IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 August 2024, accepted 11 August 2024, date of publication 14 August 2024, date of current version 26 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3443654

RESEARCH ARTICLE

Fine-Grained Fish Classification From Small to Large Datasets With Vision Transformers

RICARDO J. M. VEIGA[1,2] **AND JOÃO M. F. RODRIGUES**[2]
[1]Faculty of Science and Technology (FCT), University of Algarve, 8005-139 Faro, Portugal
[2]NOVA LINCS and Institute of Engineering (ISE), Universidade do Algarve, 8005-139 Faro, Portugal

Corresponding author: Ricardo J. M. Veiga (rjveiga@ualg.pt)

This work was supported in part by the Portuguese Foundation for Science and Technology (FCT) by NOVA LINCS under Grant UIDB/04516/2020 and Grant UIDP/04516/2020, and in part by FCT-Ph.D. under Grant 2022.11602.BD.

ABSTRACT Fish species Fine-Grained Visual Classification (FGVC) is important for ecological research, environmental management, and biodiversity monitoring, as accurate fish species identification is crucial for assessing the health of marine ecosystems, monitoring changes in biodiversity, and converting conservation plans into action. Although Convolutional Neural Network (CNN)s have been the conventional approach for FGVC, their effectiveness in differentiating visually similar species is not always satisfactory. The advent of Vision Transformer (ViT)s, in particular the Shifted window (Swin) Transformer, has demonstrated potential in addressing these issues by using sophisticated self-attention and feature extraction techniques. This paper proposes a method of combining the FGVC Plug-in Module (FGVC-PIM) and the Swin Transformer. The FGVC-PIM improves classification by concentrating on the most discriminative image regions, while the Swin Transformer acts as the framework and provides strong hierarchical feature extraction. The performance of the method was assessed on 14 different datasets, which included 19 distinct subsets with varying environmental conditions and image quality. With the proposed method it was achieved state-of-the-art results in 13 of these subsets, exhibiting better accuracy and robustness than previous methods, in 2 subsets (not yet explored by other authors) new baseline results are presented, and in the remaining 4 it was achieved results always above 83%.

INDEX TERMS Computer vision, convolutional neural networks, fine-grained visual classification, marine biodiversity monitoring, swin transformer.

I. INTRODUCTION

The accurate identification of fish species is an integral task in ecological research and environmental management, helping to assess the health of marine ecosystems, track biodiversity changes, and implement effective conservation strategies, for which the Fine-Grained Visual Classification (FGVC) of fish species can be a vital tool. FGVC methods were traditionally based on Convolutional Neural Network (CNN)s, which, despite succeeding in various image classification tasks, can struggle when differentiating between the subtle distinctions that set apart visually similar species, which is a particularly challenging task in the different aquatic environment conditions.

Advancements in Vision Transformer (ViT)s in recent times have demonstrated a significant potential to surmount these constraints. ViTs are especially well-suited for fine-grained classification tasks, as they use self-attention mechanisms to capture complex details and long-range dependencies within images. Out of all these models, the Shifted window (Swin) Transformer stands out due to its ability to extract features in a hierarchical manner and handle multiscale data efficiently, both of which are essential for FGVC high accuracy.

This paper proposes integrating the Swin Transformer [1] with the FGVC Plug-in Module (FGVC-PIM) [2] for a new domain of application for fish species detection. The present

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

method is based on the Swin Transformer, which offers strong hierarchical feature extraction: by concentrating on the most discriminative areas of the pictures, the FGVC-PIM further improves this procedure and helps the model to better classify various fish species. Thus, combining the advantages of both architectures, this integration improves performance on FGVC tasks.

The method was tested using 14 different datasets with 19 distinct subsets, all with different characteristics. The environmental conditions covered by these datasets are diverse and include images taken in controlled laboratory settings, images taken on-board vessels, and underwater images taken in natural habitats. Given real-world scenarios where fine-grained classification is necessary, this diversity provides a thorough evaluation of our method's robustness and adaptability. The experimental results obtained shows that the method in the majority of subsets outperformed previous approaches (15 times in 19). These findings emphasize the Transformers' potential for FGVC tasks to be a scalable and efficient solution for applications in the field of marine biology and environmental management. In addition, it demonstrates the method's broad applicability and effectiveness, since it adapts to various datasets and particular conditions.

The main contributions of this paper are: (i) the extensive application of the FGVC-PIM, combined with Swin Transformer as backbone on the task of FGVC of fish species; (ii) the comprehensive evaluation across 14 different datasets, resulting in 19 different experiments while maintaining the same hyperparameters, demonstrating the proposed method's versatility and robustness, covering a wide range of environmental conditions and variable image quality; (iii) the achievement of 13 state-of-the-art results, outperforming multiple previous methods, proving the ViT's potential for FGVC tasks; (iv) additionally, the inclusion of 2 new baselines, and one novel subset (counting the above, 15 state-of-the-art results in 19 datasets/subsets); (v) and a detailed performance analysis, providing multiple analysis on the datasets' characteristics and distribution, reporting a comprehensive evaluation of the proposed method's performance accompanied by additional metrics.

The remainder of this paper is structured as follows: Section II reviews related datasets and related work, establishing the context for the research and highlighting the significance of the addressed challenges. Section III details the methodology, including the organization, acquisition, and preprocessing of the datasets, and explains the proposed method integrating the Swin Transformer with the FGVC-PIM. Section IV presents the experiments and results, describing the configuration, hyperparameters, data split strategies, and a comprehensive comparison of the proposed method against state-of-the-art techniques. In this section it is also discussed the results, highlighting the method's versatility and robustness across different conditions and identifying areas for improvement. Finally, Section V concludes the paper, summarizing the findings and suggesting directions for future research to enhance model performance and broaden its application scope.

II. RELATED WORK

This section is divided into three subsections. It starts with a presentation and brief description of the most common datasets used in state-of-the-art publications, in a chronological order of their publication dates. Subsequently, the methods applied to each of them are categorized under the specific dataset, also listed chronologically by publication date. The section is closed with a discussion that synthesizes and analyzes the datasets and methods presented. Table 1 summarizes the state-of-the-art, and Figure 1 provides some example images from each dataset.

A. DATASETS

The QUT fish dataset [12] includes 3,960 real-world images of fish from 468 species, captured in different conditions defined as *controlled*, *out-of-the-water*, and *in-situ*. Out-of-water photos show fish against a variety of backgrounds and lighting conditions, while controlled shots provide crisp visuals with steady backgrounds and lighting. *In-situ* catches are made underwater with unpredictable environmental conditions. Additional bounding box annotation is provided to further differentiate the subject from the background.

The Croatian Fish Dataset [5] is made up of 794 photos that represent 12 distinct fish species. These photos were taken from high-definition footage that was shot in Croatia's Adriatic Sea. The dataset is dedicated to fine-grained natural environment visual classification. Bounding boxes and species names are labeled on fish in the dataset. The variations in the frequency of a species' appearances in the films are reflected in the number of photos for each species. For instance, *Sarpa salpa* only occurs 17 times, whereas *Diplodus vulgaris*, the most commonly encountered species, is represented in 110 photos.

The Fish4Knowledge dataset [8] is a sizable collection of data gathered over a period of 1,000 days using 12 hours of daily observation and 9 underwater cameras at 3 different sites. Twenty-three different fish species were included in this dataset. Throughout the trial, 1.44 billion individual fish instances were discovered, each large enough to facilitate identification efforts (at least 50×50 pixels). Afterwards, a total of 145 million monitored fish were created by connecting the detected fish across different video frames. The final dataset included 261,751 ten-minute video clips, or 43,625 hours of film, after accounting for many factors including false detection rates, duplicated videos with different spatial resolutions, and removal of non-fish objects. The 27.4 million fish trajectories in this final batch are usually the same fish that have been traced over a number of video frames.

The Fish-Pak dataset [7] is a collection of 915 photos of six distinct fish species from Pakistan, designed to simplify visual feature-based classification. With a transparent background and high quality images, and a resolution of

FIGURE 1. Examples of images from each dataset, from top to bottom, left to right: A Large-Scale Dataset for Fish Segmentation and Classification [3], BDIndigenousFish2019 [4], Croatian Fish Dataset [5], Fish-Gres [6], Fish-Pak [7], Fish4Knowledge [8], FishNet [9], FishNet Open Images Database [10], OBSEA [10], OzFish [11], QUT Fish Dataset [12], SEAMAPD21 [13], WildFish [14], and WildFish++ [15].

5184 × 3456 pixels, the photos highlight the three main elements of the subject: the body, head, and scale.

The WildFish dataset [14] is a large-scale, diverse benchmark collection for fish recognition in the wild, containing 54,459 unconstrained images of 1,000 fish species. Along with the dataset, two unique tasks are also included: paired text recognition for fine-grained recognition, which consists of 22 highly similar pairs of fish species, containing additional textual description between the pairs of species, and open-set fish classification, consisting of 685 fish species for training and the complete 1,000 species for testing, where the remaining 315 species should be defined as unknown.

The BDIndigenousFish2019 dataset [4] contains 2,610 photos of Bangladeshi indigenous fish species, which are divided into 8 categories. These specimens were photographed at 3968 × 2796 pixels with a mobile phone camera against a non-uniform white background.

The OzFish dataset [11] was created as part of the Australian Research Data Commons Data Discoveries program, designed to enhance machine learning research for automated fish detection in videos. It contains almost 80,000 annotated bounding boxes of fish from more than 3,000 recordings, spanning 200 genera and 70 families and more than 500 species. The dataset also includes nearly 45,000 bounding box annotations for machine learning models, spanning 1,800 frames.

A Large-Scale Dataset for Fish Segmentation and Classification [3] is a comprehensive compilation of photos from a supermarket fish counter that is intended to be used for fish segmentation and classification. There are a total of 1,000 augmented images across nine classes, with 50 unique images for seven classes and 30 for the other two. To replicate real-world settings, fresh fish were photographed at several angles and locations against a noisy, blue background. The photographs were resized to 590 × 445 pixels, almost maintaining its aspect ratio. After that, these photos underwent augmentation, which included rotation at non-repeated random angles and reflections, to provide a large dataset that makes machine learning models useful in real-world scenarios.

The Fish-Gres dataset [6] is composed of images from eight different fish species, with each species contributing between 240 and 577 images, sourced from traditional markets in Gresik, East Java, Indonesia. The assortment and quantity of images in the collection are indicative of the species' availability in these markets. The images were captured against diverse backgrounds to emulate the typical conditions of these marketplaces, aiming to accurately represent the range of fish species encountered there. To facilitate uniformity and ease of analysis, images were downscaled from an original resolution of 4160 × 3120 pixels to 390 × 520 pixels. This dataset serves as a valuable resource for academic research involving the identification and study of fish species from this region.

The FishNet Open Images Database [16] consists of 406,463 bounding boxes in 86,029 images of 34 different fish species, obtained from 73 different electronic monitoring cameras. This large-scale dataset of fish images is intended to be used for the development of computer vision algorithms in fisheries, for fish detection and fine-grained classification onboard commercial fishing vessels. The images were captured during real-world fishing trips, providing a realistic context for the development of computer vision algorithms in fisheries.

The SEAMAPD21 dataset [13] contains 90,000 annotations describing 130 species and was collected between 2018 and 2019, using baited underwater video technology. However, the bulk of the species are not as often annotated as others, with *Lutjanus campechanus* being one of the most highly represented species. The species' varying prevalence is highlighted by this bias, which presents opportunities and challenges for in-depth picture analysis aimed at improving fishery monitoring and stock assessment.

The WildFish++ dataset [15] is an enhanced dataset which extends upon the previous WildFish dataset [14], consisting in 2,348 fish species, distributed across 103,034 images in the

wild, and accompanied by 3,817 fish descriptions, totaling 213,858 words. The authors outline four key challenges: fine-grained recognition with comparison texts, exploring the multi-modal approaches, the open-set classification, continuing the work presented on the previous publication, the cross-modal retrieval, offering additional biological fish information, and the automatic fish classification, providing the largest fish dataset to date, serving as a foundational tool for both model training and methodological advancements.

