Journal of *Marine Science and Engineering*

MDPI

Article

Enhancing Phytoplankton Recognition Through a Hybrid Dataset and Morphological Description-Driven Prompt Learning

Yubo Huo [1], Qingxuan Lv [2,*] and Junyu Dong [2,*]

[1] Haide College, Laoshan Campus, Ocean University of China, No. 238 Songling Road, Laoshan District, Qingdao 266100, China; huoyubo@stu.ouc.edu.cn
[2] School of Computer Science and Technology, West Coast Campus, Ocean University of China, No. 1299 Sansha Road, Binhai Street, Huangdao District, Qingdao 266100, China
* Correspondence: lvqingxuan@stu.ouc.edu.cn (Q.L.); dongjunyu@ouc.edu.cn (J.D.)

Abstract

Phytoplankton plays a pivotal role in marine ecosystems and global biogeochemical cycles. Accurate identification and monitoring of phytoplankton are essential for understanding environmental dynamics and climate variations. Despite the significant progress made in automatic phytoplankton identification, current datasets predominantly consist of idealized laboratory images, leading to models that demonstrate persistent limitations in the fine-grained differentiation of phytoplankton species. To achieve high accuracy and transferability for morphologically similar species and diverse ecosystems, we introduce a hybrid dataset by integrating laboratory-based observations with in situ marine environmental data. We evaluate the performance of our dataset on contemporary deep learning models, revealing that CNN-based architectures offer superior stability (85.27% mAcc., 93.76% oAcc.). Multimodal learning facilitates refined phytoplankton recognition through the integration of visual and textual representations, thereby enhancing the model's semantic comprehension capabilities. We present a fine-tuned visual language model leveraging enhanced textual prompts augmented with expert-annotated morphological descriptions, significantly enhancing visual-semantic alignment and allowing for more accurate and interpretable recognition of closely related species (84.11% mAcc., 94.48% oAcc.). Our research establishes a benchmark dataset that facilitates real-time ecological monitoring and aquatic biodiversity research. Furthermore, it also contributes to the field by enhancing model robustness and transferability to diverse environmental contexts and taxonomically similar species.

Keywords: phytoplankton; deep learning; dataset; prompt learning; identification

check for updates

Academic Editor: Marco Cococcioni

Received: 28 July 2025
Revised: 25 August 2025
Accepted: 30 August 2025
Published: 1 September 2025

Citation: Huo, Y.; Lv, Q.; Dong, J. Enhancing Phytoplankton Recognition Through a Hybrid Dataset and Morphological Description-Driven Prompt Learning. *J. Mar. Sci. Eng.* **2025**, *13*, 1680. https://doi.org/10.3390/jmse13091680

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Phytoplankton, a diverse group of microscopic marine organisms, accounts for approximately 50% of global primary production and serves as the cornerstone of marine ecosystems [1]. These organisms play an essential role in organic matter production, the oceanic carbon cycle, and the foundation of marine food webs [2]. Given their sensitivity to environmental fluctuations, shifts in phytoplankton species composition and population dynamics act as reliable indicators of ecological changes and nutrient availability. Consequently, phytoplankton serves as a crucial biological marker for water quality assessment and is widely used in the monitoring of lakes, reservoirs, rivers, and oceans [3]. Therefore, accurate identification and classification of phytoplankton are vital for effective ecological surveillance and environmental monitoring.

Phytoplankton identification involves isolating organisms from background imagery and categorizing them into distinct species. This task is fundamental for the assessment of water quality [4]. Traditionally, phytoplankton identification substantially relies on manual microscopic examination, wherein experts categorize specimens based on morphological traits and taxonomic references. Nevertheless, this approach encounters considerable challenges arising from the extraordinary diversity of phytoplankton species, the morphological overlap between specific taxa, and the potential presence of undescribed species. These factors result in a labor-intensive and error-prone process [5]. Dependence on experienced experts further exacerbates inefficiencies, particularly when managing large volumes of samples for comprehensive analyses [1]. Therefore, there is a urgent need to automate the identification process to facilitate more efficient and scalable monitoring.

In recent years, the incorporation of machine learning (ML) techniques has significantly advanced the fields of aquatic ecology and marine biology. Automated computer vision systems driven by deep learning algorithms have demonstrated high effectiveness in processing large-scale image datasets with both rapidity and precision, thereby reducing dependence on human experts and improving the accuracy of ecological assessments [1]. The initial applications of deep learning in computer vision tasks employed neural networks, which excel in capturing complex feature representations through hierarchical data processing [6]. The development of convolutional neural networks (CNNs) [7] catalyzed a wave of innovation in image recognition, with architectures, such as those proposed by Krizhevsky et al. [8], achieving groundbreaking results on the ImageNet dataset. This success spurred further advancements, with CNNs being leading to the application of CNNs in various ecological image classification tasks, including the identification of microalgae and phytoplankton.

Multimodal learning, which integrates visual and textual data, has emerged as a powerful paradigm for improving classification performance across diverse domains. Recent advancements demonstrate that joint image–text representation learning enhances model generalizability and facilitates effective transfer learning. For instance, the Contrastive Language-Image Pre-training (CLIP) framework has demonstrated remarkable capabilities by aligning image representations with corresponding textual descriptions, allowing for better cross-modal learning [9]. Goyal et al. [10] proposed a contrastive fine-tuning method that converts downstream classification labels into text prompts and optimizes the contrastive loss between image and description embeddings. Xie et al. [11] introduced a novel method for active fine-tuning tasks (ActiveFT), which selects datasets that are similar to the distribution of the entire unlabeled pool and optimizes the model parameters in continuous spaces. Although multimodal learning is especially advantageous in phytoplankton recognition, existing models remain insufficient for high-precision identification of morphologically analogous species.

Despite advances in image recognition systems, current phytoplankton datasets exhibit significant limitations. The WHOI-Plankton dataset [12], developed by the Woods Hole Oceanographic Institution, constitutes a large-scale visual recognition dataset for phytoplankton classification. Although it encompasses an extensive range of phytoplankton categories, the dataset's effectiveness is overly dependent on high-contrast grayscale images. More recently, Li et al. [13] constructed the PMID2019 dataset for phytoplankton detection, further advancing the state of the art in image-based ecological monitoring. Despite its high resolution, the PMID2019 dataset contains solely isolated specimens captured under idealized conditions. Existing datasets based on idealized images result in reduced model performance in identifying field samples, primarily due to the presence of floating debris, organism overlap, and variations in lighting conditions during marine in

situ observations. Therefore, developing ecologically representative datasets has become a critical research priority.

In this paper, we present a novel hybrid dataset for phytoplankton image identification that comprises over 220,000 samples. We integrate laboratory-based observations with in situ marine environmental data, creating a more comprehensive resource for phytoplankton identification. Using the representative phytoplankton species *Ceratium* as an example, Figure 1 highlights key distinctions between existing datasets and our newly proposed hybrid dataset. Utilizing this dataset, we assess the performance of contemporary deep learning models. The results indicate that CNN-based approaches exhibit greater stability in phytoplankton recognition tasks compared with large-scale Transformer models. Moreover, we enhance the performance of a fine-tuned visual language model by incorporating expert-annotated morphological descriptions into the textual prompts, thereby improving recognition accuracy for closely related species.

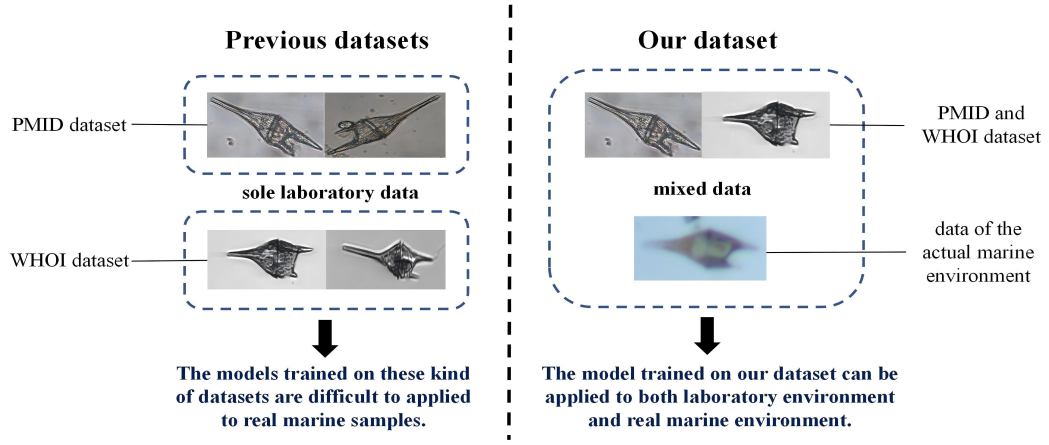

Figure 1. Our dataset offers distinct advantages over existing datasets by combining laboratory-cultured samples with field-collected marine specimens. For instance, consider the species *Ceratium*; our hybrid dataset incorporates Ceratium samples from both controlled environments (PMID and WHOI datasets) and natural marine settings. This comprehensive integration enhances the dataset's ecological relevance and demonstrates superior transferability to real-world applications.

