Fish-Vista: A Multi-Purpose Dataset for Understanding & Identification of Traits from Images

Kazi Sajeed Mehrab[1]†, M. Maruf[1], Arka Daw[3], Abhilash Neog[1], Harish Babu Manogaran[1], Mridul Khurana[1], Zhenyang Feng[2], Bahadir Altintas[6], Yasin Bakis[6], Elizabeth G Campolongo[2], Matthew J Thompson[2], Xiaojun Wang[6], Hilmar Lapp[5], Tanya Berger-Wolf[2], Paula Mabee[7], Henry Bart[6], Wei-Lun Chao[2], Wasila M Dahdul[4], Anuj Karpatne[1]†

[1]Virginia Tech, [2]The Ohio State University, [3]Oak Ridge National Laboratory,
[4]University of California, Irvine, [5]Duke University, [6]Tulane University, [7]Battelle

Abstract

We introduce Fish-Visual Trait Analysis (Fish-Vista), the first organismal image dataset designed for the analysis of visual traits of aquatic species directly from images using machine learning and computer vision methods. Fish-Vista contains 69,269 annotated images spanning 4,316 fish species, curated and organized to serve three downstream tasks: species classification, trait identification, and trait segmentation. Our work makes two key contributions. First, we provide a fully reproducible data processing pipeline to process fish images sourced from various museum collections, contributing to the advancement of AI in biodiversity science. We annotate the images with carefully curated labels from biological databases and manual annotations to create an AI-ready dataset of visual traits. Second, our work offers fertile grounds for researchers to develop novel methods for a variety of problems in computer vision such as handling long-tailed distributions, out-of-distribution generalization, learning with weak labels, explainable AI, and segmenting small objects. Dataset and code for Fish-Vista are available at https://github.com/Imageomics/Fish-Vista

1. Introduction

In much the same way as large-scale general-purpose datasets in computer vision (CV) such as ImageNet [18] have fueled the rise of deep learning for mainstream CV, the growing deluge of image datasets in organismal biology [10, 24, 25, 31, 44, 53, 60] are poised to enable similar revolutions in the field of "AI for biodiversity science" [21, 55]. Images are increasingly being considered as the "currency" for documenting the vast array of biodiverse organisms on our planet, with repositories containing millions of images of biological specimens collected by scientists in field museums or captured by drones, camera traps, or tourists posting photos on social media. This provides opportunities for CV research in biodiversity applications such as classifying species with fine-grained differences, and segmenting the entire body of organisms in natural habitat images with complex backgrounds.

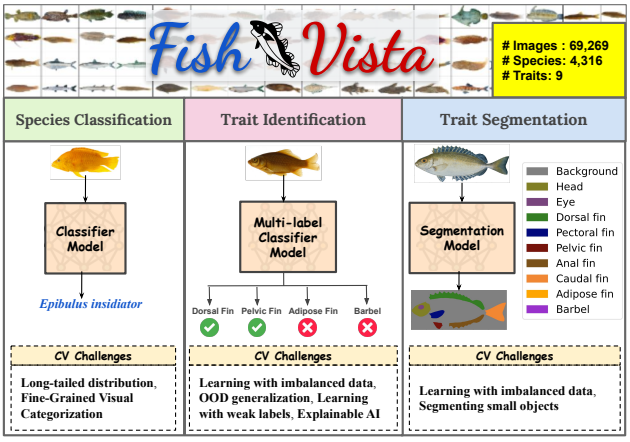

Figure 1. Overview of Fish-Vista tasks analyzing visual traits of fishes while exposing computer vision challenges.

While these applications serve critical use-cases, a scientific problem that has been relatively ignored in previous works is to discover characteristics of organisms, or *traits* (e.g., beak color, stripe pattern, and fin curvature), directly from images. Traits are the building blocks of knowledge in biodiversity science that help in discriminating between species and understanding how organisms evolve and adapt to their environment [28]. While some traits are behavioral, physiological, or related to the internal anatomy of organisms, in this work we focus on traits that are exter-

†{ksmehrab, karpatne}@vt.edu

nally visible in images, termed *visual traits*. Detecting visual traits and localizing their presence from large collections of biodiversity images offers novel opportunities for CV research to advance our understanding of the impacts of climate change on morphological features of organisms [26], and exploring the genetic and evolutionary underpinnings of their variations [28].

Current CV datasets in biodiversity science suffer from two critical gaps that limit their applicability for analyzing visual traits. *First*, most biodiversity datasets do not include trait-level annotations as they only focus on the task of species classification. Note that identifying and documenting the species of an organism from its image is relatively much easier than annotating all of its visual traits (either at the image-level or pixel-level), which often requires expert knowledge and labor-intensive manual processing. As a result, even though some datasets provide segmentation annotations of the *entire body* of organisms [42, 50, 60] (which are easier to generate using models such as the Segment Anything Model or SAM [34]), they do not provide annotations of fine-grained visual traits that are smaller in size and difficult to delineate. *Second*, most biodiversity datasets contain images taken in natural habitats [23, 50] and lack images taken in controlled environments with uniform backgrounds, necessary for the analysis of fine-grained visual traits. The challenge in using natural habitat images for visual trait analysis is that the presence of complex backgrounds such as dense foliage or underwater elements in poor-lighting environments can occlude and obfuscate visual traits that are already hard to detect, particularly for images of aquatic species taken underwater. Additionally, models trained on natural habitat images may learn to predict traits based on background patterns found in the habitats of certain species, introducing unintentional biases in the localization of visual traits.