The FishNet dataset [9] serves as a comprehensive baseline for the detection, identification, and functional trait assessment of large-scale aquatic species, comprising 94,532 images across 17,357 aquatic species. This collection is meticulously organized using an extensive aquatic biological taxonomy, featuring 8 taxonomic groups, 83 orders, 463 families, and 3,826 genera. It includes 22 features divided into habitat, ecological rule, and nutritional value to aid in the study of ecological functions and interspecies relationships. The dataset is also divided into three main categories: fish classification, fish detection, and functional trait identification.

The OBSEA dataset [10] encompasses 33,805 images with a total of 69,917 human-identified fish specimens, collected using the OBSEA underwater video platform off the coast of Barcelona, Spain. Situated 20 meters below the surface and 4km off the coast of Vilanova i la Geltrú, in a fishing protected area, this extensive collection was acquired over a two-year period (2013-2014), with recordings taken every 30 minutes to capture a wide array of seasonal and diurnal variations. The dataset not only provides high-resolution visual data but is also enriched with concurrent oceanographic and meteorological measurements, including salinity, water temperature, and solar irradiance, offering a multidimensional snapshot of the marine environment at each moment of capture.

Figure 1 shows some examples of images from each dataset. The next section presents the methods used for each specific dataset by chronologically publication date.

B. METHODS

Current research using the QUT fish dataset [12] has shown a number of novel approaches to fish species categorization, showcasing notable developments in the use of deep learning technology. In order to improve the categorization of fine-grained fish images, Qiu et al. [17] combined refined Squeeze-and-Excitation (SE) blocks and Bilinear CNN (B-CNN) with enhanced data augmentation through super-resolution reconstruction. Hridayami et al. [18] utilized a VGG16 model pre-trained on ImageNet, incorporating transfer learning with diverse preprocessing methods, like RGB color space and Canny filters. Adiwinata et al. [19] employed the Faster Region-based CNN (R-CNN) Inception-v2 architecture without the need for preprocessing images. Islam et al. [20] developed a content-based fish recognition system that fuses Local Binary Pattern (LBP), Scale Invariant Feature Transform (SIFT), and Speeded-Up Robust Features (SURF) as local features and Color Coherence Vector (CCV) as global features. Then, various machine learning models are used for the classification.

Mathur and Goel [21] applied transfer learning to a Residual Networks (ResNet)50 network without data augmentation. Guo et al. [22] presented a cross-domain approach to transfer learning where the Deep CNN (DCNN) was fine-tuned on ImageNet dataset, then on the Flowers102 dataset, and finally on the QUT fish dataset [12]. Zhang et al. [23] introduced a novel deep adversarial learning framework, named AdvFish, focusing the training on adversarially perturbed images, improving the results against noisy images. Deka et al. [24] fine-tuned a ResNet50 model and applied a Support Vector Machine (SVM) classifier. Ahmad et al. [25] used Efficient-Net, pre-trained on ImageNet, and fine-tuned with Global Average Pooling (GAP) to improve localization accuracy and Focal Loss (FL) to alleviate class imbalance.

To improve underwater fish recognition on the Croatian Fish Dataset [5], Pang et al. [26] utilized a teacher-student model, consisting of Feature Similarity Alignment (FSA) and Kullback-Leibler (KL) Divergence (KLD) to distill interference in underwater fish images, improving species recognition efficiency by reducing light absorption and scattering challenges. In order to address the problems of noise, light attenuation, and dataset imbalance, Sudhakara et al. [27] used enhanced CNN models with Underwater Image Enhanced GAN (UIEGAN), which combines a CycleGenerative Adversarial Network (GAN) with a Deep Convolutional GAN (DCGAN), and SmallerVGG.

A refined SE and a hybrid CNN-SVM framework were proposed by Veluswami and Panneerselvam [28] to improve the automatic fish species classification on both large and small-scale datasets. By combining a quantized ResNet18 model with Variational Quantum Algorithms (VQAs), Chen [29] presented a novel framework for image classification that requires fewer parameters and better classification performance than classical models. This is achieved by removing the need for global pooling in order to capture more fine-grained details and discriminative features. Fish classification accuracy was improved across multiple datasets without the need for additional parameters thanks to the introduction of the AttentionConvMixer neural network by Viet et al. [30], which used Priority Channel Attention (PChA) and Priority Spatial Attention (PSA).

For the Fish4Knowledge dataset [8], Sun et al. [31] proposed an RGB-AlexNet-SVM model as part of a knowledge transfer framework for underwater object recognition, effectively addressing low-light and high-noise challenges by extracting discriminative features from low-contrast images and utilizing a weighted probabilities decision mechanism for improved object identification in video frames.

On this dataset (Fish4Knowledge), Zhang et al. [23] also used the AdvFish framework on this dataset. By presenting two EfficientNet models for fine-grained fish classification – one with Efficient Channel Attention Module (ECA) and another with Coordinate Attention (CA) – Gong et al. [32]

addressed problems with conventional fishing supervision and permitted the reallocation of human resources to management and enforcement activities. Yang et al. [33] presented an enhanced Flow Direction Algorithm (FDA) and search agent strategy to optimize an Extreme Learning Machine (ELM) underwater target classification Fuzzy-C-Means (FCM) algorithm.

Jiang et al. [34] presented Efficient Vision Transformer with Token-Selective and Merging Strategies (TSMVT), which incorporates token-selective and merging module to reduce computational demands, and a Multi-Attention Weighting Token-Selective (MAWTS) module that dynamically adjusts attention weights to focus on key features, also applying it on this dataset (Fish4Knowledge). Qu et al. [35] introduced ConvFishNet, which incorporates large convolutional kernels and depth-wise separable convolutions to reduce the model parameters, along with PixelShuffle upsampling to enhance feature information, which rearranges elements of low-resolution, multi-channel feature map into a high-resolution feature map.

With class-balanced focal loss and SE-ResNet152, Xu et al. [36] presented a method for fish species identification that addresses the issues of small sample sizes and category imbalance on the Fish-Pak dataset [7], across the three fish image views (body, head, and scale). To overcome the difficulties of structural deformation and orientation misalignments in out-of-water fish images, N.S. et al. [37] presented a multisegmented fish classification method utilizing an AlexNet model with a naive Bayesian fusion layer. The refined ResNet50 model with a SVM classifier was further applied by Deka et al. [24] to this dataset. To address difficulties with manual identification, Deka et al. [38] used ResNet50 and AlexNet. Label smoothing and Gradient-weighted Class Activation Mapping (Grad-CAM) were used by Gong et al. [39] to improve feature extraction and model optimization in Fish-TViT, conjoining the fish classification technique based on transfer learning and ViT.

For weakly supervised fine-grained recognition, Yu et al. [40] introduced the Spatial-Channel Aware Attention Filters (SCAF) method, which, when combined with Multi-Channel Multi-Level (MCML) and non-randomly (Non-Rdn), improves discriminative regions in both spatial and channel dimensions on the WildFish dataset [14]. The AdvFish framework was also applied to this dataset by Zhang et al. [23]. A Two-Tier Knowledge Distillation (T-KD) approach with interlayer mapping similarity-preserving (IMSP) and layer tail response (LTR) was presented by Li et al. [41], in order to improve accuracy and decrease parameters, along with a novel Fish37 dataset. Jiang et al. [34] also applied TSMVT to this dataset. Manikandan and Santhanam [42] introduced Amended Dual Attention oN Self-locale and External (ADANSE) mechanism-based Vision Transformer, that combines self-locale and external attention mechanisms. Using block-tokenization and a novel dual attention approach, ADANSE extracts deep features and considers relationships among image blocks, with their outputs from the attention mechanism being fed into a Multi-Layer Perceptron (MLP).

Together with the BDIndigenousFish2019 dataset [4], Islam et al. [4] presented a Hybrid Local Binary Pattern (HLBP) for classifying indigenous fish species of Bangladesh using different SVM kernels for classification. Dey et al. [43] suggested a CNN-based automatic classification system. The outcomes of testing VGG16, Inception V3, MobileNet, and a specially designed 5-layer CNN called FishNet were similar to those of the pre-trained models. Adam and Root Mean Square Propagation (Rmsprop) optimizers outperformed the other five gradient descent-based optimizers in Smadi et al. [44], which also presented an efficient CNN-based fish classification technique. The Attention ConvMixer neural network with PChA and PSA was also applied to this dataset by Viet et al. [30].

On the Large-Scale Dataset for Fish Segmentation and Classification [3], Mampitiya et al. [45] used dimension reduction techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA); nevertheless, the authors achieved better results with Random Forest. Using the Chaotic Oppositional based Whale Optimization Algorithm (COWOA) and EfficientNet B0, Aziz et al. [46] presented a fish classification technique that outperformed models such as CNN, VGG-19, ResNet150V2, DenseNet, Inception V3, and Xception. Li et al. [41] additionally employed the T-KD technique with IMSP and LTR, on this dataset. In order to help people with seafood allergies and clinical immunology specialists, Reddy et al. [47] proposed a system that uses custom CNN algorithms for detecting and classifying fish that cause allergies. MobileNetV2 was utilized in conjunction with transfer learning strategies by Kranthi Kumar et al. [48].

Using the Fish-Gres dataset [6], Prasetyo et al. [49] proposed the MLR-VGGNet architecture, which combines low-level and high-level features using Multi-Level Residual (MLR) and Depthwise Separable Convolutions (DSC). According to Azhar et al. [50], adding Histogram Equalization (HE) and Contrast Limited Adaptive HE (CLAHE) to fish image datasets increased the classification accuracy of CNN, with the ResNet50 model delivering the best results. On this dataset, Viet et al. [30] additionally used the Attention ConvMixer neural network with PChA and PSA.

FishNet Open Image Database [16] was presented by Kay and Merrifield [16], who also published the baseline accuracy of fish detection using the Inception-V3 model. Mujtaba and Mahapatra [51] used transfer learning and data augmentation to create the Regulated Networks (RegNet)X-16GF model, which effectively classified fish species in electronic monitoring footage and showed improved accuracy on this dataset. In order to address seafood mislabeling, Mujtaba and Mahapatra [52] developed a deep-learning model using transfer learning with VGG, ResNet, and RegNet. The RegNetX-16GF model achieved the best results

in accurately distinguishing between morphologically similar fish species. Mujtaba and Mahapatra [53] used RegNetY-800MF fine-tuned with transfer learning to achieve the best results in fish species identification from electronic monitoring footage. Also on this dataset, Gong et al. [39] used label smoothing and Grad-CAM to enhance Fish-TViT.

Boulais et al. [13] introduced SEAMAPD21 [13], and presented the baseline accuracy using ResNet50 with SE blocks.

Zhuang et al. [15] developed WildFish++ [15], a large-scale vision-language benchmark focusing on fine-grained classification of fish species by leveraging comparison fish descriptions to improve CNNs' ability to identify subtle differences between similar species, and presented the baseline accuracy using Progressive Multi-Granularity (PMG) with training of Jigsaw Patches. Li et al. [41] also applied the T-KD approach, with IMSP and LTR, to this dataset.

Khan et al. [9] introduced FishNet [9], along with the baseline results using ConvNext with Class-Balanced Training.

In the course of a year, Ottaviani et al. [54] quantified the degradation in fish detection and classification performance using ResNet18 due to concept drift in visual data from the OBSEA cabled video-observatory [10] and discussed methodological solutions for effective real-time automated classification.

Finally, in the case of OzFish dataset [11], no results for the image classification task where found.

Table 1 summarizes the datasets and methods, with datasets listed in alphabetical order and methods for each dataset listed chronologically. Table 3 summarizes the accuracy results of the methods presented above by datasets. The next section presents a brief discussion about the methods.

C. DISCUSSION

The majority of state-of-the-art methods in the field of FGVC continue to rely on traditional machine learning approaches coupled with CNN. These techniques have proven to be quite effective in a number of applications, but their applicability to the generalization of this task is restricted. An exception to this pattern and a shift toward more modern architectures can be found in the suggested approach by Gong et al. [39], which utilized ViT. In addition, all methods presented have only been proven to work with a maximum of three datasets.