The main contributions of this paper are as follows:

1. Development of a Hybrid Dataset: We propose a novel dataset that combines laboratory-based data with real-world environmental samples, improving the model's applicability in both laboratory-based and natural settings.
2. Evaluation of Recognition Algorithms: We compare the performances of CNNs and Transformer models using the hybrid dataset, revealing the superior performance of CNNs in phytoplankton classification. Additionally, we demonstrate the effectiveness of incorporating enhanced prompt descriptions into a fine-tuned visual language model to improve recognition accuracy.
3. Experimental Evaluation: Extensive experiments on the hybrid dataset show that the proposed approach achieves superior recognition performance, highlighting its potential for real-world applications.

The remainder of this paper is structured as follows: Section 2 reviews the evolution of phytoplankton recognition techniques and datasets. Section 3 outlines the construction of the hybrid dataset and the strategies employed for model fine-tuning. Section 4 presents experimental results and a comparative analysis of various models. Finally, Section 5 discusses the findings and suggests directions for future research.

2. Related Work

2.1. Phytoplankton Observation and Image Identification

The manual recognition of phytoplankton in images primarily relies on microscopic observation. Initially, prepared slides are examined to detect the presence of phytoplankton organisms in water samples. Upon identifying the organisms, their taxonomic classification is determined through morphological and size comparisons with established taxonomic references [5]. This process is not only labor-intensive and resource-intensive but also prone to misidentification, particularly among morphologically similar species.

To address these limitations, automated approaches based on machine learning have been increasingly adopted for phytoplankton recognition tasks. Traditional computer vision techniques, such as thresholding [14] and edge detection [3], rely on morphological features, coloration, and contour characteristics of phytoplankton. However, these methods are susceptible to background impurities, frequently resulting in suboptimal segmentation accuracy [4]. Recent advances in deep learning have introduced sophisticated models, such as Mask R-CNN [15] and YOLOs [16], which leverage their robust feature extraction and classification capabilities to achieve superior performance in phytoplankton recognition. For instance, by fusing fluorescent and bright-field images, Jia et al. [4] automated phytoplankton segmentation, reducing manual effort in water quality assessment. Nonetheless, dependency on fluorescence and limited accuracy for small algae remain challenges. Zhao et al. [17] developed an EEM-based analytical method for the precise taxonomic identification of phytoplankton. However, its validation was restricted to laboratory-cultured samples, failing to account for the complexities of aquatic environments.

2.2. Multimodal Learning and Prompt Fine-Tuning

Multimodal learning integrates various forms of single-type data, such as images, videos, and audio, combining the perceptual capabilities of each type to produce results. This provides complementary and comprehensive information that can address shortcomings such as noise and concept absence, thereby improving model performance [18]. A prominent example of this is Contrastive Language-Image Pre-training (CLIP). CLIP is trained using a newly constructed dataset containing vast pairs of images and texts, utilizing image encoders and text encoders to learn the alignment relationship between the two. CLIP pioneered multimodal learning approaches, showing promising zero-shot transferability. However, its generalization ability critically depends on high-quality image–text pairings [9].

Following the release of multimodal learning models like CLIP, several fine-tuning techniques have been proposed to adapt the model to specific tasks and achieve efficient multimodal tasks. Zhou et al. [19] proposed Context Optimization (CoOp), which optimizes prompts by introducing learnable context tokens. Although CoOp represents an advancement through text prompt optimization, it remains constrained by fixed prompt templates, limiting its adaptability to novel domains. Current models achieve basic recognition capabilities for phytoplankton identification tasks but fail to deliver robust generalization performance and effective learning of morphologically similar species.

2.3. Current Large-Scale Datasets for Phytoplankton Identification

In recent years, several phytoplankton datasets have been established to serve as references for biological identification and environmental monitoring tasks. The WHOI-Plankton dataset [12], a large-scale benchmark for plankton classification, comprises over 3.4 million expert-labeled images across 70 distinct categories, making it well suited for classification tasks. The PMID2019 dataset [13] contains more than 10,000 high-resolution microscopic images of phytoplankton across 24 categories, with each image annotated with

bounding boxes and taxonomic labels. To enhance its utility for in situ applications, the researchers employed Cycle-GAN to perform domain transfer between preserved and live cell samples, thereby enabling training and evaluation of phytoplankton detection methods. Additionally, the Kaggle-Plankton dataset [20,21] provides 30,000 low-resolution grayscale images representing 121 categories, serving as another valuable resource for phytoplankton classification research.

However, the existing phytoplankton datasets consist exclusively of high-resolution grayscale or color images, which are applicable only for identifying phytoplankton in ideal environments [12,13,20]. These artificially clean images fail to capture the complex environmental characteristics of natural marine habitats, significantly limiting the accuracy of automated phytoplankton identification systems. Consequently, models trained on these excessively cleaned datasets struggle with fine-grained classification of morphologically similar or rare species. These challenges highlight the requirement for more ecologically representative samples to enhance the robustness and transferability of models in practical applications.

3. Methodology

3.1. Data Preparation and Preprocessing

3.1.1. Data Collection

Before conducting the model evaluations, we first carried out a comprehensive data collection process to acquire a large volume of phytoplankton imagery essential for training and validation. Our image sources were diverse and strategically selected to enhance both the realism and coverage of phytoplankton species.

First, we integrated laboratory-captured datasets from the Woods Hole Oceanographic Institution (WHOI) [12] and the Plankton Microimage Dataset 2019 (PMID2019) [13]. These datasets provided well-labeled and high-quality microscopic images under laboratory-based environments, serving as valuable references for species identification and classification tasks.

In addition to relying on public datasets, we also designed an innovative in situ sampling procedure to simulate natural marine conditions more faithfully. To this end, we employed an automated sampling device currently in the development stage, capable of capturing phytoplankton samples directly from coastal water and recording their dynamic behaviors through high-resolution video streams. This device was equipped with an integrated optical microscope, set with an observation range of 20–200 μm, which is optimal for covering the majority of marine phytoplankton species [22].

To ensure a practical balance between spatial resolution and sampling efficiency, we configured the microscope at low magnification. This setting allowed us to monitor a broader field of view [23] and facilitate high-throughput data collection, which is crucial for large-scale ecological studies and real-time detection systems.

Figure 2 illustrates the outline of this process. Seawater is subjected to automatic timed pumping, filtration, and drainage through a filter pumping and discharging device controlled by an Arduino panel. Subsequently, an injection pump adjusts the flow rate of filtered seawater, which is then sampled and captured by an optical camera device. Ultimately, sample detection and tracking are performed using deep learning algorithms on a Jetson Orin NX platform.

From July to September 2024, we conducted periodic marine observations at selected coastal sites during daylight hours (10:00 AM to 3:00 PM). During these 3 months, we successfully collected a substantial volume of video data containing live phytoplankton under natural conditions. This enriched dataset captured the complexity of natural marine environments while documenting diverse morphological and behavioral traits of phy-

toplankton populations, but also laid a solid foundation for subsequent modeling and provided benchmark data for long-term ecological analysis.

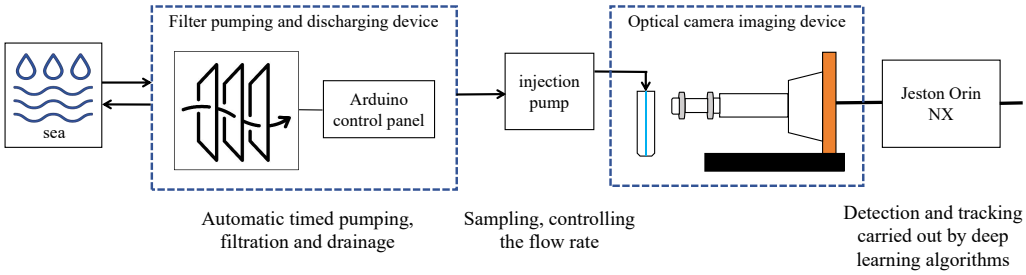

Figure 2. The process of real marine environment sample collection. We utilized an automated sampling device to capture phytoplankton samples from ocean water, with the microscope balancing the spatial resolution and sampling efficiency.

3.1.2. Data Annotation and Model Training

Next, we meticulously annotated the real-world phytoplankton detection data. More than 5000 samples were manually labeled to generate training data for our detection models. The labeling process was carried out by two annotators: one specializing in marine phytoplankton and the other possessing extensive experience in constructing phytoplankton datasets. Throughout the annotation process, we referenced the established labeling standards of the WHOI and PMID datasets [12,13], as well as authoritative handbooks on phytoplankton classification [24–26], which ensured a consistent and biologically informed categorization of samples. Each sample was assigned dual independent annotations, and discrepancies were resolved through systematic consultation of standard taxonomic references [24–26], followed by consensus discussion. This rigorous protocol guaranteed both biological accuracy and consistency within the dataset.

