. Krausman, and Michael L. Morrison. The habitat concept and a plea for standard terminology. *Wildlife Society Bulletin (1973-2006)*, 25(1):173–182, 1997.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[22] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020.

[23] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023.

[24] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm, 2024.

[25] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

[26] Guanying Huo, Ziyin Wu, and Jiabiao Li. Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data. *IEEE Access*, 8:47407–47418, 2020.

[27] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776, 2020.

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[30] Kartik Kuckreja, Muhammad S. Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S. Khan. Geochat: Grounded large vision-language model for remote sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[31] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[32] Chuanlong Li, Xiufen Ye, Dongxiang Cao, Jie Hou, and Haibo Yang. Zero shot objects classification method of side scan sonar image based on synthesis of pseudo samples. *Applied Acoustics*, 173:107691, 2021.

[33] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Tang Li, Mengmeng Ma, and Xi Peng. Deal: Disentangle and localize concept-level explanations for vlms. In *European Conference on Computer Vision*, pages 383–401. Springer, 2025.

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[36] Sylvain Lobry, Begüm Demir, and Devis Tuia. Rsvqa meets bigearthnet: a new, large-scale, visual question answering dataset for remote sensing. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 1218–1221. IEEE, 2021.

[37] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

[38] Mark A Lundine, Laura L Brothers, and Arthur C Trembanis. Deep learning for pockmark detection: Implications for quantitative seafloor characterization. *Geomorphology*, 421:108524, 2023.

[39] Xiaowen Luo, Xiaoming Qin, Ziyin Wu, Fanlin Yang, Mingwei Wang, and Jihong Shang. Sediment classification of small-size seabed acoustic images using convolutional neural networks. *IEEE Access*, 7:98331–98339, 2019.

[40] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18177–18186, June 2022.

[41] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

[42] Larry Mayer, Martin Jakobsson, Graham Allen, Boris Dorschel, Robin Falconer, Vicki Ferrini, Geoffroy Lamarche, Helen Snaith, and Pauline Weatherall. The nippon foundation—GEBCO seabed 2030 project: The quest to see the world's oceans completely mapped by 2030. *Geosciences (Basel)*, 8(2):63, February 2018.

[43] John McKay, Isaac Gerg, Vishal Monga, and Raghu G Raj. What's mine is yours: Pretrained cnns for limited training sonar atr. In *OCEANS 2017-anchorage*, pages 1–7. IEEE, 2017.

[44] Nandeeka Nayak, Makoto Nara, Timmy Gambin, Zoë Wood, and Christopher M Clark. Machine learning techniques for auv side-scan sonar data feature extraction as applied to intelligent search for underwater archaeological sites. In *Field and Service Robotics: Results of the 12th International Conference*, pages 219–233. Springer, 2021.

[45] Kien X. Nguyen, Fengchun Qiao, and Xi Peng. Adaptive cascading network for continual test-time adaptation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1763–1773, New York, NY, USA, 2024. Association for Computing Machinery.

[46] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin,

Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[47] Elizabeth A Pendleton, Wayne E Baldwin, Walter A Barnhardt, Seth D Ackerman, David S Foster, Brian D Andrews, and William C Schwab. Shallow geology, seafloor texture, and physiographic zones of the inner continental shelf from nahant to northern cape cod bay, massachusetts, 2013.

[48] Elizabeth A Pendleton, Edward M Sweeney, and Laura L Brothers. Optimizing an inner-continental shelf geologic framework investigation through data repurposing and machine learning. *Geosciences (Basel)*, 9(5):231, May 2019.

[49] Elizabeth E Pendleton, Walter A Barnhardt, Wayne E Baldwin, David S Foster, William C Schwab, Brian D Andrews, and Seth D Ackerman. Sea-floor texture and physiographic zones of the inner continental shelf from salisbury to nahant, massachusetts, including the merrimack embayment and western massachusetts bay, 2015.

[50] Elizabeth P Pendleton, Wayne E Baldwin, David S Foster, Seth Ackerman, Brian D Andrews, and Laura Brothers. Geospatial data layers of shallow geology, sea-floor texture, and physiographic zones from the inner continental shelf of martha's vineyard from aquinnah to wasque point, and nantucket from eel point to great point, 2018.