The most relevant prior work in this field is Fish-TViT by Gong et al. [39], which applied ViTs to fish classification. Fish-TViT used transfer learning, label smoothing, and Grad-CAM to enhance feature extraction and model optimization. While Fish-TViT showed improvements over CNN-based methods, our approach extends this work with the combination of the Swin Transformer with the FGVC-PIM. Our method aims to address a broader range of datasets and environmental conditions, potentially offering greater versatility and robustness in fish species classification tasks.

In conclusion, none of the methods presented a general-purpose solution for fish classification under various conditions and image quality levels. It is also clear that the majority of existing methods are highly localized, and not as flexible to new datasets or conditions, given that they are usually customized and fine-tuned for their particular dataset.

The localization of these methods is a significant drawback, as these techniques and methods are rendered less applicable to countless applications. For instance, they may lack the capacity to detect foreign species that could appear in different regions of the globe. By contrast, the development of a method that can deliver a uniform performance various datasets, handling different species and environments, would highlight its potential as a general framework for FGVC, ensuring a wider applicability and effectiveness for applications that require dependable performance in diverse contexts.

A method with the above characteristics can establish a new standard for future research in this field, while also improving upon the current state-of-the-art.

III. METHODOLOGY

The method proposed for fish classification in this study stands out due to its broad applicability across 14 distinct datasets. As shown, the previous state-of-the-art approaches have only been proven to work with a maximum of three datasets. The variety and resilience of the proposed method, which aims to provide a general-purpose solution for fish detection under various conditions and image quality levels, are demonstrated by the breadth of the datasets.

The Methodology is divided into Datasets and Methods. Section III-A outlines the preparation of the datasets used in this study and the proposed method for FGVC of fish species, i.e., it covers the datasets and the processes – data extraction, organization, and acquisition – required to prepare the images for training, including an analysis of the datasets' data distribution, particularly focusing on imbalances. Section III-B goes into great detail about our suggested approach, emphasizing in particular how the Swin Transformer [1] is integrated with the FGVC-PIM [2]. By combining the Swin Transformer as its backbone to leverage its hierarchical and efficient feature extraction capabilities with FGVC-PIM enhances the performance of Vision Transformer by focusing on the extraction of discriminative features crucial for fine-grained classification tasks.

A. DATASETS

This subsection goes into detail about the organization, acquisition, and extraction procedures for each dataset. The number of images and the distribution of classes is referred to illustrate the variations in conditions between datasets, which will be presented in descending order of the Normalized Shannon Entropy (NSE), as shown in Table 2.

TABLE 1. Summary of datasets and methods, with datasets listed in alphabetical order and methods for each dataset listed chronologically.

Dataset	Authors	Year	Model/Architecture
A Large-Scale Dataset for Fish Segmentation and Classification [3]	Mampitiya et al. [45]	2022	Random Forest
	Aziz et al. [46]	2023	EfficientNet B0 + COWOA
	Li et al. [41]	2023	T-KD - novel Fish37 dataset + IMSP + new LTR
	Reddy et al. [47]	2023	Custom CNN
	Kranthi Kumar et al. [48]	2023	MobileNetV2 + Transfer Learning
BDIndigenousFish2019 [4]	Islam et al. [4]	2019	HLBP feature with SVM
	Dey et al. [43]	2021	5-Layer CNN
	Smadi et al. [44]	2022	Custom CNN
	Viet et al. [30]	2023	Attention ConvMixer neural network with PChA and PSA
Croatian Fish Dataset [5]	Pang et al. [26]	2021	FSA + KLD
	Sudhakara et al. [27]	2022	UIEGAN + DCGAN + SmallerVGG
	Veluswami and Panneerselvam [28]	2022	CNN + Refined SE + SVM
	Chen [29]	2023	ResNet18_q + VQAs
	Viet et al. [30]	2023	Attention ConvMixer neural network with PChA and PSA
Fish-Gres [6]	Prasetyo et al. [49]	2022	MLR-VGGNet16
	Azhar et al. [50]	2023	ResNet50 pre-trained model with HE-augmented training data
	Viet et al. [30]	2023	Attention ConvMixer neural network with PChA and PSA
Fish-Pak [7]	Xu et al. [36]	2021	Combiner SE-ResNet512
	N.S. et al. [37]	2021	AlexNet with a naive Bayesian fusion layer
	Deka et al. [24]	2022	Fine-tuned ResNet50 + SVM classifier
	Deka et al. [38]	2023	ResNet50
	Gong et al. [39]	2023	Transfer Learning + ViT (Fish-TViT)
Fish4Knowledge [8]	Sun et al. [31]	2018	RGB-AlexNet-SVM
	Zhang et al. [23]	2022	AdvFish - Deep Adversarial Learning Framework
	Gong et al. [32]	2022	EfficientNet-ECA & EfficientNet-CA
	Yang et al. [33]	2022	FCMFDA-ELM
	Jiang et al. [34]	2024	TSMVT + MAWTS
	Qu et al. [35]	2024	ConvFishNet + PixelShuffle Upsampling
FishNet [9]	Khan et al. [9]	2023	ConvNeXt + Class-Balanced Training
FishNet Open Images Database [16]	Mujtaba and Mahapatra [51]	2021	RegNetX-16GF + Transfer Learning
	Mujtaba and Mahapatra [52]	2021	RegNetX-16GF + Transfer Learning (TunaConvNet-4)
	Kay and Merrifield [16]	2021	Inception-V3
	Mujtaba and Mahapatra [53]	2022	RegNetY-800MF + Fine-Tuned with Transfer Learning
	Gong et al. [39]	2023	Transfer Learning + ViT (Fish-TViT)
OBSEA [10]	Ottaviani et al. [54]	2022	ResNet18
QUT Fish Dataset [12]	Qiu et al. [17]	2018	B-CNN + Refined SE Blocks
	Hridayami et al. [18]	2019	VGG16 + Transfer Learning from ImageNet
	Adiwinata et al. [19]	2020	Faster R-CNN Inception-v2
	Islam et al. [20]	2021	CCV + LBP + SIFT + SURF
	Mathur and Goel [21]	2021	ResNet50 + Transfer Learning
	Guo et al. [22]	2021	Cross-Domain Transfer Learning
	Zhang et al. [23]	2022	AdvFish - Deep Adversarial Learning Framework
	Deka et al. [24]	2022	Fine-tuned ResNet50 + SVM Classifier
	Ahmad et al. [25]	2023	Efficient-Net Fine-Tuned with GAP and FL
SEAMAPD21 [13]	Boulais et al. [13]	2021	SE-ResNet50
WildFish [14]	Yu et al. [40]	2021	SCAF + MCML + Non-Rdn
	Zhang et al. [23]	2022	AdvFish - Deep Adversarial Learning Framework
	Li et al. [41]	2023	T-KD + novel Fish37 dataset + IMSP + new LTR
	Jiang et al. [34]	2024	TSMVT + MAWTS
	Manikandan and Santhanam [42]	2024	ADANSE Vision Transformer
WildFish++ [15]	Zhuang et al. [15]	2020	PMG + Training of Jigsaw Patches
	Li et al. [41]	2023	T-KD + novel Fish37 dataset + IMSP + new LTR

Shannon Entropy, $H(X)$, a probabilistic measurement that quantifies the degree of uncertainty or randomness in a dataset, is demonstrated by Equation 1,

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i), \tag{1}$$

with $H(X)$ being the entropy of an X dataset, n is the number of classes, $p(x_i)$ is the probability of a certain i class. It aids in figuring out how evenly distributed the class labels are in relation to class distributions. Whereas a lower value implies that some classes are significantly more common than others, a higher Shannon Entropy indicates a more uniform distribution.

Shannon's maximum entropy, defined by $H_{max}(X) = log_2 n$, represents the highest possible entropy for a given number of classes, which is the logarithm of the number of classes. Normalized Shannon Entropy (NSE or H'), is calculated using Equation 2, and normalizes the entropy value between 0 and 1,

$$H' = \frac{H(X)}{H_{max}(X)}. \tag{2}$$

A perfectly balanced dataset is indicated in this case by the maximum NSE value, which shows that every class in the dataset has the same amount of images. Alternatively, a lower value signifies a high level of imbalance, with notable differences in the quantity of images in various classes.

The Minority-Majority Ratio (MMR) calculates the relation between the number of images in the smallest class relative to the biggest class. With certain classes being notably underrepresented, a lower ratio denotes a higher degree of imbalance. The majority and minority classes are the only ones represented in this ratio – the full dataset is not included.

Box and whisker plots were used on a logarithmic scale to show the distribution of images by class in the datasets. By emphasizing the skewness and variability in the number of images across different classes, this method successfully illustrates the class probabilities. Normalizing all datasets metrics allows for meaningful comparisons between them. A better understanding of the dataset's class distribution is obtained by plotting the data logarithmically, which allows to more easily observe differences between classes, especially those with notably fewer or more images.

As already mentioned, sample images from each dataset are shown in Figure 1, where the different conditions in which the images were taken can be observed. The pre-training processed datasets, including the number of classes, images, MMR, NSE, and class distribution, are summarized in Table 2.

The Large-Scale Dataset for Fish Segmentation and Classification [3] contains 9 classes with 1,000 images each, totaling 9,000 images. For this dataset description and the ones that follows, please see Section II-A. The images are already augmented by rotation. The dataset is well-balanced, with a NSE, and a MMR, of 1.0000, indicating that each class has the same number of images. The class distribution can also be observed as a box and whisker plot on the right side of Table 2, where a single line can be observed in the logarithm scale, representing the absence of variability of the amount of images per class.

The QUT Fish Dataset [12] images are available in raw format and already cropped. The raw images were used for these experiments. The images were captured under five conditions – *controlled*, *in-situ*, *rubbish*, *sketches*, and *uncontrolled* – the previous mentioned condition *out-of-the-water* is the *uncontrolled* images. This dataset was divided into three subsets: A, B, and C. Subset A contains 27 classes, all of which are *in-situ*, which means they were taken underwater, with every class having at least 10 images. Subset B and C contain images from the *controlled*, *in-situ*, and *uncontrolled* conditions. Subset B consists of the top 50 classes by the quantity of images available.

The variation between these two subsets is minimal on the NSE, whereas the difference on the MMR is more noticeable, due to the greater disparity in the number of images per class for the majority and minority classes, with the subset A having a difference of 6 images between the minority and majority classes, and the subset B having a difference of 9.

Subset C contains all the classes with a minimum of 3 images per class, resulting in a total of 464 classes, after excluding the images from the *rubbish* and *sketches* conditions. On this subset, it is easier to observe the imbalance in the box and whisker plot in Table 2, in comparison to the other subsets, with the minority class having 3 images and the majority class having 26 images.

Although the FishNet dataset [9] contains 17,357 aquatic species, the dataset was created aiming the classification of the species' family or order. Both were considered for the experiments, with 462 family classes and 83 order classes. The NSE starts to decay as the class distribution becomes more imbalanced, as can be seen on Table 2. Either of these datasets splits have also a disparity associated to the minority and majority classes, with the family classes having a ratio of 4 by 4,782 and the order classes having a ratio of 4 by 21,827. To create a more practical subset for the experiments, only 199 classes were included.

The FishNet dataset [9] provides bounding box annotations for each image, targeting either the family or order level. To further refine the dataset, species were correlated, and only those with more than 30 images were considered, ensuring a minimum threshold of 165 images for inclusion. This experimental dataset subset contains 52,149 images, with a NSE of 0.9959, and a MMR of 0.4940, indicating a moderate imbalance in the dataset. The class distribution can be observed in the box and whisker plot in Table 2.

Both the WildFish dataset [14] and the WildFish++ dataset [15] represent a large and challenging dataset, with the first containing 1,000 classes of fish species and the second containing 2,348 classes. Both publications present different challenges associated with additional metadata or specific tasks. For this experiment, the complete set of both datasets

were used. Some of the original images were corrupted, and the dataset was pre-processed to remove them. The WildFish dataset remained with 54,453 images, while the WildFish++ filtered dataset contained 103,025 images.

The above-mentioned datasets have a slight long-tail distribution, which greatly affects the MMR as evident from Table 2, while the class distribution is relatively balanced overall, there are a few outlier classes that deviate from the general pattern, while the rest of the distribution is still condensed, even with both datasets presenting the largest number of classes. This is reflected in the NSE, with the WildFish dataset having a value of 0.9903 and the WildFish++ dataset having a value of 0.9882.

