RESEARCH-ARTICLE

Marine Biodiversity for an Inclusive Society: Unifying Species Classification through Dataset Aggregation

RICARDO J.M. VEIGA, University of Algarve, Faro, Faro, Portugal

JOÃO M.F. RODRIGUES, University of Algarve, Faro, Faro, Portugal

Published: 31 July 2025

Citation in BibTeX format

DSAI 2024: 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion
November 13 - 15, 2024
Abu Dhabi, United Arab Emirates

DSAI '24: Proceedings of the 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (November 2024)
https://doi.org/10.1145/3696593.3696637
ISBN: 9798400707292

Marine Biodiversity for an Inclusive Society: Unifying Species Classification through Dataset Aggregation

Ricardo J.M. Veiga
Faculty of Science and Technology (FCT), NOVA LINCS & Institute of Engineering (ISE), Universidade do Algarve
Portugal
rjveiga@ualg.pt

João M.F. Rodrigues
NOVA LINCS & Institute of Engineering (ISE), Universidade do Algarve
Portugal
jrodrig@ualg.pt

Figure 1: Marine biodiversity datasets examples, with their respective image proportions intact, see text for details.

Abstract

An inclusive society actively seeks the equitable and respectful participation of all its members, regardless of their differences. This concept goes beyond tolerating diversity; it involves valuing and respecting each individual, ensuring everyone has equal access to information and opportunities, and actively participating in all aspects of life. To have a full inclusive society, we have to measure and monitor our impact in the environment, specifically in the oceans and marine life. This paper addresses this challenge by proposing a framework that leverages data aggregation and advanced machine learning techniques for Fine-Grained Visual Classification of marine species. Our methodology employs the Swin Transformer architecture, enhanced with the Fine-Grained Visual Classification Plug-in Module, to process and classify diverse marine datasets. We aggregated multiple marine datasets, preprocessed them to eliminate invalid entries, and trained our model on the refined dataset. Our findings demonstrate that dataset aggregation significantly enhances model accuracy and robustness, especially for large-scale models. Notably, the aggregated data model achieved 94.75% overall accuracy on a dataset comprising 2,548 classes and 391,374 images, compared to 85.93% on individual datasets like WildFish++.

DSAI 2024, November 13–15, 2024, Abu Dhabi, United Arab Emirates

ACM ISBN 979-8-4007-0729-2/24/11
https://doi.org/10.1145/3696593.3696637

CCS Concepts

• **Computing methodologies** → **Object identification**.

Keywords

Computer Vision, Data Aggregation, Data Fusion, Fine-Grained Visual Classification, Swin Transformer

ACM Reference Format:
Ricardo J.M. Veiga and João M.F. Rodrigues. 2024. Marine Biodiversity for an Inclusive Society: Unifying Species Classification through Dataset Aggregation. In *11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2024), November 13–15, 2024, Abu Dhabi, United Arab Emirates.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3696593.3696637

1 Introduction

Regardless of differences, an inclusive society works to ensure that all of its members participate fairly and with respect. This idea encompasses more than just accepting variety; it also entails actively engaging in all facets of life, valuing and respecting each individual, and ensuring that everyone has equal access to opportunities. On the other hand, Human-Machine Collaboration (HMC) requires that machines, in the broadest sense, be designed to work together or learn how to work with humans. This means that hardware, software, and interfaces must be able to identify the user's unique needs and behaviors as well as the context in which they are used, and the surrounding environment, in order to enable on-the-spot cooperation with humans.

In that sense, to be a full inclusive society we must take advantage of HMC and monitor the human impact in the environment, specifically in the oceans and marine life, as the ability to access and utilize information is crucial for informed decision-making

and sustainable practices. In marine biodiversity monitoring, the challenge of information exclusion, or info-exclusion, can significantly impede efforts to understand and protect marine ecosystems. Information exclusion not only limits scientific research but also hinders public awareness and policymaking, which are essential for effective environmental conservation. This paper aims to bridge this gap by leveraging advanced machine learning techniques to enhance the monitoring and analysis of marine biodiversity.

Marine ecosystems are complex and dynamic, populated with a vast array of species that are often difficult to classify and monitor due to their diverse features, and sometimes subtle distinctions. Traditional methods of species identification and monitoring are labor-intensive and often require expert knowledge, which can be a limiting factor in large-scale environmental studies. Moreover, the availability of large, annotated datasets, necessary for training robust machine learning models, is often constrained by logistical and financial barriers. This scarcity of comprehensive datasets contributes to the info-exclusion problem, where valuable data remain inaccessible or underutilized.

In response to these challenges, we propose a framework that aggregates data from multiple marine biodiversity datasets to build a more comprehensive and diverse dataset for training machine learning models. Our approach utilizes the Shifted window (Swin) Transformer architecture, which has demonstrated superior performance in visual recognition tasks, enhanced with the FGVC Plug-in Module (FGVC-PIM) to improve the model's ability to distinguish between closely related species.