[51] Xi Peng, Fengchun Qiao, and Long Zhao. Out-of-domain generalization from a single source: An uncertainty quantification approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1775–1787, 2022.

[52] Fengchun Qiao and Xi Peng. Topology-aware robust optimization for out-of-distribution generalization. In *The Eleventh International Conference on Learning Representations*, 2023.

[53] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020.

[54] Xiaoming Qin, Xiaowen Luo, Ziyin Wu, and Jihong Shang. Optimizing the sediment classification of small side-scan sonar images based on deep learning. *IEEE Access*, 9:29416–29428, 2021.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[56] Nicole A Raineault, Arthur C Trembanis, and Douglas C Miller. Mapping benthic habitats in delaware bay and the coastal atlantic: Acoustic techniques provide greater coverage and high resolution in complex, Shallow-Water environments. *Estuaries Coast.*, 35(2):682–699, March 2012.

[57] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model, 2023.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[60] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

[61] Advaith Venkatramanan Sethuraman, Anja Sheppard, Onur Bagoren, Christopher Pinnow, Jamey Anderson, T. Havens, and Katherine A. Skinner. Machine learning for shipwreck segmentation from side scan sonar imagery: Dataset and benchmark. *ArXiv*, abs/2401.14546, 2024.

[62] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[63] Deepak Singh and Matias Valdenegro-Toro. The marine debris dataset for forward-looking sonar semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3734–3742, 2021.

[64] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey, 2024.

[65] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models, 2023.

[66] Arthur Trembanis, Alimjan Abla, Ken Haulsee, and Carter DuVal. Benthic habitat morphodynamics-using remote sensing to quantify storm-induced changes in nearshore bathymetry and surface sediment texture at assateague national seashore. *J. Mar. Sci. Eng.*, 7(10):371, October 2019.

[67] Arthur Trembanis, Carter DuVal, Jonathan Beaudoin, Val Schmidt, Doug Miller, and Larry Mayer. A detailed seabed signature from hurricane sandy revealed in bedforms and scour. *Geochemistry, Geophysics, Geosystems*, 14(10):4334–4340, 2013.

[68] Arthur Trembanis, Mark Lundine, and Kaitlyn McPherran. Coastal mapping and monitoring. In *Reference Module in Earth Systems and Environmental Sciences*. Elsevier, 2020.

[69] Jaskaran Singh Walia and Karthik Seemakurthy. Optimized custom dataset for efficient detection of underwater trash. *ArXiv*, abs/2305.16460, 2023.

[70] Xingmei Wang, Jia Jiao, Jingwei Yin, Wensheng Zhao, Xiao Han, and Boxuan Sun. Underwater sonar image classification using adaptive weights convolutional neural network. *Applied Acoustics*, 146:145–154, 2019.

[71] N Warakagoda and Øivind Midtgaard. Transfer-learning with deep neural networks for mine recognition in sonar images. *SAS/SAR*, 40:115–122, 2018.

[72] David P. Williams. Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2497–2502, 2016.

[73] David P. Williams. Transfer learning with sas-image convolutional neural networks for improved underwater target classification. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 78–81, 2019.

[74] Kaibing Xie, Jian Yang, and Kang Qiu. A dataset with multibeam forward-looking sonar for underwater object detection. *Scientific Data*, 9, 2022.

[75] Yichao Xu, Xingmei Wang, Kunhua Wang, Jiahao Shi, and Wei Sun. Underwater sonar image classification using generative adversarial network and convolutional neural network. *IET Image Processing*, 14(12):2819–2825, 2020.

[76] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model, 2024.

[77] Xiufen Ye, Chuanlong Li, Siyuan Zhang, Peng Yang, and Xiang Li. Research on side-scan sonar image target classification method based on transfer learning. In *OCEANS 2018 MTS/IEEE Charleston*, pages 1–6, 2018.

[78] Lyle W. Zevenbergen and Colin R. Thorne. Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12(1):47–56, 1987.

[79] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest, 2023.

[80] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2023.

[81] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation, 2024.

[82] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens, 2023.

[83] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[84] Keqing Zhu, Jie Tian, and Haining Huang. Underwater object images classification based on convolutional neural network. In *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, pages 301–305, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] The primary use of the sonar image dataset is for scientific and environmental monitoring purposes, which inherently aim to support rather than harm societal interests.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]