The Fish-Gres dataset [6] did not require any pre-processing, as it was already well-organized. Nevertheless, it is important to note that there are multiple images that contain more than one specimen, which can lead to misclassification in the task of fine-grained image classification, where only one sample of the species per image is preferred. The class distribution for these 8 classes varies from 240 images to 577 images per class.

The BDIndigenousFish2019 dataset [4] did not need any pre-processing as well. Similar to the Large-Scale Dataset for Fish Segmentation and Classification [3], the images were augmented by rotation. As the one before, this dataset also contains 8 classes, although the number of images per class varies from 120 to 500.

The Croatian Fish Dataset [5] also did not require any pre-processing. This dataset contains 12 classes, with the number of images per class varying from 17 to 111. It is important to note that the smallest image in this dataset has 36×12 pixels, while the largest has 503×231 pixels. The class distribution can be observed in the box and whisker plot in Table 2.

For the Fish-Pak dataset [7], only 4 corrupted images were removed, and from the three available subsets – body, head, and scale – only the body subset was used. This dataset contains 6 classes, with the number of images per class varying from 11 to 73. These dataset images appear to contain multiple poses of the extremely similar specimens.

The OzFish dataset [11] contains the metadata in a separate file and then bounding boxes already cropped from the original footage. The images were distributed by genus and species. From the original 594 classes, only those with more than 10 images were considered, resulting in 425 classes. The number of images per class varies from 10 to 6,095, where a long-tailed distribution can be observed. This class distribution with the existence of multiple outliers is visible in the box and whisker plot in Table 2. A lower NSE was also observed, with a value of 0.7900, and a very low MMR, with a value of 0.0003.

A long-tailed distribution is also observed in the SEAMAPD21 dataset [13], where the minority class has 3 images, while the majority class has 15,199 images. For this experiment, only classes with a minimum of 10 images were considered, with 110 classes remaining and 8 classes discarded. Nevertheless, the majority class still corresponds to more than 19% of the dataset. The metadata containing the bounding boxes' locations was used to crop the images. The missing association between some images and the corresponding bounding boxes was discarded. The images were distributed into genus and species.

The OBSEA dataset [10] follows the same long-tailed distribution as the previous datasets. The metadata containing the bounding boxes' locations was used to crop the images. The unknown species and 45 corrupted images were discarded. This experiment only considered the species with a minimum of 10 images per class, resulting in 25 classes. The class distribution observed in the box and whisker plot in Table 2 results from the minority class having 10 images and the majority class having 14,299 images, with the first and second major classes corresponding to 57.7% of the dataset.

For Fish4Knowledge dataset [8], the fish recognition ground-truth data was used, containing 27,370 fish images captured by nine cameras at three different locations in unconstrained environments. The dataset contains 23 classes, with the number of images per class varying from 16 to 12,112, with the top 5 classes representing 91% of the dataset, with a NSE of 0.5721. This imbalanced class distribution can be observed in the box and whisker plot in Table 2. The images were already cropped and prepared for training.

The FishNet Open Images Database [16] contains two subsets – L1, and L2 – with L1 consisting of 21 classes, and L2 consisting of 10 classes. The class distribution for both subsets is long-tailed, and both contain ambiguous labels. The L1 subset does not follow the conventional scientific classification – genus species – and the L2 subset contains coarser classes based on the ASFIS List of Species for Fishery Statistics Purposes. The cropped images were obtained from the bounding boxes' location available. On both L1 and L2 subsets, the majority class represents more than 50% of the dataset. There are also more than 10,000 duplicates present in this dataset, and a large portion of images containing more than one specimen. These subsets obtained the lowest NSE, with the L1 subset having a value of 0.4399, and the L2 subset having a value of 0.5546. The class distribution, especially the outliers of the top classes, can be observed in the box and whisker plot in Table 2.

In the next section, the authors' method is explained in further detail.

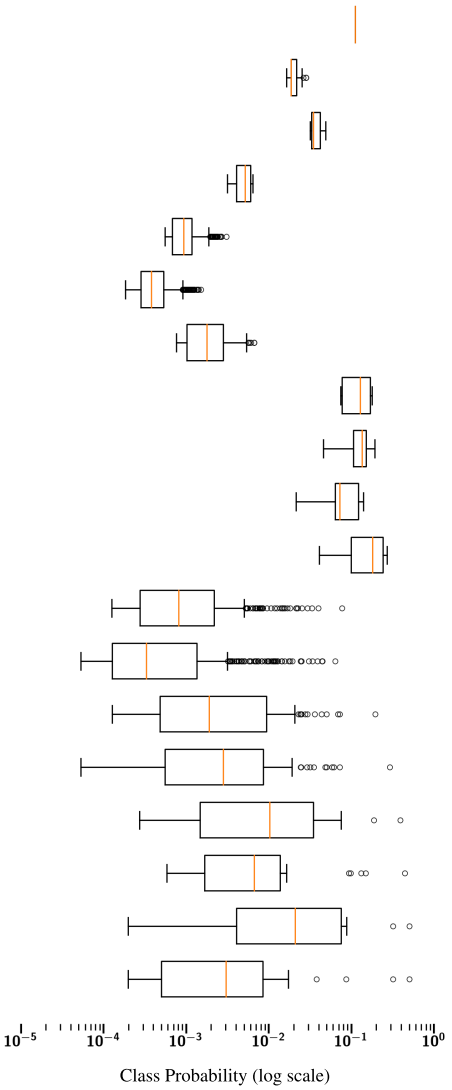
B. METHODS

Our method applies the FGVC-PIM with Swin Transformer backbone, an architecture originally designed for fine-grained visual classification. While this integrated architecture was initially optimized for bird classification, its effectiveness for fish species identification across diverse aquatic environments was explored.

To adapt this architecture for fish classification, the input pipeline was modified to handle the varied image formats and sizes found in fish datasets, implementing consistent preprocessing for underwater, above-water, and controlled

TABLE 2. Overview of datasets listed in descending order by the Normalized Shannon Entropy, also including the number of classes, images, Minority-Majority Ratio, and data distribution.

Dataset	# Classes	# Images	NSE	MMR	Data Distribution
A Large-Scale Dataset for Fish Segmentation and Classification [3]	9	9000	1.0000	1.0000	
QUT Fish Dataset – Subset B [12]	50	916	0.9975	0.5769	
QUT Fish Dataset – Subset A [12]	27	348	0.9969	0.6471	
FishNet [9] – Species (165 images minimum per class)	199	52149	0.9959	0.4940	
WildFish [14] – Complete-Set	1000	54453	0.9903	0.1796	
WildFish++ [15] – Complete-Set	2348	103025	0.9882	0.1226	
QUT Fish Dataset – Subset C [12]	464	3924	0.9746	0.1154	
Fish-Gres [6]	8	3248	0.9687	0.4159	
BDIndigenousFish2019 [4]	8	2610	0.9681	0.2400	
Croatian Fish Dataset [5]	12	794	0.9593	0.1532	
Fish-Pak [7]	6	271	0.9159	0.1507	
OzFish [11] (10 images minimum per class)	425	79503	0.7900	0.0003	
FishNet [9] – Family	462	94103	0.7669	0.0008	
SEAMAPD21 [13] (10 images minimum per class)	110	78546	0.7342	0.0002	
FishNet [9] – Order	83	94103	0.6893	0.0002	
OBSEA [10] (10 images minimum per class)	25	36710	0.6218	0.0001	
Fish4Knowledge [8]	23	27370	0.5721	0.0013	
FishNet Open Images Database – L2 [16]	10	256632	0.5546	0.0004	
FishNet Open Images Database – L1 [16]	21	257635	0.4399	0.0004	



environment images. The Feature Pyramid Network (FPN) and selection modules were adjusted to focus on fish-specific characteristics.

The training strategy involved a flexible regime based on dataset size and complexity. A balanced loss function strategy was also implemented to focus on the most relevant features for fish classification while preventing overfitting.

These adaptations allowed us to explore the full potential of the FGVC-PIM architecture in the context of fish classification, demonstrating its versatility across different aquatic environments and image qualities.

The foundation for the Fine-Grained Visual Classification method is the Swin Transformer, which was first presented by Liu et al. [1]. Each of the four stages composing the Swin Transformer architecture is responsible for processing image patches at different degrees of complexity and abstraction, using a Multi-Head Self-Attention (MSA) mechanism.

Figure 2 shows a block diagram of the method. The input image, with a width and resolution of 384 pixels and 3 channels, is first split into non-overlapping patches in the Patch Embedding block, each of which is handled as a token. A convolutional window with a kernel size and stride of 4 by 4 retrieves these small patches and feed them to a normalization layer. These patches are processed via multiple blocks of Swin Transformer. By computing self-attention within local windows, the successive Window-based MSA (W-MSA) and Shifted Window-based MSA (SW-MSA) mechanisms, contained in the Swin Transformer Block, effectively captures local dependencies. While generating low-level features, this stage preserves

fine-grained details that are essential for the original image representation.

Patch merging layers that combine features from nearby patches are incorporated in the following stages, while on the Stage 1 block, the patches enter through the Linear Embedding block. In doing so, the feature dimensionality is increased and the resolution is effectively downsampled. The number of channels is doubled at each second, third, and fourth stages of the hierarchical downsampling, while lowering the spatial resolution by a factor of two.

As a result of this design, the model can comprehend intricate patterns and hierarchical relationships within the image and is able to capture mid-level and high-level features. Modeling long-range interactions while maintaining linear computational complexity in relation to image size is made possible by the shifted window approach, which makes it possible to capture cross-window dependencies efficiently.

A residual connection, a feed-forward network (MLP layers), a MSA module, and layer normalization make up each Swin Transformer block. While layer normalization standardizes the inputs and enhances the network's convergence, residual connections guarantee stable training by addressing the vanishing gradient issue. The feed-forward network enhances the features that the successive W-MSA and SW-MSA module extracted.

The integration and enhancement of the Swin Transformer was achieved by Chou et al. [2], who proposed FGVC-PIM. By providing pixel-level feature maps and combining filtered features to improve FGVC, this module is intended to be a versatile add-on to popular backbones like CNN and transformer-based networks. In order to improve classification performance, the main purpose of this module is to locate and highlight the image regions that are the most discriminative.

To accomplish fine-grained classification, the FGVC-PIM architecture consists of a number of essential parts. First, feature maps are extracted from the input image by the backbone model. In this case, the Swin Transformer is integrated with a FPN, in order to handle multi-scale features. After that, these feature maps are run through the FGVC-PIM, which consists of a Weakly Supervised Selector (WSS) and a Combiner.

The feature maps (C1, C2, C3, C4) obtained on each stage of the Swin Transformer are passed through a FPN, with each one going directly to the correspondent Pyramid Level Blocks (P1, P2, P3, P4). Each of these blocks contains a projection layer that ensures that all the feature maps from different stages of the backbone network are converted to a common feature dimension. The WSS uses a fully connected layer to predict the category of each feature point in the extracted feature maps. It is ensured that only the most discriminative features are used in subsequent stages by keeping feature points with high confidence scores and discarding those with low confidence.

The selected feature points from the four WSS are fed into the Combiner, which uses a Graph Convolutional Network (GCN), which can efficiently combine each features points without changing the results of the backbone model. The GCN Combiner views the graph as a collection of nodes representing features at different spatial scales and locations, when receiving the feature points that have been selected, which occurs in the Graph Convolutional Layers' block. The network can comprehend the relationships between different nodes in the Attention Mechanisms block and capture the spatial and contextual relationships between the features with the aid of this graph structure. Following that, these relationships are combined into super nodes by a pooling layer. The averaged features of these super nodes are fed to a linear classifier, which produces the final class predictions.

Targeting distinct phases of the feature selection and combination process, the FGVC-PIM uses a loss function strategy that combines multiple loss functions. A process of careful modification of these loss functions' weights equalizes their contributions, which enables the model to efficiently concentrate on the most relevant features, which also prevents overfitting to the training set. The FGVC-PIM's performance of the Swin Transformer in fine-grained visual classification tasks is markedly improved by the accurate tuning, providing a reliable and accurate framework that distinguishes between detailed visual features.