However, labeling real marine samples posed significant challenges. Due to the presence of environmental noise, such as suspended particles and debris in seawater, as well as occasional issues with focus or motion blur during the image acquisition phase, some images exhibited low quality. Furthermore, the morphological differences between particular phytoplankton species were subtle and ambiguous, often leading to uncertainty during classification. These issues collectively highlight a broader gap in the current state of phytoplankton visual recognition: the field still lacks large-scale, high-quality labeled datasets to serve as ground truth for robust model development.

To address this limitation, instead of discarding ambiguous data, we retained and utilized the noisy-labeled dataset to train a lightweight object detection model based on YOLOv10 [16]. The dataset was partitioned into training and validation sets at a ratio of 9:1. This decision was informed by two considerations. First, YOLOv10 offers a favorable balance between accuracy and inference speed, which is essential for handling massive data volumes in time-sensitive ecological monitoring applications. Second, its relatively simple architecture and reduced computational demands make it suitable for deployment on edge devices or in resource-constrained environments [16], such as unmanned sampling buoys or portable field analysis tools. Our sample quality control process includes three stages: (1) automatically excluding images with severe blur artifacts, (2) conducting morphological verification against established taxonomic manuals, and (3) using YOLOv10 for verification to identify annotation discrepancies.

To ensure a rigorous evaluation and prevent potential leakage, the dataset was partitioned into training, validation, and test sets strictly by sampling source and environmental context before model training. Specifically, samples collected from the same batch were retained together to avoid interference from equipment bias. Similarly, samples collected at

different times and locations were allocated to non-overlapping partitions to ensure the absence of temporal or spatial proximity between samples. This process maintains complete independence between domains, thus guaranteeing a robust assessment of generalization ability while eliminating the risk of data leakage.

By training the model directly on imperfect annotations, we aim to explore the model's robustness to noises and assess its generalization ability under real-world conditions. This experimentation phase is crucial in laying the foundation for developing more adaptive and scalable phytoplankton detection pipelines for future marine observation systems.

Table 1 summarizes the performance of the YOLOv10 series regrading accuracy and computational efficiency on the annotated dataset. The results show that the detection accuracy gradually improves as the model's parameter size increases. However, this enhancement in accuracy comes with an associated increase in computational cost. Figure 3 presents the visualization results of YOLOv10's object detection performance on unlabeled data. The YOLOv10 model demonstrates exceptional proficiency in the detection task, accurately localizing visible phytoplankton in the images, which is attributed to its optimized convolutional layer and the introduction of a partial self-attention mechanism [16]. However, while YOLOv10 performs well in coarse-grained classification tasks, its ability in fine-grained classification is limited. Some incomplete, fuzzy, and small samples remain unidentified.

Specifically, YOLOv10 demonstrated significant improvements in both detection precision and the ability to handle noisy annotations compared with earlier versions [27,28]. This makes YOLOv10 particularly suitable for real-time applications in marine environments, where data quality may be compromised and efficiency is crucial. However, the trade-off between model accuracy and computational efficiency suggests that future improvements should focus on optimizing the balance between these factors to enhance the scalability and adaptability of the system in practical scenarios.

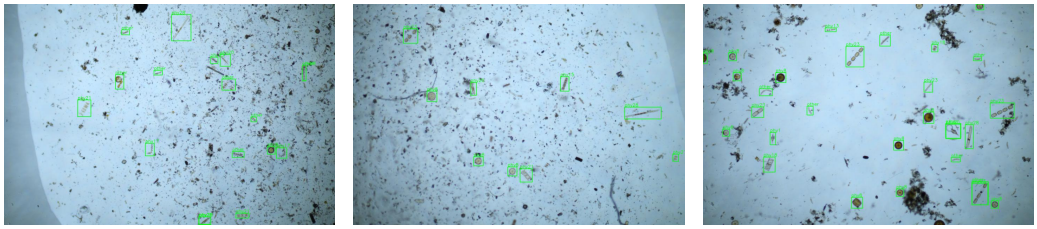

Figure 3. Visualization results of YOLOv10. The detection model achieves satisfactory object localization performance on unlabeled data, with most plankton samples accurately detected. However, its capability in fine-grained classification remains limited.

Table 1. In the evaluation and deployment efficiency comparison of phytoplankton detection models, the inference resolution is set to 640 × 640. The inference time includes both preprocessing and postprocessing, which may lead to differences from the inference speed provided by the official YOLOv10 implementation.

Method	mAP	Parameter	Infer Time (640)	FPS	Infer Time (640 Orin NX)	FPS
Yolov10s	32.80	7.2 M	31 ms	32.3	54 ms	18.5
Yolov10m	35.90	15.4 M	36 ms	27.8	55 ms	18.2
Yolov10b	43.22	19.1 M	38 ms	26.3	57 ms	17.5
Yolov10l	45.77	24.4 M	42 ms	23.8	62 ms	16.1
Yolov10x	48.60	29.5 M	45 ms	22.2	64 ms	15.6

3.1.3. Model Testing and Data Expansion

After completing the model training, we tested 5000 unlabeled images. Using the developed automated detection algorithm, we successfully identified some potential phy-

toplankton samples. Although the recognition accuracy was not ideal at this stage, this process not only provided additional sample data but also laid a valuable foundation for subsequent classification work. While the unlabeled data showed some uncertainties during initial recognition, this step was crucial for the next research phase, particularly in model fine-tuning and data annotation.

To further enrich the dataset and improve image quality, we cropped the detected phytoplankton samples. Specifically, we cropped both labeled and unlabeled data based on the bounding box information in the images. This processing method not only enhanced the clarity and quality of the images but also increased the diversity of the dataset, resulting in a dual boost in both quantity and quality.

Ultimately, we obtained approximately 60,000 high-quality phytoplankton sample images through data expansion. This data expansion significantly improved the model's generalization ability and provided abundant data support for the next research stage. At the same time, the automated detection and data augmentation strategies employed during this process laid a solid foundation for subsequent experiments and offered methods that can be referenced in similar research.

Overall, this study successfully constructed a comprehensive phytoplankton detection process by integrating dataset construction, real-world sampling, manual annotation, model training, and data expansion. This approach ensured the richness and diversity of the dataset, enhancing the model's accuracy and adaptability in real-world scenarios. This process not only offers new insights into phytoplankton detection but also serves as a methodological reference for future research.

3.1.4. Data Cleaning and Merging

After collecting and expanding the datasets, we now have three distinct datasets that cover laboratory-based environments and real ocean environments. These datasets are crucial for testing and developing a robust model for phytoplankton identification, as they encompass both theoretical and real-world data.

To enhance the contrast efficiency between theoretical samples and real samples and effectively handle the noisy samples often present in real data, we innovatively merged the three datasets into one large-scale mixed classification dataset. This merged dataset provides a more comprehensive basis for model training, as it includes a broader variety of data sources and reduces the potential for bias in sample selection.

Once the initial construction of the mixed dataset was completed, we carefully conducted manual screening and classification of the samples from these different datasets. The screening process was based on phytoplankton identification manuals [24–26], which provided important biological knowledge for classifying the samples accurately. This complex process required classifiers to possess specialized knowledge to distinguish and organize the morphological features of various phytoplankton species. Ensuring that the category information of similar samples remained consistent was essential for the reliability of the dataset and the model.

After multiple rounds of classification, check, and screening, we produced a high-quality dataset that contains approximately 220,000 images. These images are all accurately annotated with category information, ensuring that the model can learn from well-structured and reliable data. Figures 4 and 5, respectively, show all phytoplankton species detected in the training set and the validation set of our mixed database through YOLOv10 and the number of samples detected for each species.