To address these gaps, we introduce **Fish**-**Vis**ual **T**rait **A**nalysis (Fish-Vista), the first organismal dataset designed for the analysis of visual traits of fishes directly from images. Fish-Vista contains 69,269 annotated images spanning 4,316 fish species, curated and organized to serve three downstream tasks: species classification, trait identification, and trait segmentation (see Figure 1). Our work makes two key contributions to the field of CV for biodiversity science. *First*, we provide a fully reproducible data processing pipeline to process images sourced from various museum collections including GLIN [4], IDigBio [7], and MorphBank [1] and create an "AI-ready" dataset of visual traits, a novel concept in AI for biodiversity science. We annotate these images with carefully curated labels obtained from biological databases as well as manual annotations. *Second*, our work offer fertile grounds for novel CV research in a variety of problems such as handling *long-tailed distributions* (across all three downstream tasks), *out-of-distribution generalization* (for trait identification), *learning with weak labels* (for trait identification), *explainable AI* (for trait identification), and *segmenting small objects* (for trait segmentation). We benchmark the performance of state-of-the-art (SOTA) baseline methods on Fish-Vista tasks to expose current gaps and to motivate future research for answering biological questions relevant for the analysis of visual traits of organisms from images.

2. Related Works

Table 1 provides an overview of biodiversity image datasets that have been published in the last two decades covering diverse categories of organisms such as birds, cats, dogs, and fishes. While many of these datasets focus on the tasks of species classification and fine-grained visual categorization (FGVC) [64] (i.e., differentiating closely related species based on subtle visual differences), they mostly do not include annotations of visual traits either at the level of species or images (referred to as *visual trait information* in Table 1). While some datasets such as CUB [60], Oxford Pets [42], Ulucan et al. [56], and DeepFish [50] contain segmentation annotations of the entire body of organisms (*full-body segmentation*), they do not provide pixel-level annotations of individual traits that are fine-grained and smaller in size (*visual trait segmentation*). There are also datasets such as CUB [60], NABirds [57], and FishBase [23] that contain trait information at the level of images or species, but do not include trait segmentation annotations.

Another common feature in most biodiversity datasets is their focus on natural habitat images. For example, several image datasets feature fishes in their natural underwater habitats [22, 23, 50, 58]. While they are important for monitoring species populations out in the wild, they are not conducive to the analysis of visual traits of organisms because of their lack of clarity, and occlusions and obfuscations of visual traits. In contrast, museum collection images are taken in controlled environments that are easier to process (e.g., remove background and other imaging artifacts) and use for analyzing traits. While some datasets like QUT Fish [10] and Ulucan et al. [56] feature images in controlled internal environments, their backgrounds can still vary (i.e., non-uniform backgrounds). They are also limited in their number of images and species diversity. FishShapes [45] provides numeric data of the lengths of various fish parts, but does not provide image-level or pixel-level traits. Another notable dataset for studying fish traits is FishBase [23], comprising 64K images spanning 35K species. However, FishBase is limited in the number of images available per species, which poses a challenge for training AI models. FishNet [30] combines images from the iNaturalist fish dataset [58] with functional traits from FishBase [23] (such as ecological/habitat information of species). However, functional traits are not localizable in

Dataset	Organism	# Species	# Images	Full-body Segmentation	Visual Trait Information	Visual Trait Segmentation	Background
CUB-200-2011 [60]	Birds	200	11,788	✓	✓	x	Natural Habitat
Birds 525 [44]	Birds	525	89,885	x	x	x	Natural Habitat
NABirds [57]	Birds	555	48,562	x	✓	x	Natural Habitat
Stanford dogs [31]	Dogs	120	20,580	x	x	x	Natural Habitat
Oxford Pet [42]	Cats, Dogs	37	7,349	✓	x	x	Natural Habitat
FathomNet [29]	Marine Species	2244	84,454	x	x	x	Natural Habitat
Ulucan et al. [56]	Fish	9	9,000	✓	x	x	Controlled
QUT Fish [10]	Fish	468	3,960	x	x	x	Controlled/Natural Habitat
DeepFish [50]	Fish	NA	39,766	✓	x	x	Natural Habitat
Fish4Knowledge [22]	Fish	23	27,370	x	x	x	Natural Habitat
FishBase [23]	Fish	35,600	64,000	x	✓	x	Natural Habitat
iNaturalist-2021-Fish [58]	Fish	183	46,996	x	x	x	Natural Habitat
FishNet [30]	Fish	17,357	94,778	x	x	x	Natural Habitat
Fish-Vista (Ours)	Fish	4,316	69,269	✓	✓	✓	Controlled + Uniform

Table 1. Summary of commonly used fine-grained biodiversity datasets comprising images of organisms.

images, and hence fall outside our focus on visual traits. In contrast to all previous works, our proposed dataset, Fish-Vista, provides high-quality fish images with trait annotations at species, image, and pixel levels, and with controlled and uniform backgrounds from museum collections, covering a large number of images across a wide range of species, as shown in Table 1.

3. Fish-Vista Dataset

3.1. Why Fish-Vista?

Fish-Vista fixes a critical gap in current benchmark datasets available in AI for biodiversity science by bridging high-quality images cleaned and curated from diverse museum collections with labels of visual traits obtained through expert annotations and knowledge-bases. Along with enabling a range of trait-related questions in the field of biodiversity science, a primary motivation behind Fish-Vista is to expose novel problem formulations and research challenges in CV tasks involving visual traits. For example, while there has been considerable work in FGVC for species classification, the connection between the subtle differences in species discovered by AI models and visual traits known to biologists has still not been established. We hope that by focusing on visual traits, our work advances the field of CV to focus on the explainability of fine-grained features that are localized in images and grounded in knowledge of biological traits.

3.2. Data Sources used in Fish-Vista

We consider museum collections of fish images publicly available at GLIN [3–6, 8, 9, 16], IDigBio [7], and Morphbank [1] databases. We acquired these images along with their associated metadata including species names and licensing information from the FishAIR repository [2]. In total, we collected 56,481 images from GLIN, 41,505 from IDigBio, and 9,000 from MorphBank.