The aggregated dataset incorporates several marine datasets, including new additions that broaden its scope and enhance representativeness. Figure 1 provides a visual overview of the diverse images from these various datasets (Croatian Fish Dataset [12], DeepFish [17], Fish4Knowledge [9], FishCLEF-2015 [13], FishNet [14], OBSEA [10], OzFish [1], QUT Fish Dataset [2], SEAMAPD21 [4], WhaleShark [11], WildFish [25], and WildFish++ [26]; see details in Sec. 2.1), illustrating the range of species, environments, and image qualities encompassed in our study. After a thorough preprocessing, which includes the removal of corrupted and duplicated images and invalid class names, the final dataset comprises 2,548 classes and 391,374 images.

Our experiments demonstrate that models trained on this aggregated dataset achieve significantly higher accuracy and robustness compared to those trained on smaller, individual datasets. The aggregated data model achieved an overall accuracy of 94.75%, indicating the clear impact of data aggregation. Specific improvements were observed in datasets with a high diversity of species and challenging conditions for image capture. The WildFish++ dataset [26], for example, achieved a notable increase in accuracy.

By aggregating datasets, we mitigate the limitations of individual datasets, such as class imbalance and variability in image quality and capture conditions. The added diversity and volume of data allow the model to learn more robust and generalizable features, particularly beneficial for large models trained on numerous classes. This approach helps the model to better capture subtle differences between closely related species, ultimately leading to improved classification performance.

This paper introduces a novel approach to addressing information exclusion in marine biodiversity monitoring by utilizing data aggregation and advanced machine learning techniques. Our framework, while improving species classification accuracy, supports environmental sustainability and enhances public access to ecological data. By simplifying the interpretation of complex marine data, this method promotes inclusivity, enabling individuals from diverse backgrounds, including those with disabilities or from underserved communities, to engage more effectively with critical environmental information, contributing to bridge the digital divide in ecological monitoring, supporting a more informed and diverse public.

In that way, our research not only advances Fine-Grained Visual Classification (FGVC) but also plays a crucial role in democratizing access to scientific knowledge, making environmental monitoring more accessible. By offering precise and user-friendly tools, the framework fosters more equitable participation in environmental decision-making processes, fighting information exclusion. Future applications could extend these benefits, enabling communities worldwide to contribute meaningfully to the preservation and understanding of marine ecosystems, ensuring broader engagement with sustainability efforts.

The structure of this paper is as follows: Section 2 provides an overview of the related work in marine biodiversity monitoring and machine learning techniques. Section 3 describes the datasets used in this study. Section 4 details the methodology, including data aggregation, preprocessing, and the architecture of the Swin Transformer and FGVC-PIM module. Section 5 presents the experiments, results, and discussion, highlighting the improvements in accuracy and robustness achieved by the proposed approach. Finally, Section 6 concludes the paper, emphasizing the contributions to sustainability and the fight against information exclusion, and suggests directions for future work.

2 Related Work

This section reviews key literature on machine learning advancements in marine biodiversity monitoring, focusing on data aggregation, FGVC, and advanced neural architectures. We examine studies addressing challenges in species identification, including the need for large datasets and distinguishing closely related species. The review covers dataset aggregation techniques, FGVC approaches like attention mechanisms, and the application of transformer architectures in visual tasks. We also explore plug-in modules enhancing model performance in marine species classification. Throughout, we emphasize how these advancements combat info-exclusion and improve marine biodiversity research, providing insights into current state-of-the-art methods and future research directions.

2.1 Data Aggregation

Dataset aggregation has emerged as a crucial area of research in computer vision and machine learning, with significant implications for enhancing model performance and robustness. This section examines key works and methodologies related to dataset aggregation, focusing on their applications, benefits, and challenges, aiming to provide a comprehensive understanding of the current state of dataset aggregation techniques.

Wojtusiak and Baranova [23] address the challenges and methodologies of constructing predictive models from aggregated data,

particularly relevant in domains where individual data points are inaccessible due to privacy laws or proprietary restrictions. They propose a method of rule induction from aggregated data, applying it to the diagnostic of liver complications of metabolic syndrome. This approach showcases the effectiveness of working with aggregated data in healthcare applications, while complying with privacy regulations, highlighting the importance of aggregated data in fields with stringent data privacy concerns.

Prakash et al. [16] explore the impact of data aggregation in the context of policy learning for vision-based urban autonomous driving. Their study highlights the importance of aggregating diverse datasets to enhance the robustness and generalization of policy learning models. By showing that integrating data from multiple sources can significantly improve the performance of autonomous systems in complex urban environments, this work underscores the potential of dataset aggregation in developing more reliable and effective models for real-world applications.

Zhang et al. [24] address the problem of learning from aggregate observations, where individual labels are not available, but aggregated data is. They propose a framework for learning from such data, demonstrating its application in various tasks such as image classification and reinforcement learning. Their approach provides a way to leverage aggregate data effectively when fine-grained labels are unavailable, thus broadening the applicability of machine learning models in scenarios where obtaining detailed labeled data is challenging or impractical.

Sun et al. [18] introduce DEtect-rePLAce (DEPLA), a novel approach to feature aggregation in deep neural networks. DEPLA identifies and updates specific features across different layers, addressing inconsistencies and incompatibilities that arise from naive element-wise summation or concatenation. By integrating DEPLA into existing network architectures such as Residual Networks (ResNet), FishNet, and Feature Pyramid Network (FPN), significant improvements are achieved for image classification, object detection, and instance segmentation tasks. The DEPLA approach exemplifies how intelligent feature aggregation can enhance the performance of deep learning models.