The configurations, tests and findings of this method are presented in the next Section.

IV. EXPERIMENTS, RESULTS, AND DISCUSSION

The effectiveness of the FGVC-PIM, with the Swin Transformer as backbone, was assessed using the datasets listed in Section III-A, compared against the state-of-the-art methods presented in Section II-B. Now, this Section will detail the hyperparameters used in these experiments and describe how the data was split for training and testing. Following that, the results of the proposed method will be presented and compared against the existing approaches. Finally, the results will be discussed.

A. EXPERIMENTS AND RESULTS

The available splits were observed for the datasets with predefined train-test splits [9], [16], to ensure consistency with prior research. In order to ensure reproducibility across the other datasets, a uniform split strategy was implemented with a generator seed of 42. The data was divided as follows: 10% were set aside for model testing, 20% for validation, and 70% for training.

With a data input size of 384×384 pixels, the Swin Transformer serves as the foundation for FGVC in this study. The training uses 12 worker threads for data loading and a batch size of 16. The configurations for the Stochastic Gradient Descent (SGD) optimizer have a weight decay and maximum learning rate of 0.0005. There is an 800 batch warm-up phase before the training, which lasts from 10 epochs to 50 epochs, depending on the dataset size. For computational efficiency, mixed-precision training is enabled.

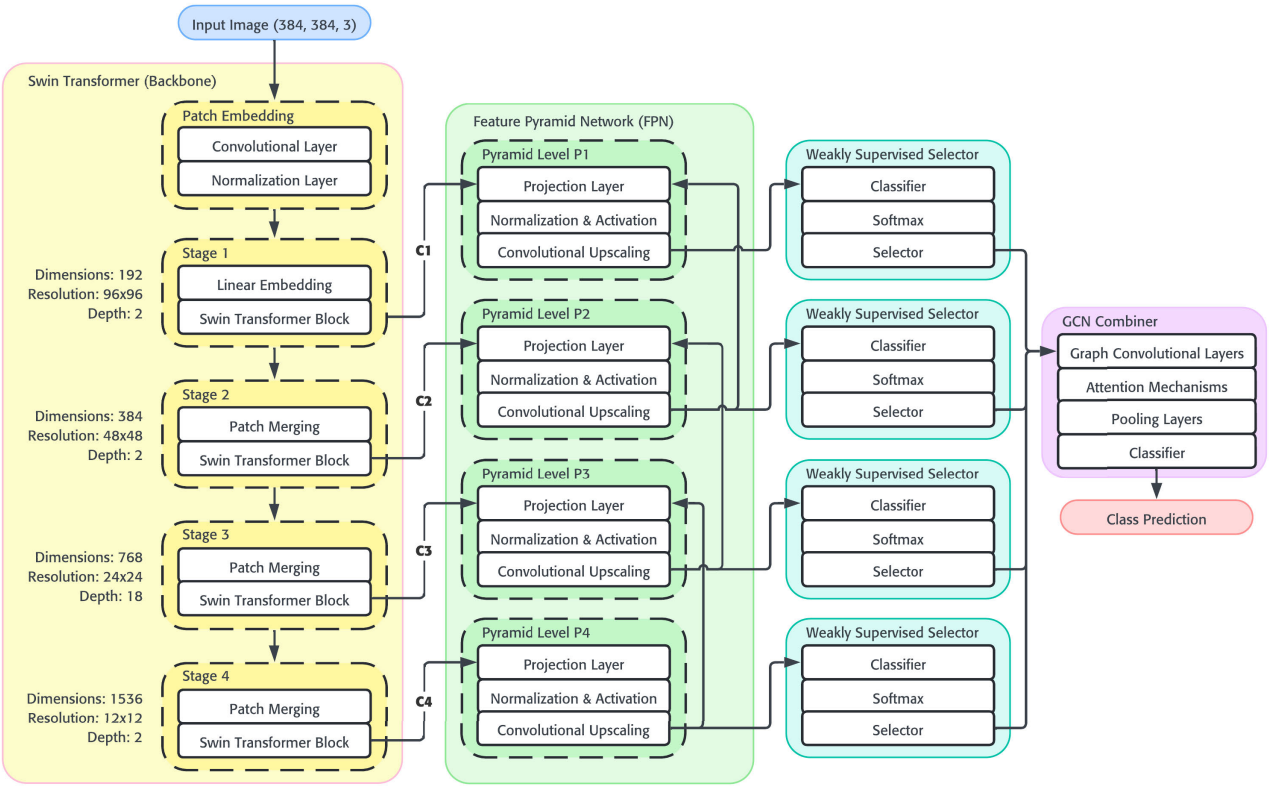


FIGURE 2. Architecture of the FGVC-PIM, a Swin Transformer-based Feature Pyramid Network with Weakly Supervised Selector and Graph Convolutional Network Combiner for robust class prediction.

To handle multi-scale features, an integrated FPN with a feature size of 1536 is used. The selection modules in the model select 2048, 512, 128 and 32 feature points from the respective layers, concentrating on the most discriminative regions. To improve classification performance, these features are combined by the Combiner module. To ensure that the contributions of the different loss components are optimized, the loss function are balanced with specific weights: base loss (0.5), selection loss (0.0), drop loss (5.0), and combiner loss (1.0).

After each batch, the model's parameters are updated, and assessments are carried out every 5 epochs. The model is intended to greatly enhance FGVC performance by utilizing the multi-scale and hierarchical properties of the Swin Transformer in conjunction with the FGVC-PIM improvements.

From the 14 datasets utilized in this study, which are presented in Section II-A, with the used data further detailed in Section III-A, 19 experiments were conducted to account for subsets within the datasets. Specifically, the FishNet dataset [9] featured the original family and order subsets, along with an additional species subset generated for this publication; the FishNet Open Images Database [16] included subsets L1 and L2; and the QUT fish dataset [12] comprised three distinct subsets. It is important to note that the task of classifying the family of order of a fish species is not the purpose of this publication, which is focused on FGVC; nevertheless, the obtained results on these tasks presents novel state-of-the-art accuracies in both in comparison to the baseline previously published. Detailed information on the data splitting for each subset can be found in Section III-A.

All the values presented with the proposed method were obtained either respecting the training and testing splits provided by the authors or by using a uniform split strategy, with the metrics corresponding to the top-1 accuracy of the combiner over the test set, which corresponds to 10% of each dataset.

The suggested approach produced state-of-the-art results in 13 out of the 19 experiments. Furthermore, in three of the experiments, the amount of classes trained are higher than the previous state-of-the-art results. Additionally, because comparative results were not available, two experiments only presented baseline results. This thorough assessment shows the stability and efficacy of the proposed method on a wide range of inconsistent visual classification, varying across image quality and conditions. Figure 3 provides a comprehensive visual comparison of these results, illustrating the consistent high performance of our method across diverse datasets.

Table 3 presents the accuracy [55] values, which range from 83.65% to 100.00% for the top-1 state-of-the-art accuracy, and Figure 4 demonstrates some of the images

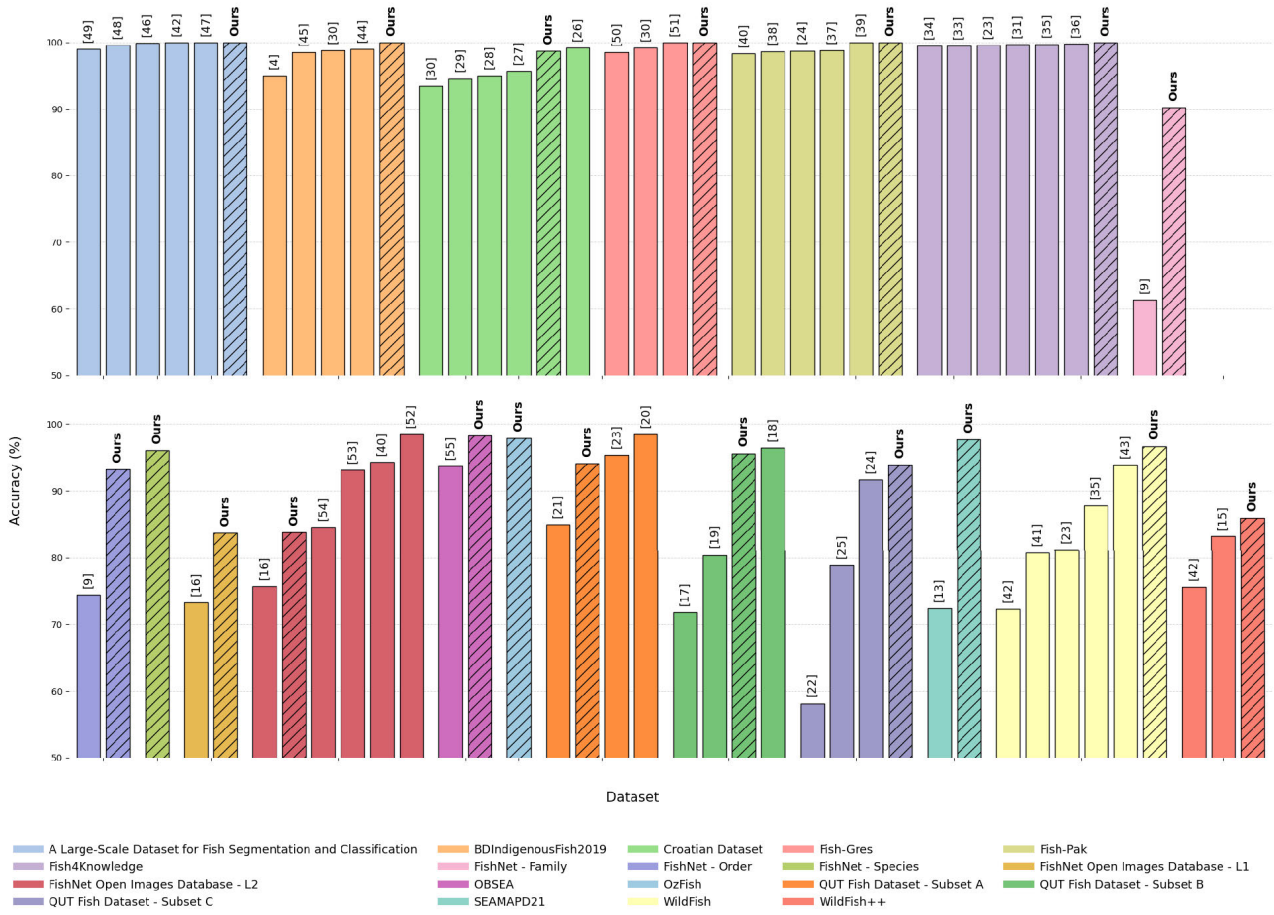


FIGURE 3. Comparison of accuracy percentages across multiple fish classification datasets, split into top and bottom panels for clarity. Each bar represents a different model, with textured bars marked "Ours" showing our current study's results. The present method consistently achieves high accuracy across diverse datasets, demonstrating robustness in various fish classification tasks. A comprehensive legend below identifies all datasets from both panels, facilitating easy comparison across the entire range of studies.

that were misclassified by the corresponding models. The experiments which achieved 100.00% in accuracy are further studied bellow, where additional tests and metrics are presented in Table 4.

In the QUT fish dataset [12] first two subsets, and the Croatian Fish Dataset [5], where state-of-the-art results was not achieved, the difference in accuracy compared to the current leading methods varied between 0.49% and 4.34%. Notably, the most significant deviation was observed in the FishNet Open Images Database [16], where this method's accuracy was 14.71% below the current state-of-the-art, achieving only 83.78%. After further analysis on the Mujtaba and Mahapatra [51] publication, where the authors used data augmentation in order to achieve a balanced dataset, it is possible to observe on the presented training accuracy and loss plots that the reported accuracy pertains to the training set rather than the test set, which could explain a part of the discrepancy.

Regarding the FishNet Open Images Database [16], the proposed method achieved an 83.65% accuracy on the L1 subset, which is 10.45% higher than the current state-of-the-art baseline. It is important to note that Gong et al. [39] achieved the second-highest accuracy in the L2 subset results employing ViT.