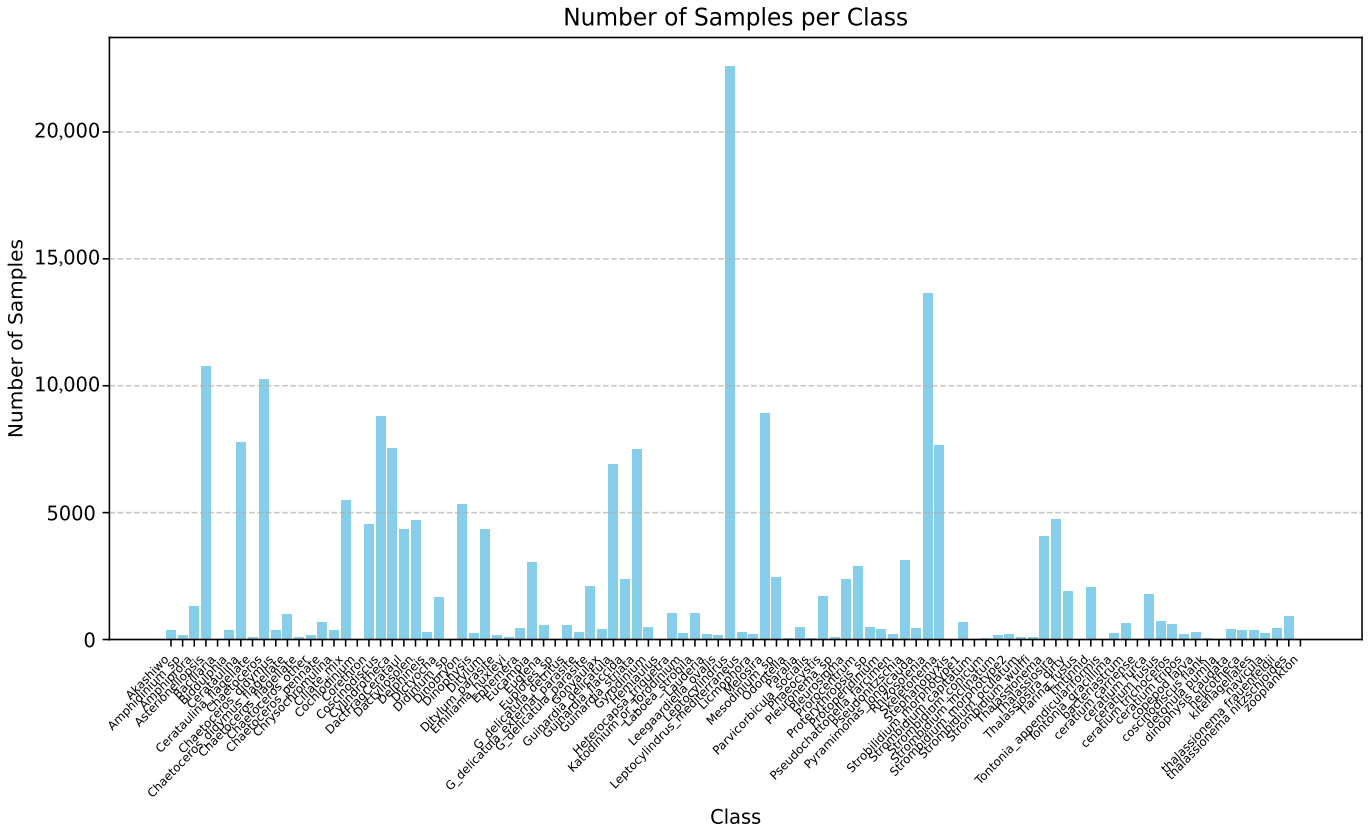

Figure 4. Category statistics of the training set for the mixed dataset.

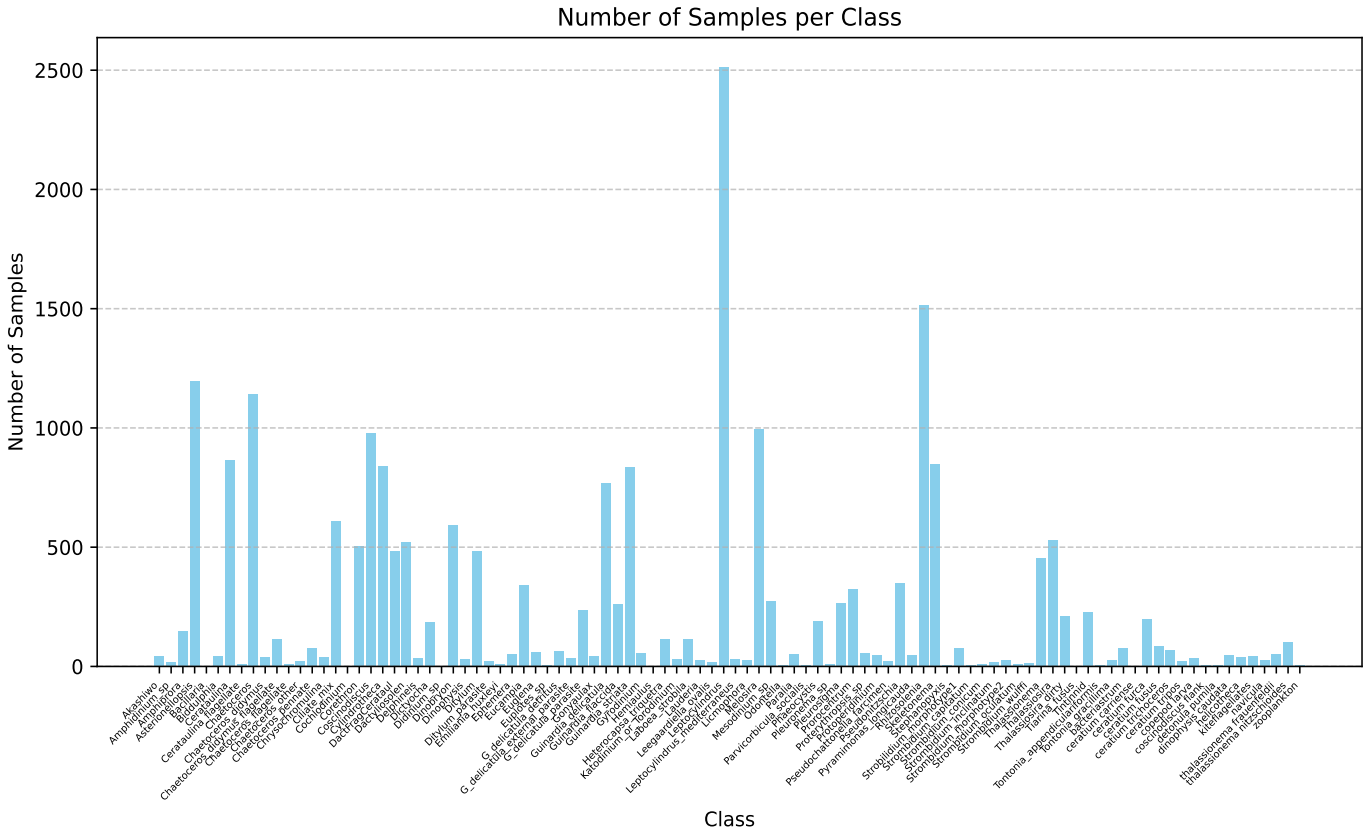

Figure 5. Category statistics of the validation set for the mixed dataset.

The development of the mixed classification dataset has proven valuable in addressing the issue of noise in real-world data. By significantly expanding the dataset, we have constructed a hybrid dataset containing 220,000 images from the WHOI dataset, the PMID dataset, and real marine environments, providing more extensive learning materials for model training. Our dataset presents significant advantages over the WHOI and PMID collections, which primarily contain idealized laboratory images, by offering authentic representations of natural marine ecosystems and in situ phytoplankton habitats. This makes our dataset particularly valuable as a reference for environmental monitoring and aquatic research. The high-quality annotations enhance the model's generalization ability, allowing it to identify and classify phytoplankton more accurately in complex and dynamic marine environments.

3.2. Description Guided Prompting Learning

3.2.1. Motivation

Traditional phytoplankton identification methods primarily rely on manual microscopic observation and morphological analysis. These approaches are not only time-consuming and labor-intensive but also highly subjective, making them unsuitable for large-scale automated identification tasks [5]. With the advancement of ocean monitoring technologies, instruments such as FlowCam [29], ZooScan [30], and IFCB [31] are now capable of continuously collecting vast amounts of phytoplankton image data, further amplifying the demand for efficient and scalable analysis methods. Against this backdrop, deep learning-based image recognition techniques have emerged as a promising solution.

In recent years, convolutional neural networks (CNNs) [32] have been widely adopted for phytoplankton image classification due to their strong visual feature extraction capabilities. Previous approaches utilized relatively simple architectures such as AlexNet [8], offering initial improvements in recognition efficiency. Subsequently, deeper models like VGGNet [33] significantly enhanced networks' abilities to perceive fine-grained morphological structures. ResNet [34], with its residual connections, addressed the vanishing gradient problem in deep networks and has since become one of the most frequently used backbones in phytoplankton identification tasks. DenseNet [35] further strengthened feature reuse and propagation by densely connecting layers, making it especially suitable for handling the diverse and imbalanced nature of phytoplankton datasets. More recently, lightweight models such as EfficientNet [36] have balanced accuracy with computational efficiency, paving the way for deploying recognition systems in edge-computing environments.

Beyond CNNs, the rapid rise of Transformer-based models in the vision domain has opened new directions for phytoplankton classification. Vision Transformer (ViT) [37] represents a landmark innovation by dividing images into patches and applying global self-attention mechanisms, enabling the model to capture long-range dependencies and subtle morphological traits more effectively. Swin Transformer [38] introduces a hierarchical design with shifted windows, balancing local and global feature modeling while maintaining high efficiency on high-resolution inputs. Efficient variants such as DeiT [39] further adapt Transformer architectures to scenarios with limited training data. This is an important consideration given the long-tail distribution and annotation scarcity in phytoplankton datasets. Hybrid models such as ConvNeXt [40], which blend CNN principles with Transformer-inspired architectural refinements, continue to broaden the design space for scientific image recognition tasks.