The aggregation of datasets from diverse sources contributes to creating more robust models that generalize better across different conditions. For instance, the integration of data from various urban environments in autonomous driving systems leads to models that can handle a wider range of scenarios, thereby improving safety and reliability. In fields like healthcare, where data privacy and accessibility are major concerns, methods that utilize aggregated data provide a viable alternative for model training. By focusing on aggregated statistics rather than individual data points, researchers can still build effective predictive models while complying with privacy regulations.

However, the aggregation of datasets presents several challenges, particularly when managing the heterogeneity of data from different sources. Variations in data quality, formats, and distributions can hinder the aggregation process and affect the performance of the resulting models. As the volume of available data continues to grow, developing scalable aggregation methods becomes increasingly important.

Combining data aggregation with advanced feature detection and learning techniques holds promise for further advancements. Approaches like DEPLA, which intelligently merge features across layers, exemplify how integrative methods can enhance the effectiveness of dataset aggregation. Future research should focus on creating efficient algorithms that can handle large-scale data aggregation without compromising on accuracy or computational feasibility.

In conclusion, dataset aggregation plays a pivotal role in advancing the capabilities of machine learning models, particularly in domains with data privacy concerns or diverse data sources. The reviewed studies underscore the benefits of aggregation to enhance model robustness, addressing privacy issues, and improving generalization across different tasks. However, challenges such as data heterogeneity and scalability need to be addressed to fully harness the potential of dataset aggregation. By meeting these challenges and leveraging emerging technologies, it is possible to develop more accurate and robust models, contributing significantly to various fields of study and paving the way for more effective and adaptable machine learning solutions.

2.2 Fine-Grained Visual Classification

FGVC has seen substantial advancements, particularly in the context of marine species identification. FGVC is focused on distinguishing between visually similar categories, often requiring the identification of subtle differences within species. Numerous methodologies have been developed to tackle these challenges, ranging from object-part-based methods to attention-based approaches.

One significant advancement in FGVC is the High-temperaturE Refinement and Background Suppression (HERBS) network proposed by Chou et al. [5]. This network addresses the challenges of fine-grained classification by incorporating two key modules: the Background Suppression (BS) module and the high-temperature refinement module. The BS module enhances discriminative features by suppressing background noise, thereby improving the accuracy of identifying fine-grained categories. The high-temperature refinement module allows the model to learn diverse features at different scales, ensuring that both global and fine-grained details are captured effectively.

HERBS [5] has demonstrated state-of-the-art performance on benchmark datasets such as CUB-200-2011 [22] and NA-Birds [20], surpassing 93% accuracy. The network's ability to integrate with various backbone architectures, including both Convolutional Neural Network (CNN)s and transformer-based models, makes it a versatile tool for FGVC. The FGVC-PIM [6] enhances existing backbone networks by focusing on the most discriminative regions within an image, further boosting classification accuracy.

Other notable approaches in FGVC include the use of Vision Transformer (ViT) and Graph Convolutional Network (GCN)s. ViTs, such as those proposed by Dosovitskiy et al. [8], utilize self-attention mechanisms to capture detailed information about object parts, improving the ability to distinguish between visually similar categories. GCNs, on the other hand, model relationships between different parts of an image as a graph, enabling the capture of complex dependencies and interactions [3].

The use of multi-task learning frameworks has also been beneficial in FGVC, as demonstrated by Diao et al. [7]. These frameworks leverage additional supervisory signals from related tasks, such as

attribute prediction, to enhance the overall accuracy and robustness of the models. Ensemble methods, which combine the predictions of multiple models, have shown promise in reducing variance and bias, leading to more reliable and accurate classifications [19].

Despite these advancements, FGVC still faces challenges related to the variability in underwater environments and the long-tailed distribution of species in many datasets. Future research should focus on developing more robust and adaptable models that can handle the diverse and dynamic nature of marine environments. This includes improving data augmentation techniques to simulate underwater conditions more accurately and exploring unsupervised and semi-supervised learning methods to leverage large volumes of unlabeled data.

Veiga and Rodrigues [21] applied the FGVC-PIM to marine datasets and achieved several state-of-the-art results, which supports the use of this method in the current study. This approach has shown great potential in addressing the unique challenges posed by marine environments and species diversity, making it particularly suitable for our research objectives in marine biodiversity assessment and ecological monitoring.

In conclusion, the field of FGVC has made remarkable strides through the development of sophisticated models and innovative methodologies. The integration of dataset aggregation techniques with FGVC methods promises to further enhance the accuracy and robustness of models, particularly in challenging domains such as marine species identification. The proposed HERBS network, with its background suppression and high-temperature refinement modules, represents a significant advancement in this field, providing a promising solution for improving the performance of FGVC tasks.

3 Datasets

This section presents an overview of the diverse fish datasets utilized in our study of FGVC for marine species. Even though they vary significantly in size, species diversity, image quality, and capture conditions, they will be fused together to create a comprehensive, large-scale dataset for our experiments.

The individual datasets consist primarily of underwater images, many extracted from video recordings, providing a realistic representation of the challenges faced in marine fish identification scenarios. They range from smaller, tailored collections to large-scale datasets encompassing thousands of species.