3.3. Data Processing Pipeline

There are two key challenges with FishAIR images that we need to address: (1) museum images contain several visual elements such as rulers and tags apart from fish specimens that need to be cropped, and (2) there are many noisy images in museum collections including hand-written notes and radiographic images that need to be dropped. Figure 2 shows a schematic of our processing pipeline to address these challenges comprising of the following five steps.

1. Removing Duplicates: Since museum collections sometimes contain duplicate images stored under different filenames, we remove duplicate images with same MD5-checksum, to avoid data leakage in training and test splits.

2. Quality Metadata-based Filtering: For a subset of the raw images ($\approx 30k$), we obtained manually annotated *quality metadata* from FishAIR that includes information about the visibility of all parts of a specimen and the orientation of the fish (e.g., side-view or top-view). We use this data to filter images where all visual traits are not visible.

3. Filtering Noisy Species Names: Scientific species names of FishAIR images sometimes contain inaccuracies like typographical errors or synonymous names. To mitigate this, we exclude entries with species names that are not valid strings, such as "*gen. sp.*". We utilize the Open Tree Taxonomy (OTT) [40] to correct typographical errors and standardize synonyms to their canonical forms, ensuring consistent categorization of species names.

4. Detecting and Cropping Fish Bounding Boxes: We use Grounding DINO [36], a SOTA zero-shot object detection model, to detect and crop tight bounding boxes around fish specimens. We discard bounding boxes with dimensions smaller than 224 pixels to avoid low-resolution images.

5. Removing Background using SAM: The backgrounds of fish bounding boxes often contain features unique to specific species or museum collections, introducing biases in the data that are not useful for analyzing visual traits. To

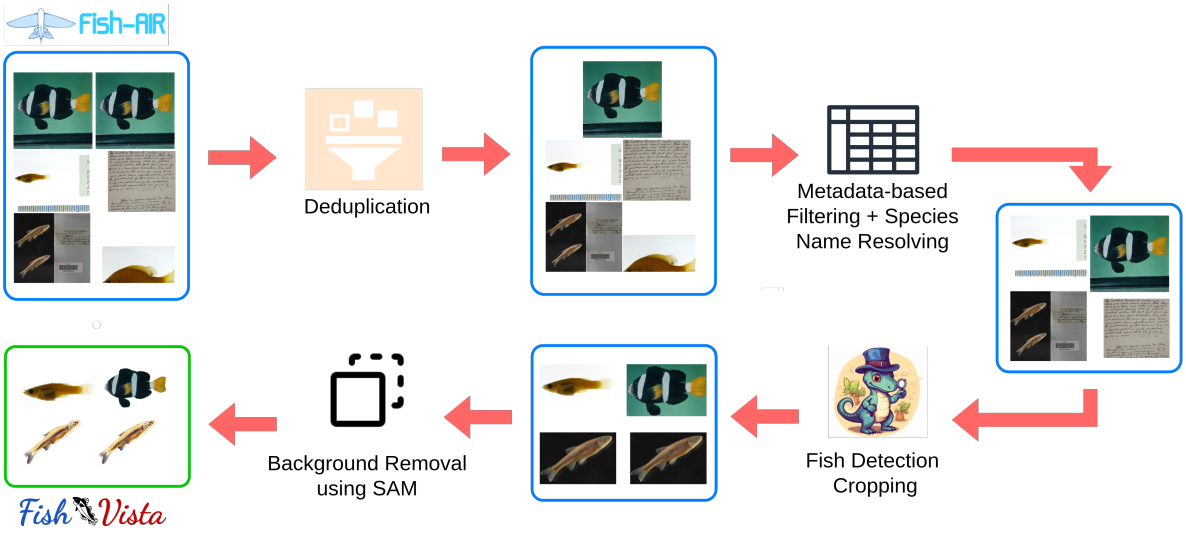

Figure 2. An overview of the data processing pipeline used to process raw museum images to obtain images in Fish-Vista.

avoid this, we use the Segment Anything Model (SAM) [34] to segment the entire body of a fish specimen from its bounding box and use a uniform white background.

Further details of the processing pipeline along with quantitative and qualitative validations for *Step 4* and *Step 5* are provided in Appendix C. Following the data processing, we obtain $\approx 100K$ images spanning $\approx 10K$ species that we further refine and annotate to create datasets for the three downstream tasks.

3.4. Fish-Vista Tasks

Figure 3 shows an overview of the process followed for creating data partitions of the three Fish-Vista tasks along with key statistics. We describe each task along with their associated key CV challenges in the following.

3.4.1. Fine-grained Species Classification

Species classification involves categorizing images to their respective species by distinguishing fine-grained visual traits. One of the key challenges in species classification with Fish-Vista is the extreme long-tailed nature of image count distributions across species classes. We adopt several steps to prepare data partitions for species classification while accounting for the long-tailed distribution of species classes. We first remove species that have less than 4 images per class, to ensure sufficient number of images for training, testing, and validation. The remaining species still suffer from a high degree of class imbalance as some species contain up to 2K images while many others have less than 10 images. To further remove rare species that do not have representative high-quality images, we manually inspect the visual quality of a randomly sampled subset of 15% images for each species. Species with predominantly low-quality images or those lacking clear visual traits are dropped from the dataset (see Appendix D for additional details). This results in the final *FV-Classification* dataset, which contains 56,360 images spanning 1,758 species.

We construct train, test, and validation splits using stratified sampling across every species with splitting fractions of 75%, 15%, and 10% respectively. We manually inspect every image in the test set to ensure that they are of high quality. To address the dataset's highly imbalanced long-tailed distribution, we subcategorize the 1,758 species in FV-Classification into four groups based on their training image counts per species: *Majority* (500 or more images), *Neutral* (100-499 images), *Minority* (10-99 images), and *Ultra-Rare* (fewer than 10 images). Figure 3 provides statistics on the four subcategories of FV-Classification and their distributions of training, test, and validation images. Note that while we only have 20 majority species, we have 1,342 Ultra-rare species, demonstrating the highly imbalanced long-tailed nature of FV-Classification. Additional details about the data splits and manual test set filtering are provided in Appendix F.1.