More recently, the emergence of vision-language models (VLMs) has brought cross-modal modeling into the spotlight for phytoplankton recognition. Unlike purely visual models, VLMs leverage textual information such as expert-provided descriptions and ecological attributes to enhance semantic understanding and generalization [41]. Among

them, CLIP [9] learns joint image–text representations through contrastive training on large-scale paired datasets, achieving impressive zero-shot performance on various visual tasks. Models like BLIP [42] were built upon this by incorporating caption generation and reconstruction tasks, further enriching cross-modal representations and improving disambiguation in fine-grained recognition scenarios.

Building upon our multi-source hybrid phytoplankton image dataset, this study aims to systematically evaluate the capabilities of deep learning models in tackling the domain-specific, fine-grained classification challenge of phytoplankton identification. We place particular emphasis on exploring model generalization and robustness across different imaging devices, environments, and class imbalances. Furthermore, we seek to integrate textual and visual information through carefully designed cross-modal alignment mechanisms, enabling the model to not only perceive visual features but also interpret semantic cues from descriptive text. Through this approach, we aim to develop a more efficient, accurate, and interpretable recognition system that advances the automation of marine ecological monitoring and offers a practical case study for the application of vision-language models in scientific domains.

3.2.2. Problem Definition

Let us denote a phytoplankton image dataset as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times 3}$ represents the i-th RGB image, and $y_i \in \{1, 2, \ldots, C\}$ is its corresponding class label, with C being the total number of phytoplankton categories. The objective is to learn a recognition function $f : \mathbb{R}^{H\times W\times 3} \to \{1, \ldots, C\}$ that can accurately classify unseen samples.

Image-Based Recognition with CNNs: In conventional image recognition pipelines, the recognition function f is typically constructed using a convolutional neural network (CNN) as the backbone. A CNN-based model extracts a visual feature representation through a parameterized mapping, as follows:

$$z_i = \phi(x_i; \theta), \quad z_i \in \mathbb{R}^d \tag{1}$$

where $\phi(\cdot)$ denotes the CNN feature extractor and θ represents the network parameters. The extracted feature z_i is then passed through a classifier (e.g., a fully connected layer) to predict the category, as follows:

$$\hat{y}_i = \arg\max_c \ \text{softmax}(Wz_i + b)_c \tag{2}$$

The model is trained by minimizing the cross-entropy loss, as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} \log p_{y_i}, \quad p_{y_i} = \frac{\exp(w_{y_i}^\top z_i)}{\sum_{c=1}^{C} \exp(w_c^\top z_i)} \tag{3}$$

This approach relies solely on visual information and does not leverage prior knowledge such as expert-written species descriptions or ecological characteristics.

Multimodal Recognition with CLIP: Recently, vision-language models (VLMs) like CLIP have introduced cross-modal alignment mechanisms that jointly model visual and textual information [9], offering a new paradigm for phytoplankton classification. Suppose that each class c is associated with a natural language description t_c, such as "a photo of *SpeciesName*, a type of phytoplankton." CLIP employs two separate encoders: an image encoder $\phi_{\text{img}}(\cdot)$ and a text encoder $\phi_{\text{text}}(\cdot)$, which project image x_i and text t_c into a shared embedding space, as follows:

$$z_i = \phi_{\text{img}}(x_i), \quad v_c = \phi_{\text{text}}(t_c), \quad z_i, v_c \in \mathbb{R}^d \tag{4}$$

The predicted label is determined by computing the cosine similarity between the image and text embeddings, as follows:

$$\hat{y}_i = \arg\max_c \cos(z_i, v_c) = \arg\max_c \frac{z_i^\top v_c}{\|z_i\| \cdot \|v_c\|} \tag{5}$$

During training, CLIP optimizes a contrastive loss function over image–text pairs, as follows:

$$\mathcal{L}_{\text{CLIP}} = -\sum_{i=1}^{N} \log \frac{\exp(\cos(z_i, v_{y_i})/\tau)}{\sum_{c=1}^{C} \exp(\cos(z_i, v_c)/\tau)}, \tag{6}$$

where τ is a temperature parameter. This framework enables the model to not only perceive visual features but also interpret rich semantic cues from text, thereby improving performance on fine-grained classification and long-tail categories.

3.2.3. Prompt Enhancement with Morphological Descriptions

In the conventional CLIP framework, the textual input for each class, t_c, is typically constructed using a generic template, such as

$$t_c^{\text{base}} = \text{“A photo of a \{label\}”} \tag{7}$$

While such templates provide a general description, they often lack the necessary semantic specificity required for fine-grained tasks like phytoplankton recognition. To address this limitation and enhance the alignment between visual features and the semantic characteristics of each class, we propose a novel approach: incorporating expert-annotated morphological descriptions into the textual prompts. These descriptions provide critical domain-specific details such as shape, size, color, structure, or arrangement—traits that are crucial for distinguishing between closely related phytoplankton species.

Let d_c represent the morphological description of class c, which encapsulates these specific features. For example, a description may specify a species as “colony-forming, needle-like, with long filaments.” We define the enhanced prompt, t_c^{enh}, as

$$\begin{aligned} t_c^{\text{enh}} &= \text{Prompt}(\texttt{template}, d_c) \\ &= \text{“A microscopic image of a \{label\},} \\ &\quad \text{which is \{description\}.”} \end{aligned} \tag{8}$$

This enhanced prompt formulation allows us to explicitly incorporate morphological details into the prompt template. For instance, consider the species *Chaetoceros*, which has the morphological description “a chain-forming diatom with long, spine-like setae.” The resulting enhanced prompt for this species is

$$\begin{aligned} t_{\texttt{Chaetoceros}}^{\text{enh}} = \text{“A microscopic image of a Chaetoceros,} \\ \text{which is a chain-forming diatom with long, spine-like setae.”} \end{aligned} \tag{9}$$

This enhanced prompt is then encoded by CLIP’s text encoder, as follows:

$$v_c = \phi_{\text{text}}(t_c^{\text{enh}}) \tag{10}$$

Subsequently, it is used in the same manner as the standard CLIP framework to compute the similarity between the text and image embeddings, as follows:

$$\hat{y}_i = \arg\max_c \cos(z_i, v_c) \tag{11}$$

where z_i represents the image embedding for the i-th image. The enhanced prompt, now informed by detailed morphological features, aims to improve the alignment between image and text representations in the joint embedding space, thereby enhancing the model's ability to distinguish among morphologically similar but taxonomically distinct phytoplankton species.

Our framework can be summarized as shown in Figure 6. In the visual component, vision classification models such as CNNs and Transformers are employed to process phytoplankton images. These models generate visual embeddings, which are then fed into a classifier to produce predictions for different classes. In the textual component, textual models leverage description-guided prompt learning. Textual descriptions of phytoplankton are processed to generate textual embeddings. These embeddings are combined with visual embeddings to produce joint representations, which are then used to refine the classification predictions.

We hypothesize that this integration of morphological descriptions into the prompt enhances the interpretability and accuracy of the model, particularly for species with subtle, but critical, morphological differences. Moreover, by grounding the prompt in domain-specific knowledge, we believe this approach could increase the model's transferability to unseen or rare species, thereby improving its generalization across diverse ecological contexts.

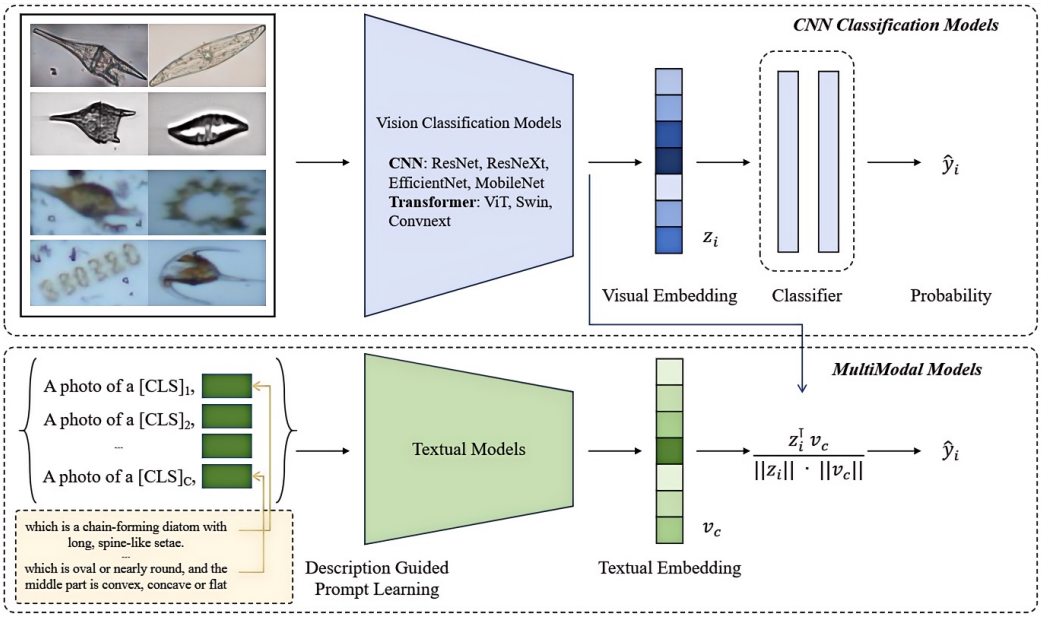

Figure 6. The processes of image recognition in CNNs and Transformers, along with prompt enhancement learning in CLIP.