Each dataset brings unique characteristics and challenges, such as class imbalance, varying image resolutions, and different levels of taxonomic granularity. Some datasets also provide additional metadata, including environmental parameters and detailed species descriptions, enhancing their value for ecological research beyond mere classification tasks.

By fusing these data, we aim to create a more robust and diverse foundation for training our FGVC models. This combined dataset will allow us to leverage the strengths of each set while potentially mitigating their individual limitations.

In the following, we detail each constituent dataset, discussing their composition, notable features, and any preprocessing steps undertaken for our experiments. Where relevant, we also highlight the class distribution characteristics, as these play a crucial role in the performance and generalization capabilities of FGVC models in underwater environments.

The paper by Holmberg et al. [11] presents a study on whale shark populations at Ningaloo Marine Park, Australia, using photo-identification and capture-mark-recapture methods from 1995 to 2008. The Whale Shark dataset contains 6,064 images representing 2,063 live captures, with 1,668 usable captures identifying 386 individual sharks. While this single-species dataset is not suitable for FGVC alone, it contributes valuable data when combined with other marine species' datasets. For our FGVC study, we used 7,693 images from this dataset, combining the training, test and validation images, adding one more species to the overall classification task when aggregated with other datasets.

The QUT Fish Dataset [2] contains 3,960 images of 468 fish species, captured under various conditions: *controlled*, *out-of-the-water* (uncontrolled), *in-situ*, *rubbish*, and *sketches*. The dataset used for experiments includes 464 classes, each with at least 3 images, after excluding the *rubbish* and *sketches* categories. This subset exhibits class imbalance, with the minority class having 3 images and the majority class having 26 images. The dataset provides both raw and cropped images, as well as bounding box annotations. The diverse capture conditions offer a range of challenges, from controlled environments with consistent lighting to unpredictable underwater settings.

The Croatian Fish Dataset [12] comprises 794 images representing 12 distinct fish species from Croatia's Adriatic Sea. Extracted from high-definition footage, this dataset is designed for FGVC in natural environments. It features bounding boxes and species labels for each fish. The dataset exhibits class imbalance, with image counts per species ranging from 17 to 111. Image sizes vary significantly, from 36×12 pixels to 503×231 pixels. No pre-processing was required for this dataset.

The FishCLEF dataset [13] consists of underwater video sequences from the Fish4Knowledge project. For this paper, both the training and test sets were utilized, in a total of 25,313 images distributed across 26 fish species, derived from 93 annotated videos (20 training + 73 test). The dataset is intentionally unbalanced; the most abundant class has 3,617 samples and the least represented, only 6, after removing one corrupted image. Some videos contain no fish to evaluate false positive rejection, although this data was irrelevant for the purpose of this paper. Temporal information was also not used to ensure identification was independent of tracking.

The Fish4Knowledge dataset [9] comprises 27,370 fish images of 23 species, captured by 9 underwater cameras at 3 sites over 1,000 days. The dataset exhibits significant class imbalance, with 16 to 12,112 images per class and the top 5 classes representing 91% of the data. Derived from 43,625 hours of video, the dataset includes pre-cropped images ready for training.

The WildFish [25] and WildFish++ [26] datasets provide large-scale, diverse benchmarks for fish recognition in unconstrained environments. WildFish contains 54,453 images of 1,000 fish species, while WildFish++ expands this to 103,025 images across 2,348 species. Both datasets present unique challenges for FGVC research.

WildFish introduces two specific tasks: paired text recognition for fine-grained classification of 22 highly similar species pairs, and open-set classification with 685 species for training and 1,000 for testing (including 315 unknown species). WildFish++ builds

upon these challenges, incorporating 3,817 fish descriptions totaling 213,858 words. Both datasets exhibit a slight long-tail distribution.

For our experiment, both datasets underwent preprocessing to remove corrupted images. The WildFish dataset was reduced from 54,459 to 54,453 images, while the WildFish++ dataset was filtered from 103,034 to 103,025 images.

WildFish++ outlines four key challenges: fine-grained recognition with comparison texts, multi-modal approaches, open-set classification, and cross-modal retrieval. It offers additional biological information and serves as the largest fish dataset to date, providing a foundational resource for both model training and methodological advancements in automatic fish classification.

The OzFish dataset [1], created for enhancing automated fish detection in videos, contains nearly 80,000 annotated crops of fish from over 3,000 videos. It covers 200 genera, 70 families, and more than 500 species. The dataset includes approximately 45,000 bounding box annotations across 1,800 frames. For experimental use, 425 classes (out of 594) containing more than 10 images each were considered. The number of images per class ranges from 10 to 6,095, exhibiting a long-tailed distribution.

The DeepFish dataset, introduced by Saleh et al. [17], is a comprehensive benchmark for underwater fish habitat analysis. It contains about 39,766 high-resolution images for classification, 3,200 for counting and localization, and 620 for segmentation, from 20 marine habitats in tropical Australia. For our research, we used the segmentation masks to extract bounding boxes for FGVC, as the original tasks were limited. From the available data we discarded the images without specimens, and the masks missing a scientific name, resulting in a subset of only 3 classes with 79, 4, and 3 images per class.