This proposed method achieved multiple state-of-the-art results using the same hyperparameters, demonstrating the potential of Transformers in the task of FGVC across diverse datasets. Furthermore, additional metrics were calculated for clarity and to facilitate future comparisons with the obtained results, as presented in Table 5. These metrics [55] include the precision, recall, and F1-score for each dataset using either the macro average or the weighted average, as well as the Matthews Correlation Coefficient (MCC) for Multiclass, and also the top-1 accuracy from the Combiner. Considering t_k representing the number of times class k actually occurred, p_k representing the number of times class k was predicted, c representing the total number of samples correctly predicted, and s representing the total number of samples, MCC is given by

$$MCC = \frac{c \times s - \sum_k^K (p_k \times t_k)}{\sqrt{(s^2 - \sum_k^K p_k^2) \times (s^2 - \sum_k^K t_k^2)}}. \quad (3)$$

TABLE 3. The datasets, and their corresponding subsets, are listed in the table alphabetically. The number of classes used by each author and the corresponding top accuracy attained are also presented. This thorough overview makes it possible to compare the effectiveness of the proposed method versus various state-of-the-art methods in detail across a range of datasets and their corresponding subsets.[1]

Dataset	Authors	# Classes	Accuracy
A Large-Scale Dataset for Fish Segmentation and Classification [3]	**Veiga, Ricardo et al.**	**9**	**100.00%**
	Aziz et al. [46]	9	99.99%
	Li et al. [41]	9	99.96%
	Mampitiya et al. [45]	9	99.89%
	Reddy et al. [47]	9	99.60%
	Kranthi Kumar et al. [48]	9	99.00%
BDIndigenousFish2019 [4]	**Veiga, Ricardo et al.**	**8**	**100.00%**
	Dey et al. [43]	8	99.00%
	Viet et al. [30]	8	98.77%
	Smadi et al. [44]	8	98.46%
	Islam et al. [4]	8	94.97%
Croatian Fish Dataset [5]	Pang et al. [26]	12	99.22%
	Veiga, Ricardo et al.	**12**	**98.73%**
	Sudhakara et al. [27]	12	95.64%
	Veluswami and Panneerselvam [28]	12	94.99%
	Chen [29]	12	94.55%
	Viet et al. [30]	12	93.40%
Fish-Gres [6]	**Veiga, Ricardo et al.**	**8**	**100.00%**
	Azhar et al. [50]	8	100.00%
	Viet et al. [30]	8	99.20%
	Prasetyo et al. [49]	8	98.46%
Fish-Pak [7]	**Veiga, Ricardo et al.**	**6**	**100.00%**
	Deka et al. [38]	6	100.00%
	Xu et al. [36]	6	98.80%
	Deka et al. [24]	6	98.74%
	N.S. et al. [37]	6	98.64%
	Gong et al. [39]	6	98.34%
Fish4Knowledge [8]	**Veiga, Ricardo et al.**	**23**	**100.00%**
	Qu et al. [35]	23	99.80%
	Jiang et al. [34]	23	99.71%
	Sun et al. [31]	23	99.68%
	Zhang et al. [23]	23	99.60%
	Gong et al. [32]	23	99.50%
	Yang et al. [33]	23	99.50%
FishNet - Family [9]	**Veiga, Ricardo et al.**	**462**	**90.18%**
	Khan et al. [9]	463	61.38%
FishNet - Order [9]	**Veiga, Ricardo et al.**	**83**	**93.25%**
	Khan et al. [9]	83	74.38%
FishNet - Species [9]	**Veiga, Ricardo et al.**	**199**	**96.05%**
FishNet Open Images Database - L1 [16]	**Veiga, Ricardo et al.**	**21**	**83.65%**
	Kay and Merrifield [16]	21	73.20%
FishNet Open Images Database - L2 [16]	Mujtaba and Mahapatra [51]	10	98.49%
	Gong et al. [39]	12	94.33%
	Mujtaba and Mahapatra [52]	4	93.15%
	Mujtaba and Mahapatra [53]	9	84.56%
	Veiga, Ricardo et al.	**10**	**83.78%**
	Kay and Merrifield [16]	10	75.70%
OBSEA [10]	**Veiga, Ricardo et al.**	**25**	**98.28%**
	Ottaviani et al. [54]	14	93.70%
OzFish [11]	**Veiga, Ricardo et al.**	**425**	**97.90%**
QUT Fish Dataset - Subset A [12]	Islam et al. [20]	21	98.46%
	Zhang et al. [23]	27	95.38%
	Veiga, Ricardo et al.	**27**	**94.12%**
	Mathur and Goel [21]	37	84.92%
QUT Fish Dataset - Subset B [12]	Hridayami et al. [18]	50	96.40%
	Veiga, Ricardo et al.	**50**	**95.60%**
	Adiwinata et al. [19]	50	80.40%
	Qiu et al. [17]	60	71.80%
QUT Fish Dataset - Subset C [12]	**Veiga, Ricardo et al.**	**464**	**93.88%**
	Deka et al. [24]	395	91.63%
	Ahmad et al. [25]	463	78.91%
	Guo et al. [22]	482	58.09%
SEAMAPD21 [13]	**Veiga, Ricardo et al.**	**110**	**97.67%**
	Boulais et al. [13]	54	72.40%
WildFish [14]	**Veiga, Ricardo et al.**	**1000**	**96.68%**
	Manikandan and Santhanam [42]	1000	93.90%
	Jiang et al. [34]	1000	87.72%
	Zhang et al. [23]	1000	81.12%
	Yu et al. [40]	1000	80.75%
	Li et al. [41]	1000	72.23%
WildFish++ [15]	**Veiga, Ricardo et al.**	**2348**	**85.93%**
	Zhuang et al. [15]	2348	83.20%
	Li et al. [41]	2348	75.54%

The precision measures the accuracy of positive predictions; the recall, or sensitivity, measures the models' ability to identify all relevant instances, and the F1-score combines both metrics into a single value, providing a balance by taking their harmonic mean. By averaging metrics determined for each class, the macro average applies an equal treatment to all classes. However, the weighted average accounts for class size, favoring larger classes, which is helpful for datasets that are unbalanced.

When dealing with imbalanced datasets, the MCC for Multiclass metric is particularly helpful in assessing the quality of a classification model. False positives, false negatives, true positives, and true negatives are all considered by the metric. The additional metrics displayed in Table 5 demonstrate how well the suggested approach performs on a variety of datasets, including the FishNet dataset [9] subsets.

The results suggest that the full potential of the Transformer architecture is not being realized in the classification

[1]The values presented by the authors of this publication are all from unseen data during the training phase, either by the available published splits or by using a uniform split strategy, consistently considering 10% for testing.

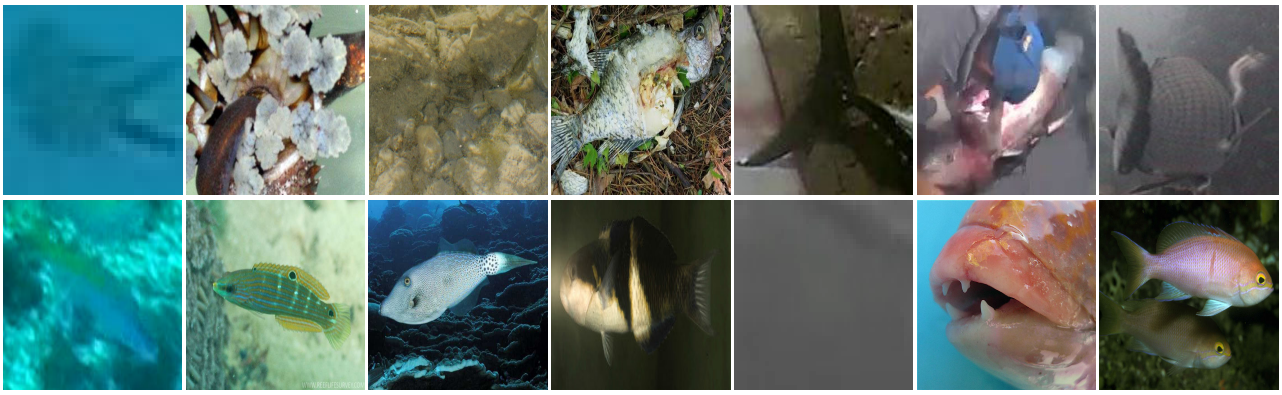


FIGURE 4. Failed detection samples from the proposed method. The top row, from left to right, includes failure samples from the Croatian dataset [5], FishNet – Family, Order, and Species [9], FishNet Open Images Database – L1, and L2 [10], and OBSEA [10]. The bottom row features failure samples from OzFish [11], QUT Fish Dataset Subsets A, B, and C [12], SEAMAPD21 [13], WildFish [14], and WildFish++ [15].

tasks for the family and order subsets of the FishNet dataset [9]. These broader taxonomic categories likely do not require the fine-grained feature detection capabilities that Transformers excel at. In contrast, species-level classification tasks better leverage the Transformer's ability to capture subtle visual distinctions. While the method produced state-of-the-art results, the macro average results, when the classes are treated equally, show that some classes are being severely misclassified. This suggests that there are significant differences in the model's classification performance, indicating a struggle with some classes. These difficulties are lessened, nevertheless, when using the FishNet dataset [9]'s novel species subset, where the suggested approach performs admirably on all metrics. This implies that the technique works better for more detailed classification tasks, where the fine-grained capabilities of the Transformer can be fully utilized.

According to the metrics obtained for the FishNet Open Image Database [16], the suggested method did not perform as well as the other methods on this dataset. The macro average precision, recall, and F1-score show this clearly, being notably lower than the values for other datasets. Because of the inherent difficulties in the data, these results imply that the method struggles with the FishNet Open Image Database [16]. Moreover, a lower MCC denotes a less optimal balance between true and false positives as well as negatives across the classes, which reflects the general decline in effectiveness of the model. These findings show how challenging it is to obtain a good classification performance on this specific dataset using the proposed method, suggesting that further data preprocessing may be necessary to obtain better results.

Comparing the WildFish++ dataset to the WildFish dataset, a decrease can also be seen in the overall performance metrics in Table 5. The more than twofold increase in the number of available classes significantly contributed to the increased complexity and the decreased performance. A future challenge would be the development of a foundational model, combined with additional metadata, aimed at classification of the maximum number of species. However, for both datasets, the results are still cutting edge and highly relevant, showcasing the adaptability and room for improvement of the proposed method when working with large and varied species datasets.

It was previously noted that some datasets achieved 100% accuracy, as shown in Table 3 and Table 5. In all of these occurrences, the datasets had fewer than 24 classes, proving that using Transformers for small-scale datasets outperforms the previous methods published. These results were obtained following the same experimental configuration and hyperparameters as the other datasets presented in this publication. To ensure the robustness of these findings, additional training was performed using random seeds for each training session, with the experiment repeated ten times per dataset. The results of these additional experiments are presented in Table 4.

With Table 4 it is possible to confirm that for the datasets: A Large-Scale Dataset for Fish Segmentation and Classification [3]; BDIndigenousFish2019 dataset [4]; and the Fish-Gres dataset [6], the 100% accuracy remains unchallenged, including the macro and weighted averages, as well as the MCC. For the Fish-Pak dataset [7], and the Fish4Knowledge dataset [8], it is possible to observe a small reduction over the previous obtained metrics, although these values still present state-of-the-art results.

Finally, Figure 5 provides some examples of how Grad-CAM [56] is used to show how well the model can differentiate between various species. The Grad-CAM visualization technique highlights the critical areas of an image for class prediction, and can be used to gain insight into the model's decision-making process. This method falls under the category of Explainable Artificial Intelligence (XAI), which aims to improve human comprehension and responsibility towards AI systems by enhancing the transparency and comprehensibility of AI model outputs.

TABLE 4. Additional experimental results for datasets with previously attained accuracy of 100%. Each dataset was trained ten times with randomized seeds.