4. Results

4.1. Implementation Detail

All experiments were conducted on a single NVIDIA RTX 4080 GPU, using PyTorch 2.1.0 as the training and evaluation framework. For **CNN-based models**, including ResNet, EfficientNet, and related architectures, all input images were resized to 224 × 224 pixels without applying any additional data augmentation. A batch size of 32 was used throughout the training process. The models were optimized using the Adam optimizer [7] with a fixed learning rate of 1×10^{-4}, and trained for 50 epochs. For **Transformer-based models**, including the Vision Transformer (ViT) and its variants, we adopted the same image preprocessing strategy as in CNN models by resizing all input images to 224 × 224 pixels. Similar to the CNN experiments, no additional data augmentation techniques were applied [43].

A batch size of 32 was used, and all models were trained for 50 epochs using the Adam optimizer with a fixed learning rate of 1×10^{-4}. For **vision-language models**, including CLIP and its prompt-learning variants such as CoOp and CoCoOp, we followed a similar experimental protocol. All input images were resized to 224 × 224 pixels without applying any additional data augmentation. A batch size of 32 was used during training. For CLIP, we used the pre-trained ViT-B/16 backbone and evaluated it in a zero-shot setting to assess its baseline generalization ability [9]. For CoOp [19] and CoCoOp [44], we fine-tuned the text prompts while keeping the image encoder frozen, following the standard training strategy proposed in the original papers. The training was conducted for 50 epochs using the Adam optimizer with a fixed learning rate of 1×10^{-4}. This experimental setup was designed to provide a consistent and laboratory-based environment for evaluating the baseline performance of CNN models on phytoplankton classification tasks, without the influence of enhanced preprocessing or data-level modifications.

4.2. Comparison Methods

To comprehensively evaluate the performance of our proposed method, we compare it with three categories of baseline approaches: (1) CNN-based architectures, (2) Transformer-based models, and (3) vision-language models (VLMs). These representative models cover both unimodal and multimodal paradigms, as well as classic and state-of-the-art techniques in cross-domain recognition.

(a) CNN-Based Architectures

ResNet-50/ResNet-101: ResNet-50 and ResNet-101 are deep convolutional neural networks designed to address the vanishing gradient problem in very deep architectures. They utilize residual blocks with skip connections to enable stable training and efficient gradient flow, even in networks with 50 or 101 layers. The architecture employs a bottleneck structure (1 × 1–3 × 3–1 × 1 convolutions) to reduce computational complexity while maintaining performance. Key techniques include residual learning, which allows the network to learn residual mappings, and batch normalization, which stabilizes training and accelerates convergence [34]. While ResNet excels in classification tasks and is widely used in computer vision, its reliance on local convolutions limits its ability to model global context [45], which can affect performance in complex tasks such as phytoplankton classification.

EfficientNet-b0/EfficientNet-b4: EfficientNet-b0 and EfficientNet-b4 are convolutional neural networks optimized through Neural Architecture Search (NAS) and a compound scaling strategy that balances depth, width, and resolution for superior accuracy–efficiency trade-offs. The architecture adopts inverted residual blocks from MobileNetV2, using depthwise separable convolutions to reduce computational overhead, while integrating Squeeze-and-Excitation (SE) modules to enhance feature representation through channel attention. Key techniques include compound scaling, which jointly optimizes network dimensions, and NAS-based architecture discovery to automate design and improve efficiency. EfficientNet achieves high classification accuracy with fewer parameters and lower computational cost, making it suitable for resource-constrained environments, such as mobile and embedded devices [36]. However, its reliance on specialized components such as SE modules and inverted residuals may impact generalization in certain tasks, and optimal performance often depends on hardware accelerators like GPUs and TPUs [46].

(b) Transformer-Based Vision Models

ViT-Small/ViT-Base: ViT-Small and ViT-Base are variants of the Vision Transformer (ViT) architecture, which replaces traditional convolutional layers with self-attention mechanisms to enable global context modeling and long-range dependency capture. The architecture divides input images into non-overlapping patches, embeds them into a high-dimensional feature space, and processes them through multiple Transformer encoder

layers with multi-head self-attention (MHSA) and feedforward networks. Key techniques include MHSA for global relationship modeling, patch embedding for sequence representation, layer normalization (LN) for training stability, and pre-training on large-scale data to compensate for the lack of inductive biases like convolutions. ViT excels in capturing fine-grained morphology and offers a global receptive field, achieving superior performance on large-scale image classification tasks with abundant training data [37]. However, it requires extensive computational resources and large datasets for optimal performance, struggles with cross-domain generalization, and incurs higher computational costs than traditional CNNs, leading to slower inference on resource-constrained hardware [47].

Swin-Small/Swin-Base: Swin Transformer is a hierarchical Vision Transformer that improves the scalability and efficiency of ViT by introducing window-based self-attention. It computes self-attention within local windows and shifts the windows between layers to enable cross-window information exchange. The architecture adopts a multistage design with patch merging to gradually reduce spatial resolution and build multiscale representations, making it suitable for dense prediction tasks. Key components include Window-based Multi-Head Self-Attention (W-MHSA), shifted windowing, and patch merging. Compared with ViT, Swin Transformer achieves better computational efficiency, stronger generalization across resolutions, and improved performance on downstream tasks [38]. However, its local attention in early layers may limit the modeling of long-range dependencies [48].

(c) Vision-Language Models (VLMs)

CLIP/CoOp: CLIP is a vision-language model trained via contrastive learning on large-scale image–text pairs, using a dual-tower architecture with a vision encoder (e.g., ViT) and a text encoder (e.g., Transformer) to align visual and textual features. It excels in zero-shot transfer and open-vocabulary recognition but requires high-quality image–text pairs for effective generalization [9]. CoOp enhances CLIP by optimizing context tokens in text prompts, improving zero-shot classification without retraining the entire model, though it struggles with unseen domains due to fixed prompts [19].

4.3. Experiment Results

CNN-Based Models: As summarized in Table 2, compared with traditional convolutional neural networks (CNNs) such as ResNet, ResNeXt, and EfficientNet, our method shows consistent improvements across multiple evaluation settings. For instance, our CLIP-based ResNet101 achieves the highest mean accuracy (85.27%) among all models, and matches or exceeds performance in class-specific tasks such as Cls15 and Cls48. This indicates not only superior accuracy but also better adaptability to datasets with varying numbers of classes. In contrast, conventional CNNs often struggle with small-class tasks or show less balanced performance across different settings.

The improvement can be attributed to the semantic richness introduced by the morphological prompt design. While CNNs rely heavily on large amounts of labeled data to learn discriminative features, our method benefits from CLIP's pre-trained vision-language knowledge and enhanced prompt expressiveness. This results in stronger generalization and robustness, making it a promising direction for vision tasks where semantic understanding and limited supervision are crucial.

Transformer-Based Models: In this experiment, we further evaluate our method against various Vision Transformer (ViT)-based models on the mixed phytoplankton dataset. The results are presented in Table 3. Specifically, we compare our CLIP-enhanced approach (ResNet50, ResNet101, and ViT-Base backbones) with ViT-Small, ViT-Base, ConvNext, and Swin Transformer variants. As shown in Table 3, although our method using ResNet101 achieves lower mean accuracy (85.27%) than Convnext-Base (85.72%), it outperforms

Convnext-Base in class-specific performance for larger class groups such as Cls92 (89.67% vs. 89.20%) and maintains strong results across other class groups.

Table 2. Comparison of different CNN methods on the mixed dataset.

Model	mAcc.	oAcc.	Cls8	Cls15	Cls48	Cls92
ResNet50 [34]	83.53	94.48	91.67	70.00	93.36	85.92
ResNet101 [34]	84.11	94.48	75.00	85.00	93.36	91.55
ResNeXt50 [49]	84.36	94.41	58.33	85.00	94.01	86.85
ResNeXt101 [49]	77.62	92.14	41.67	65.00	94.66	89.20
MobileNet [28]	83.33	93.92	66.67	82.50	94.01	89.20
Efficient-b0 [36]	83.74	94.40	83.33	92.50	94.14	84.98
Efficient-b4 [36]	84.54	94.54	83.33	80.00	93.75	88.73
Ours-ResNet50	84.14	93.20	91.67	72.50	88.54	90.12
Ours-ResNet101	85.27	93.76	91.67	75.00	88.54	89.67

Notably, the performance of ViT-based models on this dataset is less stable than CNN-based counterparts. For example, ViT-Base yields only 80.66% mean accuracy and shows poor classification performance on smaller class subsets like Cls15 (52.50%) and Cls8 (58.33%). In contrast, CNN-based architectures such as Convnext-Base and Swin-Base consistently deliver more balanced results across all subsets, indicating a stronger adaptability to the characteristics of this dataset. Our CLIP-based ViT model also demonstrates limited performance (72.62% mAcc., 50.00% on Cls15), further supporting this trend.