The SEAMAPD21 dataset [4] comprises 90,000 annotations of 130 species, collected using baited underwater video technology between 2018 and 2019. It exhibits a significant long-tailed distribution, with the minority class having 3 images and the majority class (*Lutjanus campechanus*) having 15,199 images. For experimental use, only classes with a minimum of 10 images were considered, resulting in 110 classes. The majority class represents over 19% of the dataset. Images were cropped using bounding box metadata and distributed by genus and species.

The FishNet dataset [14] is a comprehensive collection of 94,532 images across 17,357 aquatic species, organized into 8 taxonomic groups, 83 orders, 463 families, and 3,826 genera. It includes 22 features related to habitat, ecological role, and nutritional value. Veiga and Rodrigues [21] refined a species subset for FGVC from the FishNet dataset [14]. This subset includes 199 classes, considering only species with more than 30 images and ensuring a minimum of 165 images per class. The resulting experimental dataset contains 52,149 images. This refinement was done to create a more practical and balanced subset for classification tasks.

The OBSEA dataset [10] contains 33,805 images with 69,917 hand-identified fish specimens, collected off the coast of Barcelona, Spain. The data was gathered over two years, between 2013 and 2014, using an underwater video platform, capturing images every 30 minutes to reflect seasonal and diurnal variations. The dataset includes high-resolution visual data and concurrent oceanographic and meteorological measurements.

For experimental purposes, the images were cropped using bounding box metadata. Unknown species and 45 corrupted images were discarded. Only species with a minimum of 10 images were considered, resulting in 25 classes. The dataset exhibits a long-tailed distribution, with the minority class having 10 images and the majority class having 14,299 images. The two largest classes account for 57.7% of the dataset.

Our previous publication [21] can be consulted for a detailed description of the datasets' preprocessing.

In the following chapter, we present our suggested methodology for addressing the challenges of FGVC in underwater fish habitats.

4 Methodology

The Fine-Grained Visual Classification methodology is underpinned by the Swin Transformer architecture, a groundbreaking approach introduced by Liu et al. [15]. This sophisticated framework is structured into four distinct stages, each meticulously designed to process image patches at progressively complex levels of abstraction using an advanced Multi-Head Self-Attention (MSA) mechanism.

The process initiates with a high-resolution input image, specifically 384x384 pixels with 3 color channels. This image undergoes initial processing in the patch embedding block, where it is systematically divided into non-overlapping patches. Each of these patches is then treated as an individual token, setting the stage for deeper analysis. These tokens are subsequently fed through a series of Swin Transformer blocks, where the Window-based MSA (W-MSA) and Shifted Window-based MSA (SW-MSA) mechanisms work in tandem to capture local dependencies effectively. This is achieved by computing self-attention within localized windows, allowing the model to grasp fine-grained details that are elemental for accurate classification.

As we progress through the subsequent stages, the architecture incorporates patch merging layers, which play a pivotal role in aggregating features from neighboring patches, effectively increasing the feature dimensionality while simultaneously reducing spatial resolution. This clever design choice results in a hierarchical downsampling effect, where each of the second, third, and fourth stages doubles the number of channels while halving the spatial resolution. This approach enables the model to construct a rich, multi-scale representation of the input image.

The backbone of each Swin Transformer block is composed of several key components: a residual connection, a feed-forward network (comprising Multi-Layer Perceptron (MLP) layers), a MSA module, and layer normalization. The residual connections serve a critical function in ensuring stable training by mitigating the notorious vanishing gradient problem. Concurrently, the layer normalization standardizes inputs, significantly enhancing the network's convergence properties.

Building upon this robust foundation, Chou et al. [6] introduced the innovative FGVC-PIM module. This versatile enhancement is designed to seamlessly integrate with various backbone networks, including both CNNs and transformer-based architectures. The primary objective of this module is to pinpoint and accentuate the most discriminative regions within an image, thereby boosting classification accuracy.

In the context of fine-grained classification, the FGVC-PIM architecture first extracts feature maps from the input image. This is accomplished by leveraging the Swin Transformer in conjunction with a FPN, allowing for multi-scale feature extraction. These rich feature maps are then channeled through the FGVC-PIM, which consists of two main components: the Weakly Supervised Selector (WSS) and the Combiner.

The WSS employs a sophisticated fully connected layer to predict categories for each feature point within the extracted maps. This process effectively filters out less relevant information, retaining only the most salient features. Subsequently, the Combiner, powered by a GCN, efficiently integrates these carefully selected feature points. This integration is performed in a manner that preserves the integrity of the backbone model's output, ensuring that the additional processing enhances, rather than alters, the original predictions.

One of the key strengths of the FGVC-PIM lies in its implementation of a multi-loss function strategy. This approach involves a delicate balancing act, where the weights of various loss functions are meticulously adjusted to equalize their contributions. The result is a model that can focus intensely on the most relevant features while simultaneously protecting against overfitting. This precise tuning mechanism significantly amplifies the Swin Transformer's performance in FGVC tasks.

For a detailed description of the proposed methodology, readers are referred to our previous publication [21]. In the following section, we present our experimental methodology, encompassing dataset aggregation and post-processing techniques, followed by a comprehensive analysis of results and subsequent discussion.