Key CV challenges: Given the large number of species that have varying evolutionary and anatomical similarities, classification models must differentiate subtle, fine-grained visual differences among highly similar species, making this a challenging FGVC task. The dataset's long-tailed distribution further adds the challenge of performing highly imbalanced classification.

3.4.2. Trait Identification

Trait identification is the task of predicting the presence or absence of visual traits from an image (Figure 1). There are four key points to consider for this task. *First*, predicting presence/absence of all possible traits in images is unnecessary; we only need to predict traits that vary across

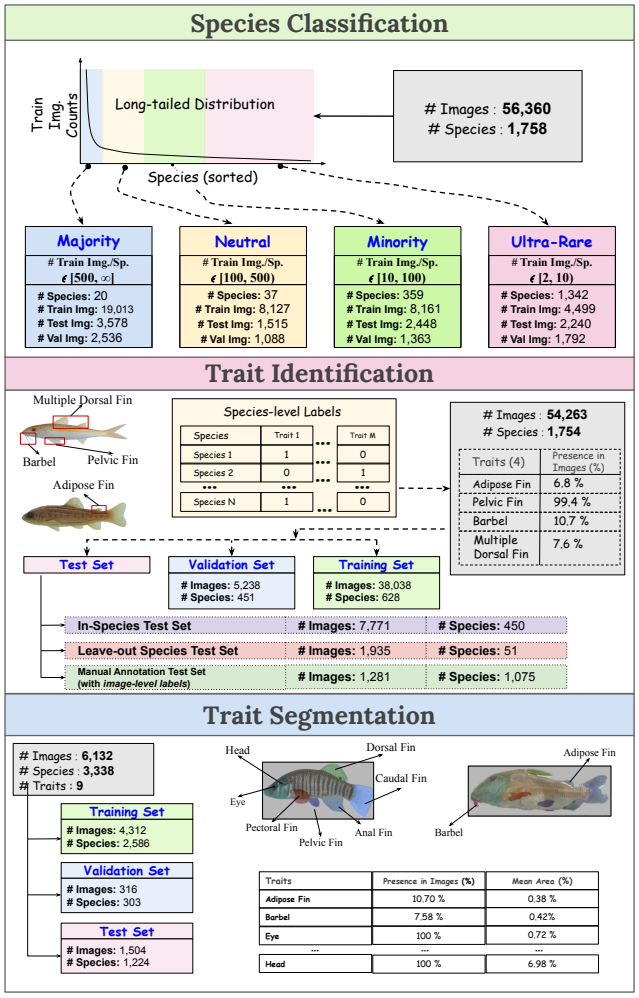

Figure 3. An overview of the key statistics of Fish-Vista Dataset.

species and are considered biologically *interesting*. Traits deemed *interesting* are often "rare" traits – those that are present or absent in only a few species. For example, the presence of eyes, which is nearly universal across all fish species, is neither informative nor biologically significant to predict, whereas rare traits, such as the presence of an adipose fin (see Figure 3 for definition), offers more scientific value. *Second*, manually annotating thousands of images for trait presence/absence requires biological expertise, is time-consuming, and difficult to scale. *Third*, the effectiveness of trait identification models needs to be evaluated on out-of-distribution data containing *species never seen during training*. This would ensure that the models learn to generalize based on the visual appearance of traits, rather than memorizing species names and predicting traits known to be associated with every species. *Fourth*, we should also evaluate the ability of models to accurately identify or *visually localize* the traits within the image while predicting their presence or absence.

We create the trait identification dataset with the aforementioned key points in mind involving a number of steps as outlined in Figure 3. *First*, we select four scientifically significant traits that vary across species – adipose fin, pelvic fin, barbel, and multiple dorsal fins. From the *Presence in Images (%)* column, we see that while some of these traits are present over a large percentage of images (e.g., pelvic fin), there are traits that are rare such as adipose fin and multiple dorsal fin that are present only 6.8% and 7.6%, respectively. *Second*, instead of annotating each image manually, we gather species-level trait labels from the Phenoscape KnowledgeBase (KB) [38] and FishBase [23]. Similar to FV-Classification, we discard species containing predominantly low-quality images through manual observation (Appendix D). This results in the retrieval of trait information for 682 species that are mapped to their corresponding images (about 53K). Note that since FishVista images are manually filtered to include complete fish specimens with all traits visible, species-level labels provide a reasonable basis for image-level trait identification. The use of species-level labels also introduces the challenge of learning with *weak labels*, since traits in images are identified based on coarse-grained labels at the species level.

We split the $\approx$ 53K images containing trait-level information into training, validation, and test sets with the goal of evaluating out-of-distribution (OOD) generalization (that is, evaluating on species never encountered during training). Toward this goal, we create a *leave-out-species* set by holding out 51 species (1,935 images) that are only used for testing. Images from the remaining species are split into training, validation, and an *in-species* test set stratified by trait labels. We ensure that every image in the *in-species* test set comes from a species that has been seen during training. Additionally, to further evaluate the generalization performance of trait identification across unseen species, we create a *manual-annotation* test set consisting of 1,281 manually annotated images across 1,075 species (that have no overlap with the training set species), labeled by expert biologists for the presence or absence of the four target traits. The *manual-annotation* set differs from the *leave-out-species* set in three ways. First, it contains manual annotations of trait labels at the level of individual images rather than at the species level. Second, it contains a much larger diversity of species compared to the *leave-out-species* set. Third, the *manual-annotation* set also includes pixel-level segmentation annotations for every image, enabling the evaluation of *trait localization* within the body of fish images. Adding the *manual-annotation* set results in the complete identification dataset, *FV-Id*, with a total of 54,263 images across 1,754 species. We manually inspect all test sets to ensure quality. Key statistics of FV-Id are summarized in Figure 3 and details are in Appendix F.2.