The observed discrepancy in performance can be attributed to the inherent bias present in phytoplankton image data. In contrast with natural scene datasets, these samples are captured with minimal background variation and contain limited semantic context, focusing solely on morphological differences among phytoplankton species. ViT models, which rely heavily on global attention mechanisms and benefit from rich spatial and contextual diversity in images, struggle to extract discriminative features in such constrained settings. In contrast, CNNs, with their local receptive fields and inductive biases, are better suited for capturing fine-grained morphological patterns, resulting in more stable and effective learning for this task. The characteristics of CNNs render them particularly suitable for deployment on autonomous sampling platforms, as they preserve local feature extraction capabilities while optimizing the architecture for phytoplankton identification. These findings demonstrate that fine-grained morphological patterns provide greater discriminative power than contextual information in phytoplankton taxonomy, thereby contributing to advancements in automated phytoplankton identification technologies.

Table 3. Comparison of different ViT methods on the mixed dataset.

Model	mAcc.	oAcc.	Cls8	Cls15	Cls48	Cls92
ViT-Small [37]	80.44	94.03	66.67	67.50	94.01	84.51
ViT-Base [37]	80.66	93.60	58.33	52.50	93.10	85.92
Swin-Small [38]	83.39	94.19	83.33	67.50	94.53	91.08
Swin-Base [38]	84.12	94.61	84.33	68.03	94.98	92.02
Convnext-Small [40]	84.71	94.68	66.67	82.50	93.36	90.14
Convnext-Base [40]	85.72	95.05	66.67	82.50	93.88	89.20
Ours-ResNet50	84.14	93.20	91.67	72.50	88.54	90.12
Ours-ResNet101	85.27	93.76	91.67	75.00	88.54	89.67
Ours-ViTBase	72.62	90.10	66.67	50.00	88.80	92.02

Vision-Language Models: In Table 4, we compare our proposed method with several vision-language models (VLMs), including CLIP and CoOp variants, across different back-

bone architectures. While CLIP-based models such as CLIP-Res101 demonstrate relatively high performance (e.g., 84.92% mAcc., 88.28% on Cls48), their effectiveness remains limited when applied directly to phytoplankton classification. Notably, CoOp-based models, which rely on learnable context without introducing domain-specific semantics, perform poorly across all class subsets, with mAcc. values near 40% and complete failure in subsets (Cls8: 0.00%). We hypothesize that the performance differences between CLIP and CoOp are mainly attributable to their distinct adaptation strategies for phytoplankton recognition. The pre-trained visual language alignment of CLIP enables superior generalization when addressing ecological diversity within mixed datasets. In contrast, the fixed prompt template employed by CoOp fails to adequately capture essential taxonomic and morphological features. This divergence becomes particularly evident when processing real-time samples that contain debris or overlapping organisms. CoOp's context learning mechanism demonstrates pronounced sensitivity to label quality due to its reliance on clean categorical annotations. In contrast, CLIP's contrastive pre-training architecture offers inherent robustness against such noise.

Table 4. Comparison of different VLM-based methods on the mixed dataset.

Backbone	mAcc.	oAcc.	Cls8	Cls15	Cls48	Cls92
CLIP-Res50 [9]	83.74	93.12	83.33	75.00	87.22	90.12
CLIP-Res101 [9]	84.92	93.55	100.0	80.00	88.28	88.73
CLIP-ViT-Base [9]	72.14	89.08	66.67	50.00	87.92	86.85
CoOp-Res50 [19]	40.82	68.06	0.00	7.50	74.34	65.73
CoOp-Res101 [19]	40.88	69.23	0.00	7.50	72.34	63.85
CoOp-ViT/Base [19]	10.72	40.85	0.00	0.00	62.50	12.68
Ours-ResNet50	84.14	93.20	91.67	72.50	88.54	90.12
Ours-ResNet101	85.27	93.76	91.67	75.00	88.54	89.67
Ours-ViTBase	72.62	90.10	66.67	50.00	88.80	92.02

These results highlight a key limitation of existing pre-trained or prompt-tuned VLMs: they lack explicit morphological guidance crucial for fine-grained biological classification. Phytoplankton categories are characterized primarily by subtle morphological variations rather than diverse visual or contextual cues. Therefore, even when combined with automatic prompt tuning, general-purpose vision-language pre-training fails to capture the critical features needed for accurate classification in this domain.

By contrast, our method introduces morphology-aware semantic information directly into the prompt design. This targeted integration provides the model with more relevant and discriminative cues tailored to the structural features of phytoplankton species. As shown in Table 4, our approach with a ResNet101 backbone achieves the highest overall performance (85.27% mAcc., 93.76% oAcc.) and maintains consistent strength across all class groups. These results demonstrate that injecting domain-specific morphological knowledge into prompt construction significantly enhances the classification capability of multimodal models for specialized tasks like phytoplankton analysis.

Prompt Template Study: Table 5 presents an ablation study comparing the effectiveness of basic category prompts t_c^{base} and morphology-enhanced prompts t_c^{enh} within our description-guided prompting framework. The results clearly demonstrate the advantage of incorporating morphological descriptions into the prompt design. Specifically, the model using t_c^{enh} achieves higher mean accuracy (84.11% vs. 83.74%) and overall accuracy (94.48% vs. 93.12%), suggesting that even a modest refinement in the textual prompt can lead to more accurate classification.

Table 5. Ablation study for the proposed description-guided prompting learning.

Model	mAcc.	oAcc.	Cls8	Cls15	Cls48	Cls92
t_c^{base}	83.74	93.12	83.33	75.00	87.22	90.12
t_c^{enh}	84.11	94.48	91.67	72.50	88.54	90.12

Beyond enhancing accuracy, the morphological prompt improves model interpretability and biological fidelity. The performance gains are especially notable in subsets with smaller class counts or finer-grained distinctions. For example, in Cls8, the accuracy improves significantly from 83.33% to 91.67%. This suggests that morphology-enhanced prompts provide more discriminative guidance by emphasizing structural differences among species. For morphologically similar species, the enhanced prompts explicitly encode diagnostic features, which is a capability particularly crucial in ecological studies where misclassification may compromise biodiversity assessments. In contrast, simple category names lack this depth and fail to inform the model of subtle inter-class variations sufficiently.

Moreover, the framework incorporates domain knowledge via natural language processing, enabling adaptation to newly discovered species through prompt modifications without architectural changes. This positions our approach as a performance-enhancing methodology and an interpretability interface between computational languages and biological knowledge.

Overall, this comparison highlights the importance of domain-specific semantic enrichment in prompt construction. By embedding morphological knowledge directly into the language input, the model is better equipped to leverage the multimodal alignment capabilities of vision-language models, resulting in more robust and semantically aware classification performance. However, the effectiveness of this method needs to be further verified through multiple rounds of experiments.

5. Conclusions

This paper presents a novel approach to phytoplankton recognition by integrating deep learning techniques with a hybrid dataset. By combining laboratory observations and in situ marine environmental data, we construct a hybrid dataset containing over 200,000 samples, addressing the limitations of existing datasets and offering broader references for environmental monitoring while improving model generalization. We evaluated the performance of our dataset on contemporary deep learning models, revealing that CNN-based architectures show higher stability. Moreover, we fine-tuned a visual language model with expert-annotated morphological descriptions, which achieved improved accuracy and stability for morphologically similar species classification. Overall, our method offers a scalable and efficient solution for automated ecological monitoring and water quality assessment, providing novel insights for phytoplankton identification.

Although our experimental results demonstrate that this proposed approach outperforms traditional methods, the results represent a single experimental trial without repeated validation. Verification through multiple independent test cycles to establish statistical reliability and reproducibility is required. Future work will systematically evaluate the method's accuracy through repeated experiments incorporating morphological prompt enhancement across diverse dataset configurations. Moreover, the efficacy of our approach remains limited in identifying novel species due to dependence on comprehensive datasets containing all targeted species. Based on our hybrid dataset, our future work plans to employ unsupervised representation learning with large-scale unlabeled real-time observations to enhance the model's generalization.

Author Contributions: Conceptualization, Y.H. and J.D.; methodology, Y.H.; validation, Y.H.; formal analysis, Y.H.; investigation, Y.H.; data curation, Q.L.; writing—original draft preparation, Y.H.; writing—review and editing, Y.H., Q.L., and J.D.; visualization, Y.H.; supervision, J.D.; project administration, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Sanya Science and Technology Special Fund under Grant 2022KJCX92.