5 Experiments, Results and Discussion

This section details the experimental methodology employed to assess the efficacy of dataset aggregation for FGVC in marine species identification. Our approach leverages the Swin Transformer architecture, enhanced by the FGVC-PIM, to optimize model performance. The experiments were designed to evaluate the Dataset Aggregation Hypothesis, which posits that combining diverse datasets enhances model robustness and accuracy, while significantly improves performance over baseline models.

5.1 Data Preparation

We aggregated datasets from multiple sources, each presenting unique characteristics and challenges. Initially, our dataset was comprised of 2,862 classes and 450,197 images. After rigorous preprocessing to remove corrupted and duplicated images, and invalid class names, the dataset was sifted into 2,734 classes and 392,152 images. To ensure scientific accuracy, we retained only classes with genus and species' scientific names. Classes with fewer than 10 images were excluded, resulting in a final dataset of 2,548 classes and 391,374 images.

The aggregated dataset incorporates the Croatian Fish Dataset [12], DeepFish [17], Fish4Knowledge [9] , FishCLEF-2015 [13], FishNet [14], OBSEA [10], OzFish [1], QUT Fish Dataset [2], SEAMAPD21 [4], Whale Shark [11], WildFish [25], and WildFish++ [26]. Notably, FishCLEF-2015 [13], DeepFish [17], and WhaleShark [11] are new additions compared to Veiga and Rodrigues [21]. Table 1 presents a comparative analysis of the original and contributed class counts from various fish-related datasets, illustrating the impact of data curation processes on class representation in the aggregated dataset.

Table 1: Analysis of datasets: Unique class contributions to the final curated dataset after removal of duplicates and image validation.

Dataset	Original Classes	Contributed Classes
Croatian [12]	12	10
DeepFish [17]	3	2
Fish4Knowledge [9]	23	22
FishCLEF-2015 [13]	26	23
FishNet [14]	199	198
OBSEA [10]	25	22
OzFish [1]	425	404
QUT [2]	464	333
SEAMAPD21 [4]	110	82
Whale Shark [11]	1	1
WildFish [25]	1,000	987
WildFish++ [26]	2,348	2310

5.2 Model Training and Evaluation

Our experimental protocol employed a uniform split strategy (generator seed: 42), partitioning data into 10% testing, 20% validation, and 70% training sets. The Swin Transformer, with 384×384 pixel input, formed the FGVC foundation. Training utilized 12 worker threads, a batch size of 16, and an Stochastic Gradient Descent (SGD) optimizer (weight decay and maximum learning rate: 0.0005). An 800-batch warmup preceded 10-50 epochs of training, varying by dataset size, with mixed-precision training for efficiency.

To address multi-scale features, we incorporated an FPN (feature size: 1536). Selection modules targeted discriminative regions, selecting 2048, 512, 128, and 32 feature points from respective layers, subsequently combined via the Combiner module. Loss function components were weighted: base loss (0.5), selection loss (0.0), drop loss (5.0), and combiner loss (1.0). The model parameters were updated after each batch, with assessments every 5 epochs. This approach, leveraging the Swin Transformer's multi-scale and hierarchical properties alongside FGVC-PIM enhancements, aimed to significantly boost FGVC performance.

5.3 Results and Discussion

The experimental results demonstrate that dataset aggregation significantly enhances FGVC model performance. Our aggregated data model achieved a remarkable 94.75% overall accuracy, underscoring the efficacy of combining diverse datasets (see Table 2). Notable improvements were observed in datasets characterized by high species diversity and challenging capture conditions, exemplified by the substantial accuracy increase (+6.25%) for the WildFish++ [26] dataset.

Table 2: Comparison of individual fish datasets and their performance metrics, including the number of classes, number of valid images, baseline accuracy, and accuracy when used in an aggregated data model. The final column shows the difference between baseline and the aggregated data model, and the bottom line shows the totals for the aggregated dataset combining all individual datasets.

Dataset	Number of Classes	Number of Valid Images	Baseline Accuracy	**Aggreg. Data Model Accur.**	Difference
Croatian [12]	12	794	98.73% [21]	95.95%	-2.78%
DeepFish [17]	3	86	———	100.00%	———
Fish4Knowledge [9]	23	27,370	100.00% [21]	99.93%	-0.07%
FishCLEF-2015 [13]	26	25,313	99.68%	99.56%	-0.12%
FishNet [14]	199	52,149	96.05% [21]	93.42%	-2.63%
OBSEA [10]	25	36,710	98.28% [21]	97.55%	-0.73%
OzFish [1]	425	79,503	97.90% [21]	94.95%	-2.95%
QUT [2]	464	3,924	93.88% [21]	93.52%	-0.36%
SEAMAPD21 [4]	110	78,546	97.67% [21]	93.81%	-3.86%
Whale Shark [11]	1	7,693	———	100.00%	———
WildFish [25]	1,000	54,453	96.68% [21]	92.83%	-3.85%
WildFish++ [26]	2,348	103,025	85.93% [21]	92.18%	+6.25%
Aggregated Datasets	**2,548**	**391,374**	———	**94.75%**	———

Table 2 provides a comprehensive comparison of these datasets, comparing the baseline accuracy and accuracy within the aggregated model, based on the 10% test set. The Table shows in the first column fish datasets and their performance criteria, such as the quantity of valid images, the number of classes, the accuracy of the baseline, and the accuracy of the data model when combined. The difference between the aggregated data model and the baseline is displayed in the final column, and the totals for the aggregated dataset – which combines all the different datasets – are displayed on the bottom line.