Dataset	Macro Average			Weighted Average			MCC (%)	Top 1 Accuracy (%)
	Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)		
A Large-Scale Dataset for Fish Segmentation and Classification [3]	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
BDIndigenousFish2019 [4]	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
Fish-Gres [6]	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
Fish-Pak [7]	97.54±5.95	97.83±5.20	97.64±5.60	98.67±2.36	98.89±1.79	98.72±2.11	98.60±2.26	98.89±1.79
Fish4Knowledge [8]	98.77±2.09	98.78±1.89	98.71±1.94	99.91±0.05	99.90±0.04	99.90±0.04	99.86±0.06	99.90±0.04

TABLE 5. Extended assessment of the suggested method which also includes the precision, recall, and F1-score for each dataset using either the macro average or the weighted average. The corresponding top-1 accuracies from the combiner, the MCC for Multiclass is also provided.

Dataset	Macro Average			Weighted Average			MCC	Top 1 Accuracy (%)
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
A Large-Scale Dataset for Fish Segmentation and Classification [3]	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
BDIndigenousFish2019 [4]	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Croatian [5]	97.92%	99.44%	98.52%	99.05%	98.73%	98.80%	98.57%	98.73%
Fish-Gres [6]	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Fish-Pak [7]	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Fish4Knowledge [8]	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
FishNet - Family [9]	73.36%	71.45%	71.15%	90.16%	90.18%	89.99%	89.99%	90.18%
FishNet - Order [9]	84.94%	82.27%	83.12%	93.26%	93.25%	93.22%	92.40%	93.25%
FishNet - Species [9]	96.03%	96.14%	96.02%	96.17%	96.05%	96.05%	96.03%	96.05%
FishNet Open Images Database - L1 [16]	64.85%	68.88%	61.95%	82.10%	83.65%	81.85%	75.61%	83.65%
FishNet Open Images Database - L2 [16]	82.97%	68.58%	70.95%	82.05%	83.78%	81.64%	75.89%	83.78%
OBSEA [10]	92.82%	92.13%	92.38%	98.28%	98.28%	98.27%	97.84%	98.28%
OzFish [11]	96.21%	95.81%	95.66%	98.02%	97.90%	97.88%	97.87%	97.90%
QUT Fish Dataset - Subset A [12]	90.83%	92.50%	90.67%	93.63%	94.12%	92.94%	93.91%	94.12%
QUT Fish Dataset - Subset B [12]	95.16%	94.88%	94.39%	97.89%	95.60%	96.12%	95.51%	95.60%
QUT Fish Dataset - Subset C [12]	89.11%	88.75%	88.39%	95.24%	93.88%	93.92%	93.86%	93.88%
SEAMAPD21 [13]	95.23%	93.92%	94.13%	97.70%	97.67%	97.65%	97.51%	97.67%
WildFish - Complete-Set [14]	96.16%	96.16%	95.76%	97.12%	96.68%	96.61%	96.67%	96.68%
WildFish++ - Complete-Set [15]	83.82%	83.99%	82.34%	87.78%	85.93%	85.72%	85.92%	85.93%

B. DISCUSSION

The proposed FGVC achieved 13 state-of-the-art results across 14 distinct datasets, encompassing 19 experiments, and also presented 2 baseline results for a novel subset, and a dataset lacking published results. The results underscore the robustness and versatility of the method, as it consistently outperformed several existing state-of-the-art approaches across diverse datasets, each with unique characteristics. These datasets included varied environments such as onboard boats, underwater, and neutral backgrounds.

The model underperformed when multiple subjects were present in the same data, or when the classes weren't narrow enough, with the same subjects showing up in multiple classes. This restriction suggests that the proposed method performs best when each image primarily depicts a single subject and classes are clearly defined. However, the method's consistent top-1 accuracy scores, ranging from 83.65% to 100%, show that it has a great potential for a wide use in FGVC tasks. The results, which are shown in Table 3, highlight the effectiveness of the suggested method to handle various image qualities and conditions, establishing new industry standards.

In summary, the comparative results are presented across three key tables. Table 3 offers a comprehensive comparison of the method against various state-of-the-art approaches across multiple datasets. It details the number of classes and accuracy for each method, with the results highlighted in bold for easy identification. Each dataset is delineated to facilitate direct comparisons within specific contexts. Table 4 focuses on the datasets where the method achieved 100% accuracy. To demonstrate the robustness of these results, ten training runs with different random seeds were conducted for each of these datasets, providing additional performance metrics. Table 5 extends the analysis by presenting a wide range of performance metrics for all datasets used in the study. Finally, Figure 3 provides a comprehensive visual comparison of the accuracy achieved by the proposed method and existing approaches across all datasets examined in this study. This comprehensive view allows for a deeper understanding of the

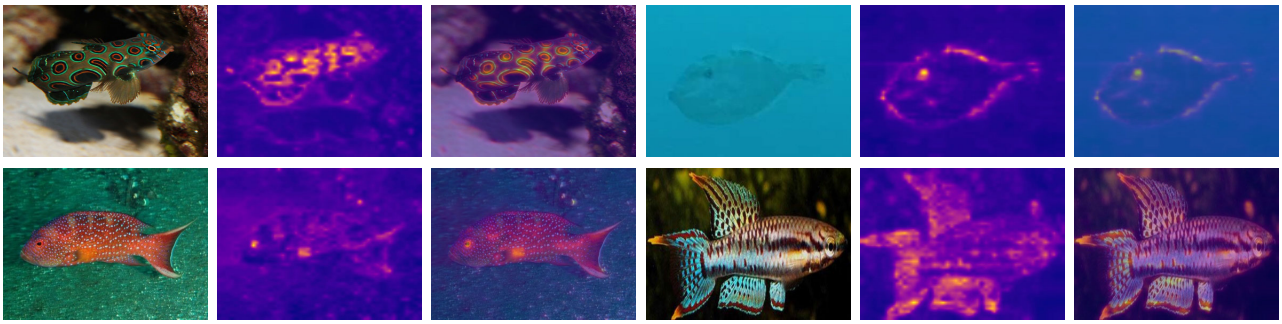


FIGURE 5. Examples of Grad-CAM technique to highlight the regions of interest. Four examples are depicted, each with three images: the original, the Grad-CAM result, and the Grad-CAM superimposed on the original. The top row contains an example from the WildFish++ dataset [15] and OzFish dataset [11]. The bottom row contains and example from the QUT Fish Dataset [12] and the FishNet Dataset [9].

method's effectiveness across various conditions and dataset sizes.

V. CONCLUSION

This study applies a new FGVC technique that utilizes the Swin Transformer combined with the FGVC-PIM module. After extensive testing on 14 different datasets with 19 different subsets, the method continuously outperformed other approaches and reached state-of-the-art outcomes in 13 of the experiments. These datasets validated the robustness and adaptability of the method across a wide range of conditions, including controlled laboratory settings, underwater environments, and on-board vessels.

This research demonstrates that Transformers are a useful tool for FGVC tasks, especially when working with a range of complex datasets. Because of the integration of the FGVC-PIM module with the Swin Transformer, the model was able to achieve high levels of classification accuracy by recognizing subtle differences between fish species. Nevertheless, the approach faced difficulties with datasets that included multiple subjects in the images or with class labels that were not sufficiently specific, suggesting areas that need to be further refined.

In summary, this work presents a comprehensive performance study that offers different analyses based on the distribution and features of the datasets, delivering a thorough assessment of the suggested method's performance, while also providing supplementary metrics. The broad use of the FGVC-PIM in conjunction with Swin Transformer as the foundation for the task of fish species FGVC was validated with a rigorous assessment over a variety of different datasets and subsets, using identical hyperparameters, demonstrating the adaptability and resilience of the suggested approach, covering a broad range of environmental conditions and variable image quality. The results highlighted the ViT's capability for FGVC tasks, introducing two new baselines and one unique subset, making a total of 19 datasets/subsets with 15 state-of-the-art outcomes, when counting the aforementioned.

Future work will concentrate on addressing the above-mentioned limitations by applying sophisticated preprocessing data techniques and fine-tuning the model to improve performance on unbalanced and ambiguous datasets. Additionally, it was planned to leverage the Fish-Vista dataset [57], which is based on museum images rather than *in-situ* captures and apply knowledge transfer techniques before proceeding to *in-situ* images. This intermediate step using Fish-Vista dataset [57] could potentially bridge the gap between controlled museum environments and more challenging real-world underwater conditions, potentially improving the model's generalization capabilities. Also, efforts will be made to combine the previous fish detection model [58] (previous work) with the classification (present work) to develop a complete framework that can work underwater in real-time, under real-world conditions.

The ultimate goal is to create a foundational model for marine life species that can manage an ample spectrum of FGVC tasks. To that end, this research aims to expand and add more metadata and classify additional species.

ACKNOWLEDGMENT

The authors would like to thank 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] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[2] P.-Y. Chou, C.-H. Lin, and W.-C. Kao, "A novel plug-in module for fine-grained visual classification," 2022, *arXiv:2202.03822*.

[3] O. Ulucan, D. Karakaya, and M. Turkan, "A large-scale dataset for fish segmentation and classification," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, Oct. 2020, pp. 1–5.

[4] M. A. Islam, M. R. Howlader, U. Habiba, R. H. Faisal, and M. M. Rahman, "Indigenous fish classification of Bangladesh using hybrid features with SVM classifier," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (ICME)*, Jul. 2019, pp. 1–4.

[5] J. Jäger, M. Simon, J. Denzler, V. Wolff, K. Fricke-Neuderth, and C. Kruschel, "Croatian fish dataset: Fine-grained classification of fish species in their natural habitat," in *Proc. BMVC*, vol. 2, Dec. 2015, pp. 1–7.

[6] E. Prasetyo, N. Suciati, and C. Fatichah. (2021). *Fish-Gres Dataset for Fish Species Classification*. [Online]. Available: https://data.mendeley.com/datasets/76cr3wfhff/1

[7] S. Z. H. Shah, H. T. Rauf, M. IkramUllah, M. S. Khalid, M. Farooq, M. Fatima, and S. A. C. Bukhari, "Fish-pak: Fish species dataset from Pakistan for visual features based classification," *Data Brief*, vol. 27, Dec. 2019, Art. no. 104565, doi: 10.1016/j.dib.2019.104565.

[8] R. B. Fisher, D. Giordano, and F.-P. L. Editors, *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*, vol. 104. Berlin, Germany: Springer, 2016. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84961753059

[9] F. F. Khan, X. Li, A. J. Temple, and M. Elhoseiny, "FishNet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20496–20506.

[10] M. Francescangeli, S. Marini, E. Martínez, J. Del Río, D. M. Toma, M. Nogueras, and J. Aguzzi, "Image dataset for benchmarking automated fish detection and classification algorithms," *Sci. Data*, vol. 10, no. 1, p. 5, Jan. 2023.

[11] *OzFish Dataset—Machine learning Dataset for Baited Remote Underwater Video Stations*, Austral. Inst. Mar. Sci. (AIMS), Univ. Western Aust. (UWA) Curtin Univ., 2019.

[12] K. Anantharajah, Z. Ge, C. McCool, S. Denman, C. Fookes, P. Corke, D. Tjondronegoro, and S. Sridharan, "Local inter-session variability modelling for object classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 309–316.

[13] O. Boulais, S. Y. Alaba, J. E. Ball, M. Campbell, A. T. Iftekhar, R. Moorhead, J. Prior, F. Wallace, H. Yu, and A. Zheng, "SEAMAPD21: A large-scale reef fish dataset for fine-grained categorization," in *Proc. 8th Workshop Fine-Grained Visual Categorization*, vol. 25, 2021, pp. 1–6. [Online]. Available: https://github.com/SEFSC/

[14] P. Zhuang, Y. Wang, and Y. Qiao, "WildFish: A large benchmark for fish recognition in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1301–1309.

[15] P. Zhuang, Y. Wang, and Y. Qiao, "Wildfish++: A comprehensive fish benchmark for multimedia research," *IEEE Trans. Multimedia*, vol. 23, pp. 3603–3617, 2021.

[16] J. Kay and M. Merrifield, "The fishnet open images database: A dataset for fish detection and fine-grained categorization in fisheries," 2021, *arXiv:2106.09178*.

[17] C. Qiu, S. Zhang, C. Wang, Z. Yu, H. Zheng, and B. Zheng, "Improving transfer learning and squeeze- and-excitation networks for small-scale fine-grained fish image classification," *IEEE Access*, vol. 6, pp. 78503–78512, 2018.