Key CV Challenges: We need to test the generalization performance of models on both species seen during

training (in-distribution performance), and unseen species (out-of-distribution performance). The traits are also highly imbalanced. By training models using species-level labels and evaluating them on image-level annotations using the *manual-annotation* set, FV-Id is enabling the study of learning with weak labels. Additionally, segmentation annotations in the *manual-annotation* set enable analysis of model explainability, i.e., whether models attend to the correct regions on the image when predicting trait presence.

3.4.3. Trait Segmentation

Going beyond trait identification at the image level, we introduce the task of trait segmentation, where the goal is to delineate traits within the fish images (Figure 1). We focus on segmenting nine visible traits on fish bodies as shown in Figure 3. It is worth noting that while certain traits, such as the *eye*, may be uninformative for image-level presence/absence prediction in trait identification, their localization on images is still significant. We create the segmentation dataset, *FV-Segmentation*, by manually labeling the 9 traits across 6,132 images. The annotation process is conducted by expert biologists by utilizing the CVAT tool [15]. Due to the labor-intensive nature of the annotation process and the need for biological expertise, the segmentation dataset is smaller than its classification and identification counterparts. To enhance models' ability to generalize from a limited number of images, we include images from a highly diverse set of 3,338 species in FV-Segmentation. Key statistics of the dataset are provided in Figure 3. We can see that certain traits, like the adipose fin and barbel, occupy very small pixel areas on average (0.38 % and 0.42% respectively), while also being present very rarely (10.7% and 7.58% respectively), making it a challenging dataset for *segmenting small and rare objects*. Additional details in the creation of FV-Segmentation are provided in Appendix F.3.

Key CV Challenges: The trait segmentation task presents several unique challenges. *First*, because of the relatively smaller size of the dataset, the segmentation models must be adept at learning to generalize using limited number of labels. Second, the various fins of the fish can appear visually similar in shape and texture. This means models must rely on positional cues alone to distinguish between these traits, which can cause misclassifications. This is further complicated by anatomical variations across diverse species. For example, the adipose and dorsal fins can look similar, and also appear in similar positions across various species. Third, some traits are very small, posing the well-known challenge of small object segmentation. For example, barbels are whisker-like projections near the fish's mouth that occupy a small area. Finally, as with identification, the presence of certain traits is very rare, creating high imbalance. For example, adipose fin and barbel, which are both small traits, are also very rare, combining the challenges of small object segmentation with data imbalance.

Type	Model	F1	Major Acc.	Neutral Acc.	Minor Acc.	Ultra-R Acc.
CNN	**VGG-19** [51]	49.7	93.5	83.0	74.2	45.9
	ResNeXt-50 [63]	44.4	91.4	78.3	69.8	39.1
	RegNetY-4G [47]	43.7	89.8	77.4	68.5	38.5
ViT	**ViT-B16** [19]	48.3	88.7	82.3	73.3	43.4
	Swin-B-22k [37]	55.1	92.6	86.2	79.6	50.4
	CvT-13 [62]	49.3	92.0	83.3	73.5	44.7
	MaxViT-T [54]	57.8	94.4	86.7	81.4	53.9
	PVT-v2-b0 [61]	51.0	92.0	83.4	75.7	45.8
Foundation Models	**BioCLIP-ZS** [52]	4.6	1.1	1.4	10.2	5.6
	CLIP-ZS [46]	0.1	0.0	0.3	0.4	0.2
	BioCLIP-LP [52]	38.2	75.5	65.2	61.3	31.1
	CLIP-LP [46]	25.4	55.9	49.8	46.7	20.9
	DINOv2-LP [41]	53.1	89.9	78.04	76.04	47.02
FGVC	**INTR** [43]	6.1	92.2	73.2	22.6	0.62
	TransFG [27]	50.3	94.5	86.6	75.5	45.3

Table 2. Comparison of species classification performance (in %). Results are color-coded as Best, Second best, Worst, Second worst (excluding zero-shot (ZS) methods).

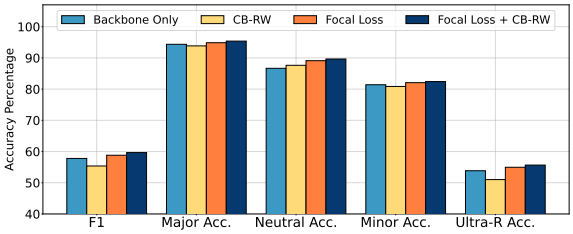

Figure 4. Comparison of classification performance of different imbalanced classification methods on MaxViT-T

4. Experiments and Results

We compare results of baseline methods on the three Fish-Vista tasks in the following. Implementation details of all methods are provided in Appendix G. A cross-cutting objective of our experiments is to discover insights and highlight key challenges that current CV methods encounter in Fish-Vista, rather than determining the best-performing method for each task.

4.1. Species Classification

We evaluate a wide range of approaches for species classification including CNN-based and vision transformer-based (ViT) backbones, zero-shot (ZS) and linear probing (LP) methods on vision foundation models, fine-grained categorization (FGVC) methods, and techniques for handling class-imbalance. We report the overall macro-averaged F1-score and mean accuracy for each species subcategory to assess performance across the imbalanced distribution of species. Key results are shown in Table 2, with additional results in Appendix H.1. As expected, most methods perform well on *Majority* species ($\approx 90\%$) and *Neutral*

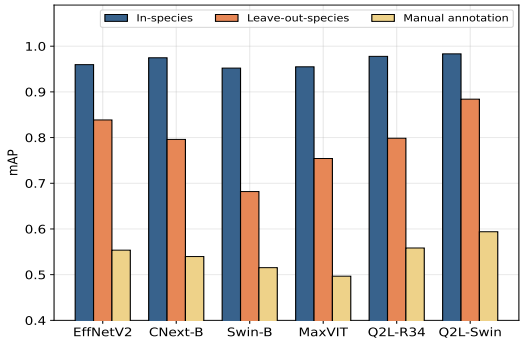

Figure 5. Trait identification performance of different multi-label classification methods. Details of models are in Appendix G.2.

species ($\approx$ 80%), but the performance drops significantly on rare categories, with *Ultra-rare* species reaching only about 50% accuracy.