Intriguingly, accuracy differences between models trained on smaller class subsets versus our proposed aggregated data model were generally minimal, although not positive. In some instances, particularly with the largest dataset, data aggregation even yielded accuracy improvements over the baseline model. This highlights the profound impact of larger, more diverse datasets on the model's generalization capabilities and classification accuracy across a wide range of classes.

Dataset aggregation effectively mitigates individual dataset limitations, such as class imbalance and variability in image quality and capture conditions. The enhanced diversity and volume of data enable the model to learn more robust and generalizable features. This proves particularly advantageous for large-scale models like the one employed in this study, which was trained on 2,548 classes. The increased training data volume allows the model to better capture subtle distinctions between closely related classes, ultimately leading to improved classification performance.

Our experiments substantiate Dataset Aggregation Hypothesis. By aggregating datasets, we mitigate individual dataset limitations, providing a more comprehensive representation of marine species. This approach effectively addresses class imbalance and image quality variability issues, resulting in more robust and generalizable models.

The integration of FGVC-PIM with the Swin Transformer significantly enhances model capabilities. FGVC-PIM's ability to focus on discriminative image regions improves fine-grained classification, particularly in complex underwater environments. Despite these advancements, challenges persist. Data heterogeneity from diverse sources complicates the aggregation process, necessitating sophisticated preprocessing and normalization techniques. Furthermore, the scalability of these methods to larger datasets and more diverse marine environments requires further investigation.

6 Conclusion

This study demonstrates the efficacy of dataset aggregation and advanced model integration techniques in enhancing FGVC for marine species. By leveraging diverse datasets and innovative methodologies like FGVC-PIM, we achieved substantial improvements in model performance. These findings underscore the critical role of data diversity and advanced feature extraction in developing robust, accurate classification models for complex real-world applications.

Our research contributes significantly to sustainability and environmental awareness by providing more precise tools for marine biodiversity monitoring, thus combating information exclusion. Improved marine species classification enables better-informed conservation decisions, fostering a deeper understanding of marine ecosystems and their health. Access to reliable, detailed marine life information is crucial for developing sustainable practices and policies that protect our oceans for future generations. To further support this goal and promote open science, we plan to make our code for dataset preparation and model training publicly available in the near future. This step will enable other researchers to replicate our methodology, utilize our aggregated dataset, and potentially extend our work, thereby amplifying the impact of our research on marine conservation efforts and environmental sustainability.

Moreover, our approach aligns with HMC objectives. By enhancing machines' ability to accurately classify and monitor marine

biodiversity, we facilitate improved human-machine cooperation. This enhanced interaction ensures machines provide correct information and functionality, tailored to users' individual characteristics, needs, tasks, and contexts. Such collaboration is essential for improving accessibility, promoting sustainability, and combating information exclusion.

Our work not only advances FGVC but also supports broader efforts to create more interpersonal relationships between the digital world and humans, ultimately contributing to a more sustainable and informed society. It also exemplifies how advanced machine learning and data aggregation can make complex environmental data more accessible and interpretable, providing an approach for reducing information barriers in other research areas. This aligns with creating inclusive technologies that empower scientists, biologists, and people without particular technical skills to aid in their research or monitoring efforts.

Future research should prioritize developing adaptive and scalable aggregation algorithms, enhancing data augmentation techniques, and leveraging unsupervised and semi-supervised learning methods to effectively utilize unlabeled data. Addressing these challenges will facilitate the development of more accurate and reliable FGVC models, significantly advancing marine biodiversity assessment and ecological monitoring efforts. It is also important to stress that future work also should explore adapting the presented techniques to other domains, further promoting equitable access to information across various fields.

Acknowledgments

This work is supported by the Portuguese Foundation for Science and Technology (FCT), by NOVA LINCS ref. UIDB/04516/2020 https://doi.org/10.54499/UIDB/04516/2020 and ref. UIDP/04516/2020 https://doi.org/10.54499/UIDP/04516/2020 with the financial support of FCT.IP, and by the FCT - PhD grant 2022.11602.BD. A sincere thank you to Dra. Elda Veiga for her diligent proofreading of the article, and constructive feedback that greatly improved the overall quality of this manuscript.

References

[1] Australian Institute of Marine Science (AIMS), University of Western Australia (UWA), and Curtin University. 2019. OzFish Dataset - Machine learning dataset for Baited Remote Underwater Video Stations. https://doi.org/10.25845/5e28f062c5097

[2] Kaneswaran Anantharajah, Zong Yuan Ge, Chris McCool, Simon Denman, Clinton Fookes, Peter Corke, Dian Tjondronegoro, and Sridha Sridharan. 2014. Local inter-session variability modelling for object classification. *2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014* (2014), 309–316. https://doi.org/10.1109/WACV.2014.6836084

[3] Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. 2022. SR-GNN: Spatial Relation-Aware Graph Neural Network for Fine-Grained Image Categorization. *IEEE Transactions on Image Processing* 31, 0 (2022), 6017–6031. https://doi.org/10.1109/TIP.2022.3205215 arXiv:2209.02109