Data Availability Statement: The hybrid dataset and related morphological descriptions can be accessed at https://pan.baidu.com/s/1i_p3CDPs0yVBtEdUqeLNig?pwd=quss (accessed on 27 May 2025) or in Figshare at https://doi.org/10.6084/m9.figshare.29447705 (accessed on 13 August 2025). Relevant codes and data presented in this study are available upon request from the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Marzidovšek, M.; Mozetič, P.; Francé, J.; Podpečan, V. Computer Vision Techniques for Morphological Analysis and Identification of Two Pseudo-nitzschia Species. *Water* **2024**, *16*, 2160. [CrossRef]
2. Yu, Y.; Li, Y.; Sun, X.; Dong, J. MPT: A large-scale multiphytoplankton tracking benchmark. *Intell. Mar. Technol. Syst* **2024**, *2*, 35. [CrossRef]
3. Natchimuthu, S.; Chinnaraj, P.; Parthasarathy, S.; Senthil, K. Automatic identification of algal community from microscopic images. *Bioinform. Biol. Insights* **2013**, *7*, 327–334. [CrossRef]
4. Jia, R.; Yin, G.; Zhao, N.; Chen, X.; Xu, M.; Hu, X.; Huang, P.; Liang, P.; He, Q.; Zhang, X. Phytoplankton Image Segmentation and Annotation Method Based on Microscopic Fluorescence. *J. Fluoresc.* **2025**, *35*, 369–378. [CrossRef] [PubMed]
5. Dimitrovski, I.; Kocev, D.; Loskovska, S.; Džeroski, S. Hierarchical classification of diatom images using ensembles of predictive clustering trees. *Ecol. Inform.* **2012**, *7*, 19–29. [CrossRef]
6. Yann, L.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
7. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization, Ver. 9. *arXiv* **2017**, arXiv:1412.6980.
8. Krizhevsky, A.; Sutskever, I.E.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Commun* **2017**, *60*, 84–90. [CrossRef]
9. Radford, A.; Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision, Ver. 1. *arXiv* **2021**, arXiv:2103.00020.
10. Goyal, S.; Kumar, A.; Garg, S.; Kolter, Z.; Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In Proceedings of the CVPR, Vancouver, BC, Canada, 18–22 June 2023; pp. 19338–19347.
11. Xie, Y.; Lu, H.; Yan, J.; Yang, X.; Tomizuka, M.; Zhan, W. Active Finetuning: Exploiting Annotation Budget in the Pretraining-Finetuning Paradigm. In Proceedings of the CVPR, Vancouver, BC, Canada, 18–22 June 2023; pp. 23715–23724.
12. Orenstein, E.C.; Beijbom, O.; Peacock, E.E.; Sosik, H.M. WHOI-Plankton- A Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification, Ver. 1. *arXiv* **2015**, arXiv:1510.00745.
13. Li, Q.; Sun, X.; Dong, J.; Song, S.; Zhang, T.; Liu, D.; Zhang, H.; Han, S. Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. *ICES J. Mar. Sci.* **2020**, *77*, 1427–1439. [CrossRef]
14. Bi, H.; Guo, Z.; Benfield, M.C.; Fan, C.; Ford, M.; Shahrestani, S.; Sieracki, J.M. A Semi-Automated Image Analysis Procedure for In Situ Plankton Imaging Systems. *PLoS ONE* **2015**, *10*, e0127121. [CrossRef] [PubMed]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
16. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection, Ver. 2. *arXiv* **2024**, arXiv:2405.14458.
17. Zhao, N.; Zhang, X.; Yin, G.; Yang, R.; Hu, L.; Chen, S.; Liu, J.; Liu, W. On-line analysis of algae in water by discrete three-dimensional fluorescence spectroscopy. *Opt. Express* **2018**, *26*, A251–A259. [CrossRef] [PubMed]
18. Bayoudh, K.; Knani, R.; Hamdaoui, F.; Mtibaa, A. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.* **2022**, *38*, 2939–2970. [CrossRef]
19. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to Prompt for Vision-Language Models, Ver. 6. *arXiv* **2022**, arXiv:2109.01134.
20. Zheng, H.; Wang, R.; Yu, Z.; Wang, N.; Gu, Z.; Zheng, B. Automatic plankton image classification combining multiple view features via multiple kernel learning. *BMC Bioinform* **2017**, *18*, 57. [CrossRef]
21. Cowen R.K.; Guigand C.M. In situ ichthyoplankton imaging system (*I*SIIS): System design and preliminary results. *Limnol. Oceanogr. Methods* **2008**, *6*, 126–132. [CrossRef]

22. Brotas, V.; Tarran, G.A.; Veloso, V.; Brewin, R.; Wooodward, E.S.; Airs, R.; Beltran, C.; Ferreira, A.; Groom, S.B. Complementary Approaches to Assess Phytoplankton Groups and Size Classes on a Long Transect in the Atlantic Ocean. *Sec. Mar. Ecosyst. Ecol.* **2022**, *8*. [CrossRef]
23. Bartlett, B.; Santos, M.; Dorian, T.; Moreno, M.; Trslic, P.; Dooly, G. Real-Time UAV Surveys with the Modular Detection and Targeting System: Balancing Wide-Area Coverage and High-Resolution Precision in Wildlife Monitoring. *Remote Sens.* **2025**, *17*, 879. [CrossRef]
24. China Species Library. Available online: https://species.sciencereading.cn/biology/v/botanyIndex/122/ZLZK.html (accessed on 6 March 2025)
25. Database of Plant Bioresources in the Coastal Zone of China (NBSDC-DB-22). Available online: http://algae.yic.ac.cn/taxa/doku.php?id=start (accessed on 8 March 2025)
26. Qian, S.; Liu, D.; Sun, J. *Marine Phycology*; China Ocean University Press: Qingdao, China, 2014.
27. Yu, Y.; Lv, Q.; Li, Y.; Wei, Z.; Dong, J. PhyTracker: An Online Tracker for Phytoplankton. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *35*, 2932–2944. [CrossRef]
28. Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
29. Sieracki, C.K.; Sieracki, M.E.; Yentsch, C.S. An imaging-in-flow system for automated analysis of marine microplankton. *Mar. Ecol. Prog. Ser.* **1998**, *168*, 285–296. [CrossRef]
30. Grosjean, P.; Picheral, M.; Warembourg, C.; Gorsky, G. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. *ICES J. Mar. Sci.* **2004**, *61*, 518–525. [CrossRef]
31. Olson, R.J.; Sosik, H.M. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. *Limnol. Oceanogr. Methods* **2007**, *5*, 195–203. [CrossRef]
32. Nardelli, S.C.; Gray, P.C.; Schofield, O. A Convolutional Neural Network to Classify Phytoplankton Images Along the West Antarctic Peninsula. *Mar. Technol. Soc. J.* **2022**, *56*, 45–57. [CrossRef]
33. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition, Ver. 6. *arXiv* **2015**, arXiv:1409.1556.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition, Ver. 1. *arXiv* **2015**, arXiv:1512.03385.
35. Huang, G.; Liu, Z.; Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks, Ver. 5. *arXiv* **2018**, arXiv:1608.06993.
36. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, Ver. 5. *arXiv* **2020**, arXiv:1905.11946.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Ver. 2. *arXiv* **2021**, arXiv:2010.11929.
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, Ver. 2. *arXiv* **2021**, arXiv:2103.14030.
39. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention, Ver. 2. *arXiv* **2021**, arXiv:2012.12877.
40. Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s, Ver. 2. *arXiv* **2022**, arXiv:2201.03545.
41. Li, Z.; Wu, X.; Du, H.; Liu, F.; Nghiem, H.; Shi, G. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges, Ver. 6. *arXiv* **2025**, arXiv:2501.02189.
42. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, Ver. 2. *arXiv* **2022**, arXiv:2201.12086.
43. O'Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks, Ver. 2. *arXiv* **2015**, arXiv:1511.08458.
44. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional Prompt Learning for Vision-Language Models, Ver. 2. *arXiv* **2022**, arXiv:2203.05557.
45. Zhang, Y.; Wei, D.; Qin, C.; Wang, H.; Pfister, H.; Fu, Y. Context Reasoning Attention Network for Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 4258–4267.
46. Feng, Z.; Yang, J.; Chen, L.; Chen, Z.; Li, L. An Intelligent Waste-Sorting and Recycling Device Based on Improved EfficientNet. *Int. J. Environ. Res. Public Health* **2022**, *19*, 15987. [CrossRef] [PubMed]
47. Ishikawa, M.; Ishibashi, R.; Lin, M. Norm-Regularized Token Compression in Vision Transformer Networks. In Proceedings of the 6th International Symposium on Advanced Technologies and Applications in the Internet of Things, Shiga, Japan, 19–22 August 2024.

48. Wei, C.; Duke, B.; Jiang, R.; Aarabi, P.; Taylor, G.W.; Shkurti, F. Sparsifiner: Learning Sparse Instance-Dependent Attention for Efficient Vision Transformers, Ver. 1. *arXiv* **2023**, arXiv:2303.13755.
49. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5987–5995.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.