We also evaluated the zero-shot performance of CLIP and BioCLIP, a foundation model for biodiversity images. Both models performed near random on Fish-Vista, with BioCLIP outperforming CLIP, particularly on *Rare* species, but still with low accuracy. Next, we use pre-trained features from CLIP, BioCLIP, and DINOv2 to perform linear probing by training a single classification layer. BioCLIP features outperform CLIP, while DINOv2 performs significantly better than both. However, it still performs worse than the best-performing backbone models. These results highlight the limitations of current vision foundation models – including those trained on biodiversity images like BioCLIP – in capturing fine-grained species variation, underscoring the need for specialized approaches.

We also experiment using two FGVC methods – INTR and TransFG, and observe similar patterns. Both models struggle with the minority and ultra-rare species, with INTR obtaining the worst accuracies on these categories. This indicates that FGVC methods may struggle to handle the long-tailed distribution of our dataset effectively.

Given the dataset's highly imbalanced nature, we evaluate the impact of well-known imbalance-handling techniques – class-balanced re-weighting (CB-RW) [14] and focal loss [35]. Results using our top-performing backbone, MaxViT, are shown in Figure 4. We observe that even after combining both these techniques, the improvement in performance is only marginal, highlighting the challenge of long-tailed distribution in our FV-classification dataset.

Summary of Insights: Standard classification techniques, including fine-grained and class-imbalance methods, may not perform optimally on Fish-Vista – especially for the rare species that constitute the majority of the species in the dataset – due to its challenging long-tailed distribution and fine-grained categorization requirements.

Q2L Backbone	# Attention Heads	IoU ($\times$100) *Adipose*	*Pelvic*	*Barbel*	*Dorsal*	mIOU ($\times$100)
ResNet	1	0.014	1.482	0.0008	1.731	0.81
ResNet	4	0.003	0.963	0.0067	0.908	0.47
SWIN-B	1	0.007	1.447	0.005	1.945	0.85
SWIN-B	4	0.048	1.844	0.002	0.967	0.72

Table 3. IoU of Query2Label attention maps for each trait in the manual annotation set. IoUs and mIoUs are amplified 100 times.

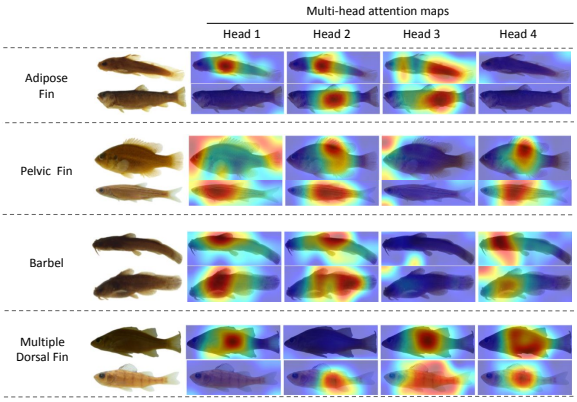

Figure 6. Attention maps from the Query2Label-SWIN model corresponding to the four traits.

4.2. Trait Identification

We train an extensive range of baseline models using a multi-label classification objective to predict the presence or absence of the four traits in the FV-Id dataset. Figure 5 compares the Mean Average Precision (mAP) of top-performing models across all three test sets. Trait-wise results and additional metrics for all models are provided in Appendix H.2.

As expected, our best-performing models achieve high accuracy on the *in-species* test set, given that these species were included in training. However, performance drops substantially on the *leave-out-species* test set, and this decline is even more pronounced on the highly diverse *manual-annotated* test set. This indicates that existing methods struggle to generalize to traits on unseen species, which is a key requirement for the task of trait identification.

A crucial aspect of evaluating trait identification performance is to determine whether models can visually *attend* to the correct traits, i.e., can we achieve trait localization just using trait presence/absence labels on images? To assess this, we examine our top-performing identification model, Query2Label (Q2L) with SWIN backbone, using the model's transformer attention maps for each trait on the manual annotation dataset. We calculate the Intersection over Union (IoU) between ground-truth trait segmentations and the model's attention maps (Table 3) and visualize the maps for images where the model correctly predicts traits (Figure 6). Despite high accuracy, Query2Label demon-

Model	mIoU	Trait-wise IoU									
		BG	*Head*	*Eye*	*Dorsal*	*Pectoral*	*Pelvic*	*Anal*	*Caudal*	*Adipose*	*Barbel*
PSPNet [65]	73.8	94.6	84.3	77.7	85.1	67.1	80.5	83.0	88.7	56.9	20.1
DeepLabV3 [11]	74.9	95.0	85.4	78.1	86.0	71.2	83.0	85.3	88.8	58.1	18.2
DeepLabV3Plus [12]	77.0	95.4	86.0	79.1	88.1	71.0	84.7	86.2	89.9	66.1	23.7
UNet [49]	77.5	95.6	86.2	79.9	88.2	70.6	84.8	86.6	90.8	69.7	22.3
Semantic FPN [33]	77.6	95.5	86.1	79.2	88.2	71.4	85.0	86.4	90.3	67.6	26.3
Mask2Former [13]	81.6	95.5	86.4	79.1	87.6	74.2	76.1	84.7	88.8	59.6	0.0
YOLOv8 [59]	83.1	96.8	84.5	83.1	88.0	78.0	77.5	85.6	89.6	66.8	26.7
Molmo+SAM (ZS) [17, 48]	39.1	85.3	37.4	29.8	50.3	29.7	38.5	36.1	83.4	0.4	0.6

Table 4. Performance (in %) of seven mainstream segmentation models on the Segmentation dataset, along with a zero-shot architecture. Results are color-coded as Best, Second best, Worst & Second worst (excluding zero-shot method Molmo+SAM).

strates extremely low mIoUs and scattered attention maps, indicating a failure to attend to the correct traits.