[4] O. Boulais, S. Y. Alaba, J. E. Ball, M. Campbell, A. T. Iftekhar, R. Moorhead, J. Prior, F. Wallace, H. Yu, and A. Zheng. 2021. SEAMAPD21: a large-scale reef fish dataset for fine-grained categorization. *FGVC8: The Eight Workshop on Fine-Grained Visual Categorization - CVPR 2021* 25 (2021). https://github.com/SEFSC/

[5] Po-Yung Chou, Yu-Yung Kao, and Cheng-Hung Lin. 2023. Fine-grained Visual Classification with High-temperature Refinement and Background Suppression. March (2023), 1–9. arXiv:2303.06442 http://arxiv.org/abs/2303.06442

[6] Po-Yung Chou, Cheng-Hung Lin, and Wen-Chung Kao. 2022. A Novel Plug-in Module for Fine-Grained Visual Classification. *arXiv preprint arXiv:2202.03822* (feb 2022). arXiv:2202.03822 http://arxiv.org/abs/2202.03822

[7] Qishuai Diao, Yi Jiang, Bin Wen, Jia Sun, and Zehuan Yuan. 2022. MetaFormer: A Unified Meta Framework for Fine-Grained Recognition. (2022). arXiv:2203.02751 http://arxiv.org/abs/2203.02751

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Robert B Fisher, Daniela Giordano, and Fang-pang Lin Editors. 2016. *Fish4Knowledge : Collecting and Analyzing Massive Coral Reef Fish Video Data*. Vol. 104. Springer. 321 pages. http://www.scopus.com/inward/record.url?eid=2-s2.0-84961753059{&}partnerID=tZOtx3y1

[10] Marco Francescangeli, Simone Marini, Enoc Martínez, Joaquín Del Río, Daniel M. Toma, Marc Nogueras, and Jacopo Aguzzi. 2023. Image dataset for benchmarking automated fish detection and classification algorithms. *Scientific Data* 10, 1 (dec 2023). https://doi.org/10.1038/s41597-022-01906-1

[11] Jason Holmberg, Bradley Norman, and Zaven Arzoumanian. 2009. Estimating population size, structure, and residency time for whale sharks Rhincodon typus through collaborative photo-identification. *Endangered Species Research* 7, 1 (2009), 39–53. https://doi.org/10.3354/esr00186

[12] Jonas Jäger, Marcel Simon, Joachim Denzler, Viviane Wolff, Klaus Fricke-Neuderth, and Claudia Kruschel. 2015. Croatian fish dataset: Fine-grained classification of fish species in their natural habitat. *Swansea: Bmvc* 2 (dec 2015), 6.1–6.7. https://doi.org/10.5244/c.29.mvab.6

[13] Alexis Joly, Concetto Spampinato, Pierre Bonnet, Willem-pier Vellinga, Robert Planqu, Andreas Rauber, and Simone Palazzo. 2015. LifeCLEF 2015 : Multimedia Life Species Identification Challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 462–483.

[14] Faizan Farooq Khan, Xiang Li, Andrew J Temple, and Mohamed Elhoseiny. 2023. FishNet: A Large-scale Dataset and Benchmark for Fish Recognition, Detection, and Functional Trait Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20496–20506.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986 arXiv:2103.14030

[16] Aditya Prakash, Aseem Behl, Eshed Ohn-Bar, Kashyap Chitta, and Andreas Geiger. 2020. Exploring data aggregation in policy learning for vision-based Urban autonomous driving. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2020), 11760–11770. https://doi.org/10.1109/CVPR42600.2020.01178

[17] Alzayat Saleh, Issam H. Laradji, Dmitry A. Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. 2020. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports* 10, 1 (aug 2020), 1–10. https://doi.org/10.1038/s41598-020-71639-x arXiv:2008.12603

[18] Shuyang Sun, Xiaoyu Yue, Xiaojuan Qi, Wanli Ouyang, Victor Prisacariu, and Philip Torr. 2021. Aggregation with Feature Detection. *Proceedings of the IEEE International Conference on Computer Vision* (2021), 507–516. https://doi.org/10.1109/ICCV48922.2021.00057

[19] Mingxing Tan and Quoc V Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2019). arXiv:arXiv:1905.11946v5

[20] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 595–604.

[21] Ricardo J.M. Veiga and João M.F. Rodrigues. 2024. Fine-Grained Fish Classification from small to large datasets with Vision Transformers.

[22] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[23] Janusz Wojtusiak and Ancha Baranova. 2011. Model learning from published aggregated data. *Studies in Computational Intelligence* 375 (2011), 369–384. https://doi.org/10.1007/978-3-642-22913-8_17

[24] Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. 2020. Learning from aggregate observations. *Advances in Neural Information Processing Systems* 2020-December, NeurIPS (2020). arXiv:2004.06316

[25] Peiqin Zhuang, Yali Wang, and Yu Qiao. 2018. Wildfish: A large benchmark for fish recognition in the wild. In *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, Vol. 2. Association for Computing Machinery, Inc, 1301–1309. https://doi.org/10.1145/3240508.3240616

[26] Peiqin Zhuang, Yali Wang, and Yu Qiao. 2021. Wildfish++: A Comprehensive Fish Benchmark for Multimedia Research. *IEEE Transactions on Multimedia* 23 (2021), 3603–3617. https://doi.org/10.1109/TMM.2020.3028482