[18] P. Hridayami, I. K. G. D. Putra, and K. S. Wibawa, "Fish species recognition using VGG16 deep convolutional neural network," *J. Comput. Sci. Eng.*, vol. 13, no. 3, pp. 124–130, Sep. 2019.

[19] Y. Adiwinata, A. Sasaoka, I. P. Agung Bayupati, and O. Sudana, "Fish species recognition with faster R-CNN inception-v2 using QUT FISH dataset," *Lontar Komputer, Jurnal Ilmiah Teknologi Informasi*, vol. 11, no. 3, p. 144, Dec. 2020.

[20] S. M. M. Islam, S. I. Bani, and R. Ghosh, "Content-based fish classification using combination of machine learning methods," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 1, pp. 62–68, Feb. 2021.

[21] M. Mathur and N. Goel, "FishResNet: Automatic fish classification approach in underwater scenario," *Social Netw. Comput. Sci.*, vol. 2, no. 4, p. 273, Jul. 2021.

[22] C. Guo, B. Wei, and K. Yu, "Deep transfer learning for biology cross-domain image classification," *J. Control Sci. Eng.*, vol. 2021, pp. 1–19, Dec. 2021.

[23] Z. Zhang, X. Du, L. Jin, S. Wang, L. Wang, and X. Liu, "Large-scale underwater fish recognition via deep adversarial learning," *Knowl. Inf. Syst.*, vol. 64, no. 2, pp. 353–379, Feb. 2022.

[24] J. Deka, S. Laskar, and B. Baklial, "Freshwater fish species classification using deep CNN features," *ICTACT J. Image Video Process.*, vol. 12, no. 4, pp. 2721–2729, 2022.

[25] U. Ahmad, M. Junaid Ali, F. A. Khan, A. A. Khan, A. U. Rehman, M. Muhammad Ali Shahid, M. A. Haq, I. Khan, Z. S. Alzamil, and A. Alhussen, "Large scale fish images classification and localization using transfer learning and localization aware CNN architecture," *Comput. Syst. Sci. Eng.*, vol. 45, no. 2, pp. 2125–2140, 2023.

[26] J. Pang, W. Liu, B. Liu, D. Tao, K. Zhang, and X. Lu, "Interference distillation for underwater fish recognition," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13188. Springer, 2022, pp. 62–74.

[27] M. Sudhakara, M. J. Meena, K. R. Madhavi, P. Anjaiah, and L. N. Prakash, "Fish classification using deep learning on small scale and low-quality images," *Int. J. Intell. Syst. Appl. Eng.*, vol. 10, no. 1s, pp. 282–293, 2022.

[28] J. R. S. Veluswami and N. Panneerselvam, "Multi-species fish identification using hybrid DeepCNN with refined squeeze and excitation architecture," *Aquatic Sci. Eng.*, vol. 37, no. 4, pp. 220–228, 2022.

[29] Y. Chen, "A novel image classification framework based on variational quantum algorithms," 2023, *arXiv:2312.07932*.

[30] T. V. Le, H.-M.-Q. Le, V. Y. Vu, T.-T. Tran, and V.-T. Pham, "Attention ConvMixer model and application for fish species classification," *EAI Endorsed Trans. Ind. Netw. Intell. Syst.*, vol. 10, no. 3, p. e2, Sep. 2023.

[31] X. Sun, J. Shi, L. Liu, J. Dong, C. Plant, X. Wang, and H. Zhou, "Transferring deep knowledge for object recognition in low-quality underwater videos," *Neurocomputing*, vol. 275, pp. 897–908, Jan. 2018, doi: 10.1016/j.neucom.2017.09.044.

[32] Y. Gong, S. Gu, and L. Guan, "Fine-grained classification network for fish based on the attention mechanism and EfficientNet," in *Proc. 4th Int. Conf. Robot., Intell. Control Artif. Intell.*, Dec. 2022, pp. 1010–1015.

[33] J. Yang, M. Cai, X. Yang, and Z. Zhou, "Underwater image classification algorithm based on convolutional neural network and optimized extreme learning machine," *J. Mar. Sci. Eng.*, vol. 10, no. 12, p. 1841, Dec. 2022.

[34] Y. Jiang, Y. Zhang, Y. Wang, Q. Guo, M. Zhao, and H. Qin, "Efficient vision transformer with token-selective and merging strategies for autonomous underwater vehicles," *IEEE Internet Things J.*, 2024.

[35] H. Qu, G.-G. Wang, Y. Li, X. Qi, and M. Zhang, "ConvFishNet: An efficient backbone for fish classification from composited underwater images," *Inf. Sci.*, vol. 679, Sep. 2024, Art. no. 121078, doi: 10.1016/j.ins.2024.121078.

[36] X. Xu, W. Li, and Q. Duan, "Transfer learning and SE-ResNet152 networks-based for small-scale unbalanced fish species identification," *Comput. Electron. Agricult.*, vol. 180, Jan. 2021, Art. no. 105878, doi: 10.1016/j.compag.2020.105878.

[37] N. S. Abinaya, D. Susan, and R. Kumar, "Naive Bayesian fusion based deep learning networks for multisegmented classification of fishes in aquaculture industries," *Ecological Informat.*, vol. 61, Mar. 2021, Art. no. 101248, doi: 10.1016/j.ecoinf.2021.101248.

[38] J. Deka, S. Laskar, and B. Baklial, "Automated freshwater fish species classification using deep CNN," *J. Inst. Eng. (India), Ser. B*, vol. 104, no. 3, pp. 603–621, Jun. 2023, doi: 10.1007/s40031-023-00883-2.

[39] B. Gong, K. Dai, J. Shao, L. Jing, and Y. Chen, "Fish-TViT: A novel fish species classification method in multi water areas based on transfer learning and vision transformer," *Heliyon*, vol. 9, no. 6, Jun. 2023, Art. no. e16761, doi: 10.1016/j.heliyon.2023.e16761.

[40] N. Yu, L. Huang, Z. Wei, W. Zhang, and B. Wang, "Weakly supervised fine-grained recognition based on spatial-channel aware attention filters," *Multimedia Tools Appl.*, vol. 80, no. 9, pp. 14409–14427, Apr. 2021.

[41] B. Li, Y. Liu, and Q. Duan, "T-KD: Two-tier knowledge distillation for a lightweight underwater fish species classification model," *Aquaculture Int.*, vol. 32, no. 3, pp. 3107–3128, Jun. 2024, doi: 10.1007/s10499-023-01314-1.

[42] D. L. Manikandan and S. M. Santhanam, "Underwater species classification using deep learning technique," vol. 34, no. 2, pp. 7–20, 2024.

[43] K. Dey, M. M. Hassan, M. M. Rana, and M. H. Hena, "Bangladeshi indigenous fish classification using convolutional neural networks," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, Jul. 2021, pp. 899–904.

[44] A. Al Smadi, A. Mehmood, A. Abugabah, E. Almekhlafi, and A. M. Al-Smadi, "Deep convolutional neural network-based system for fish classification," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 12, no. 2, pp. 2026–2039, Apr. 2022.

[45] L. I. Mampitiya, R. Nalmi, and N. Rathnayake, "Performance comparison of sea fish species classification using hybrid and supervised machine learning algorithms," in *Proc. Moratuwa Eng. Res. Conf. (MERCon)*, Jul. 2022, pp. 1–6.

[46] R. M. Aziz, R. Mahto, A. Das, S. U. Ahmed, P. Roy, S. Mallik, and A. Li, "CO-WOA: Novel optimization approach for deep learning classification of fish image," *Chem. Biodiversity*, vol. 20, no. 8, Aug. 2023, Art. no. e202201123.

[47] P. S. V. Reddy, M. V. Krishna, R. Aishwarya, R. Yogitha, and K. A. Kumar, "Fish species classifier for allergic people using CNN algorithm," in *Proc. 7th Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Feb. 2023, pp. 489–494.

[48] R. K. Kumar, D. Ashritha, D. Siddharth, N. Ashish, and M. Akshara, "Automated fish species detection," *Int. Res. J. Eng. Technol.*, pp. 500–506, Jul. 2023. [Online]. Available: www.irjet.net

[49] E. Prasetyo, N. Suciati, and C. Fatichah, "Multi-level residual network VGGNet for fish species classification," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5286–5295, Sep. 2022, doi: 10.1016/j.jksuci.2021.05.015.
[50] D. M. Azhar, N. Suciati, and C. Fatichah, "Analysis of effect of image augmentation with image enhancement on fish image classification using convolutional neural network," in *Proc. 14th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, Oct. 2023, pp. 129–134.
[51] D. F. Mujtaba and N. R. Mahapatra, "Fish species classification with data augmentation," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2021, pp. 1588–1593.
[52] D. F. Mujtaba and N. R. Mahapatra, "Convolutional neural networks for morphologically similar fish species identification," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2021, pp. 1553–1559.
[53] D. F. Mujtaba and N. R. Mahapatra, "Hierarchical deep learning models for identification of fish species," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2022, pp. 1558–1562.
[54] E. Ottaviani, M. Francescangeli, N. Gjeci, J. del Rio Fernandez, J. Aguzzi, and S. Marini, "Assessing the image concept drift at the OBSEA coastal underwater cabled observatory," *Frontiers Mar. Sci.*, vol. 9, pp. 1–13, Apr. 2022.
[55] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive Into Deep Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2023.
[56] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[57] K. S. Mehrab, M. Maruf, A. Daw, H. B. Manogaran, A. Neog, M. Khurana, B. Altintas, Y. Bakis, E. G. Campolongo, M. J. Thompson, X. Wang, H. Lapp, W.-L. Chao, P. M. Mabee, H. L. Bart, W. Dahdul, and A. Karpatne, "Fish-vista: A multi-purpose dataset for understanding & identification of traits from images," 2024, *arXiv:2407.08027*.
[58] R. J. M. Veiga, I. E. Ochoa, A. Belackova, L. Bentes, J. P. Silva, J. Semião, and J. M. F. Rodrigues, "Autonomous temporal pseudo-labeling for fish detection," *Appl. Sci.*, vol. 12, no. 12, p. 5910, Jun. 2022.

RICARDO J. M. VEIGA received the bachelor's degree in electric and electronic engineering, in 2017, and the master's degree from the University of Algarve (UAlg), in 2020, where he is currently pursuing the Ph.D. degree in computer engineering. He has contributed to various research projects, including as a Ph.D. Student Fellow in the MARREAL project, and as a Research Fellow in INSPECT, KTTSeaDrones, Mirar, and ACCESS4ALL. He has authored 26 publications, including conference papers, journal articles, book chapters, and patents. His notable works encompass areas, such as fish species image classification, self-supervised underwater fish detection, and augmented reality applications. Additionally, he has contributed to the academic community through peer review activities for IEEE Access. In his academic role, he has taught various subjects in Information and Communication Technology and Computer Science at UAlg, participating in workshops and seminars aimed at enhancing pedagogical practices and promoting innovation. His primary areas of research encompass computer vision, augmented reality, and machine learning, with a particular focus on applications in fish species image classification, underwater fish detection, and room layout estimation. His work also explores the development of innovative augmented reality frameworks and portable devices for enhanced user experiences in various environments.

JOÃO M. F. RODRIGUES received the Ph.D. degree in electronic and computer engineering, in 2008, and the Habilitation degree in electrical and computer engineering, in 2021. He is currently the Vice-Rector for Transfer, Innovation, and Digital University, University of Algarve (UAlg), Portugal, where he has also held positions, such as the Pro-Rector of Transfer and Innovation, the Director of the Department of Electrical Engineering, and the Director of the bachelor's degree in ICT and the master's degree in electrical and electronic engineering. Since 1994, he has taught in computer science and computer vision courses. He is a member of the Research Centre NOVA LINCS (Algarve), classified as Excellent by the Portuguese Foundation for Science and Technology. He had participated in more than 30 nationally or internationally funded scientific projects, several of which he served as coordinator. He is the co-author of more than 200 scientific publications. He has also organized or participated in the organization of special issues, tracks, workshops, and international scientific conferences in the fields of computer science and computer vision. His main interests include computer vision, human–computer interaction, human-centered artificial intelligence, and affective computing. He is a member of the editorial board of several international journals.