Summary of Insights: Existing methods struggle to generalize traits on unseen species. Additionally, existing methods may predict presence/absence of traits with decent performance at the image level but may not focus on relevant image regions, lacking explainability and spatial awareness necessary for localization.

4.3. Trait Segmentation

We evaluate several baseline methods for image segmentation for this task, including semantic segmentation architectures, instance segmentation models, and a zero-shot segmentation method. Table 4 presents the overall mIoU and individual trait-wise IoUs for each method. Traits that have higher presence and occupy larger areas (e.g., head, dorsal fin, and caudal fin) generally achieve high IoUs of over 85%. However, performance drops significantly for smaller traits like the eye, and rarer traits like the adipose fin and barbel. Traits that are located over the body, such as the pectoral fin, are also harder to localize.

All methods struggle particularly with the adipose fin and barbel. Notably, Mask2Former entirely fails to detect the barbel. This difficulty is likely due to both traits being rare (low presence) and occupying minimal area (see Figure 3). Further inspection of the confusion matrix (Appendix H.3) reveals that the barbel, located beneath the head, is frequently misclassified as the head, while the adipose fin, which is small and often near the dorsal fin, is misclassified as either the background or the dorsal fin. These results underscore the challenges that current methods face in accurately segmenting small, rare, and fine-grained traits.

Finally, we investigate the zero-shot segmentation capabilities of the Segment Anything Model (SAM-v2) [48], coupled with a large vision-language model (LVLM), Molmo [17]. We direct Molmo to identify trait location points in images through textual prompts. The points generated by Molmo serve as input prompts to SAM-v2, which relies on spatial prompts (e.g., points) to generate segmentation masks of the traits (details in Appendix G.3.1). While we did not expect high performance, results demonstrate promising mIoU on traits like the dorsal and caudal fins.

Summary of Insights: Conventional segmentation methods face significant challenges in localizing small, rare and fine-grained traits in Fish-Vista. Moreover, large foundational models like LVLMs and SAM have the potential to localize scientifically relevant visual traits.

5. Limitations of Fish-Vista and Future Work

While the processing of Fish-Vista includes a range of automated and manual filtering steps, there may still be some images that are noisy and do not clearly exhibit visual traits, such as those with deformed fins (see Appendix D). Additionally, for the FV-Id dataset, while we assume that species-level labels of the presence or absence of traits are representative over all images of the species, this may not be true especially when certain traits in an image are occluded due to poor data quality. Finally, the segmentation dataset contains a relatively smaller number of annotated images compared to the other two tasks, which is due to the labor-intensive nature of generating pixel-level annotations.

There are several directions of future work with Fish-Vista. First, Fish-Vista can serve as a valuable resource to train foundation models for biology, similar to BioCLIP [52], that can be explored in future research. Second, future research can explore ways of incorporating structured biological knowledge (e.g., the tree of life or taxonomic grouping of species) in the training of models based on Fish-Vista, to ground the learning of visual traits in scientifically meaningful concepts. This would scale previous works in the emerging field of knowledge-guided ML (KGML) for biodiversity science [20, 32, 39] over a large and diverse dataset that enables novel hypothesis generation and discovery of visual traits directly from images.

Acknowledgments

This research is supported by grants from the National Science Foundation (NSF) for the HDR Imageomics Institute (OAC-2118240).

We acknowledge the following members from Battelle for their time and effort in manual annotations: Suvi Birch, James Boudreau, Casey Chen, Isa DuMond, Mina Esphahanian, Taylor Jarvis, Cesar Ortiz, Rebecca Osborne, Shelley Rider, Zachary Shappell, Alma Suarez and Jerry Tatum.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains, and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).

We are thankful for the support of computational resources provided by the Advanced Research Computing (ARC) Center at Virginia Tech and the Ohio Supercomputer Center.

References

[1] Morphbank: Biological imaging (https://www.morphbank.net/). *Florida State University, Department of Scientific Computing, Tallahassee, FL 32306-4026 USA*. 2, 3

[2] Multimedia of fish specimen and associated metadata. fishair. *Biology guided Neural Network. Tulane University Biodiversity Research Institute (https://fishair.org)*. 3

[3] Fmnh field museum of natural history (zoology) fish collection. *Field Museum. https://fmipt.fieldmuseum.org/ipt/resource?r=fmnh_fishes*. 3

[4] Great lakes invasives network project. *https://greatlakesinvasives.org/portal/index.php*. 2

[5] University of wisconsin-madison zoological museum - fish. *http://zoology.wisc.edu/uwzm/*.

[6] Ummz university of michigan museum of zoology, division of fishes. *https://ipt.lsa.umich.edu/resource?r=ummz_fish*. 3

[7] idigbio. *http://www.idigbio.org/portal*, 2020. 2, 3

[8] The illinois natural history survey's biological collections. *http://biocoll.inhs.illinois.edu/portal/index.php*, 2022. 3

[9] Jfbm bell atlas. *http://bellatlas.umn.edu/index.php.*, 2022. 3

[10] Kaneswaran Anantharajah, ZongYuan Ge, Christopher McCool, Simon Denman, Clinton B Fookes, Peter Corke, Dian W Tjondronegoro, and Sridha Sridharan. Local inter-session variability modelling for object classification. In *Winter Conference on Applications of Computer Vision (WACV), 2013 IEEE Conference on*, 2014. 1, 2, 3

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 8

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 8

[13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 8

[14] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 7

[15] CVAT.ai Corporation. Computer vision annotation tool (cvat) (v2.4.3), 2023. 6

[16] Johnson N Daly M. Ohio state university fish division (osum). *Museum of Biological Diversity, The Ohio State University. Occurrence dataset, https://doi.org/10.15468/subsl8*, 2018. 3

[17] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 8

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[20] Mohannad Elhamod, Mridul Khurana, Harish Babu Manogaran, Josef C Uyeda, Meghan A Balk, Wasila Dahdul, Yasin Bakis, Henry L Bart Jr, Paula M Mabee, Hilmar Lapp, et al. Discovering novel biological traits from images using phylogeny-guided neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3966–3978, 2023. 8

[21] Paul Fergus, Carl Chalmers, Steven Longmore, and Serge Wich. Harnessing artificial intelligence for wildlife conservation. *Conservation*, 4(4):685–702, 2024. 1

[22] Robert B. Fisher, Yun-Heh Chen-Burger, Daniela Giordano, Lynda Hardman, and Fang-Pang Lin, editors. *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*. Springer, 2016. 2, 3

[23] R. Froese and D. Pauly. Fishbase, 2024. World Wide Web electronic publication. Version 02/2024. 2, 3, 5

[24] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, Iuliia Zarubiieva, Joakim Bruslund Haurum, Scott Lowe, Jaclyn

McKeown, Chris Ho, Joschka McLeod, Yi-Yun Wei, et al. A step towards worldwide biodiversity assessment: The bioscan-1m insect dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[25] Zahra Gharaee, Scott C Lowe, ZeMing Gong, Pablo Millan Arias, Nicholas Pellegrino, Austin T Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Lila Kari, Dirk Steinke, et al. Bioscan-5m: A multimodal dataset for insect biodiversity. *arXiv preprint arXiv:2406.12723*, 2024. 1

[26] CT Graham and Chris Harrod. Implications of climate change for the fishes of the british isles. *Journal of Fish Biology*, 74(6):1143–1205, 2009. 2

[27] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 852–860, 2022. 6

[28] David Houle and Daniela M Rossoni. Complexity, evolvability, and the process of adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 53:137–159, 2022. 1, 2

[29] Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, et al. Fathomnet: A global image database for enabling artificial intelligence in the ocean. *Scientific reports*, 12(1):15914, 2022. 3

[30] Faizan Farooq Khan, Xiang Li, Andrew J Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20496–20506, 2023. 2, 3

[31] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings CVPR workshop on fine-grained visual categorization (FGVC)*, 2011. 1, 3

[32] Mridul Khurana, Arka Daw, M Maruf, Josef C Uyeda, Wasila Dahdul, Caleb Charpentier, Yasin Bakış, Henry L Bart Jr, Paula M Mabee, Hilmar Lapp, et al. Hierarchical conditioning of diffusion models using tree-of-life for studying species evolution. In *European Conference on Computer Vision*, pages 137–153. Springer, 2024. 8

[33] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 8

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 4

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[36] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[38] Paula M Mabee, Wasila M Dahdul, James P Balhoff, Hilmar Lapp, Prashanti Manda, Josef Uyeda, Todd Vision, and Monte Westerfield. Phenoscape: semantic analysis of organismal traits and genes yields insights in evolutionary biology. In *Application of Semantic Technology in Biodiversity Science*, pages 207–224. IOS Press, 2018. 5

[39] Harish Babu Manogaran, M Maruf, Arka Daw, Kazi Sajeed Mehrab, Caleb Patrick Charpentier, Josef C Uyeda, Wasila Dahdul, Matthew J Thompson, Elizabeth G Campolongo, Kaiya L Provost, et al. What do you see in common? learning hierarchical prototypes over tree-of-life to discover evolutionary traits. *arXiv preprint arXiv:2409.02335*, 2024. 8

[40] Open Tree of Life, Karen A. Cranston, Benjamin Redelings, Luna Luisa Sanchez Reyes, Jim Allman, Emily Jane McTavish, and Mark T. Holder. Open Tree of Life Taxonomy (3.2). Zenodo, 2019. 3

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[42] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505, 2012. 2, 3

[43] Dipanjyoti Paul, Arpita Chowdhury, Xinqi Xiong, Feng-Ju Chang, David Carlyn, Samuel Stevens, Kaiya L Provost, Anuj Karpatne, Bryan Carstens, Daniel Rubenstein, et al. A simple interpretable transformer for fine-grained image classification and analysis. *arXiv preprint arXiv:2311.04157*, 2023. 6

[44] Gerald Piosenka. Birds 525 species - image classification. 2023. 1, 3

[45] Samantha A Price, Sarah T Friedman, Katherine A Corn, Olivier Larouche, Kasey Brockelsby, Anna J Lee, Maya Nagaraj, Nick G Bertrand, Mailee Danao, Megan C Coyne, et al. Fishshapes v1: Functionally relevant measurements of teleost shape and size on three dimensions, 2022. 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[47] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 6

[48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 8

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 8

[50] Alzayat Saleh, Issam H Laradji, Dmitry A Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671, 2020. 2, 3

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[52] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. *arXiv preprint arXiv:2311.18803*, 2023. 6, 8

[53] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 1

[54] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 6

[55] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):1–15, 2022. 1

[56] Oguzhan Ulucan, Diclehan Karakaya, and Mehmet Turkan. A large-scale dataset for fish segmentation and classification. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE, 2020. 2, 3

[57] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 595–604, 2015. 2, 3

[58] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 2, 3

[59] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6. IEEE, 2024. 8

[60] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2, 3

[61] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 6

[62] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 6

[63] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[64] Yong Zhang, Weiwen Chen, and Ying Zang. Fine-grained vision categorization with vision transformer: A survey. In *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, pages 1910–1915. IEEE, 2022. 2

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 8