SHELLFISHNET: A DOMAIN-SPECIFIC BENCHMARK FOR VISUAL RECOGNITION OF MARINE MOLLUSCS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable visual recognition of shellfish is important for ecological monitoring and aquaculture, but real-world imagery exhibits substantial variation in acquisition conditions, appearance, and posture. Existing marine vision resources cover diverse taxa and imaging settings, but fine-grained mollusc recognition remains underexplored in a group-disjoint benchmark that jointly evaluates class prediction, representation-space retrieval, and condition sensitivity. To address this gap, we introduce ShellfishNet, a provenance-aware, group-disjoint benchmark of 40,130 images from 100 marine mollusc classes, organized into 30,879 acquisition or observation groups. Rather than treating recognition as a single clean-accuracy leaderboard, ShellfishNet evaluates model configurations across classification, image-to-image retrieval, taxonomic error structure, and controlled condition sensitivity. We benchmark representative convolutional, transformer-based, hybrid and efficient, self-supervised, and vision-language models on classification and retrieval. Separately, we evaluate open image-captioning and vision-language models on the captioning subset and assess selected fixed checkpoints under controlled generic and marine-inspired synthetic transformations. Across the benchmark, image-level and class-aware classification largely agree, whereas retrieval provides complementary model rankings and taxonomy projection reveals that a non-negligible subset of exact-class errors remains correct at higher taxonomic ranks. Selected fixed checkpoints with similar clean accuracy can also exhibit distinct sensitivity profiles under controlled transformations. These results show that model comparison on shellfish imagery should account for both task and image condition, rather than rely on a single clean classification score. The dataset and evaluation protocols will be publicly released.

1 INTRODUCTION

The management of nearshore waters and benthic ecosystems is important for maintaining biodiversity and supporting coastal economies. Detailed knowledge of shellfish can clarify their ecological and economic contributions and inform conservation efforts. Ecologically, a global synthesis found that structural restoration of oyster habitats was followed by rapid increases in the biodiversity and abundance of reef-associated fauna during the first two years, although recovery remained 35% below predisturbance levels (Hemraj et al., 2022). Economically, a global review estimated substantial market and non-market value associated with bivalve aquaculture, including food production and non-food ecosystem services (van der Schatte Olivier et al., 2020).

Maintaining ecological resilience requires early, proactive monitoring to enable timely interventions and adaptive strategies that sustain ecosystem functions. However, traditional benthic monitoring relies heavily on specialized taxonomic expertise for morphological identification. This labor-intensive and time-consuming process creates a bottleneck, inherently limiting its scalability for broad-scale ecological assessments (Lejzerowicz et al., 2015). In recent years, rapid advances in computer vision technologies have created substantial opportunities for automated surveys in ecology (Weinstein, 2018). Specifically, foundation models have demonstrated broad transfer capabilities across visual recognition tasks (Radford et al., 2021). Unlike many common visual recognition targets, marine shellfish can exhibit high intra-class variation and subtle inter-class differences (Zhang et al., 2019; Valverde et al., 2025). Existing studies suggest that, when confronted with specialized biological taxa exhibiting high morphological similarity, general-purpose founda-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

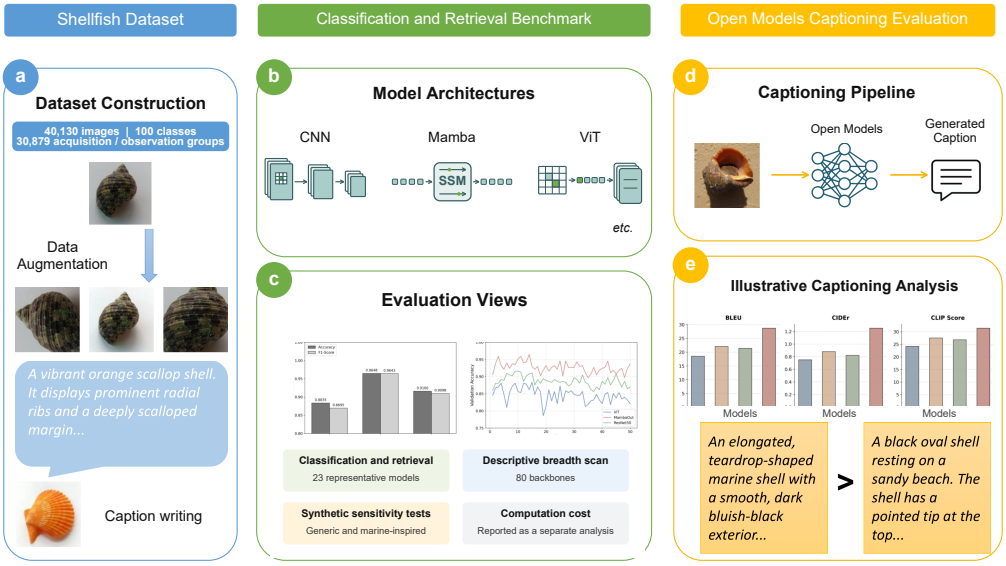


Figure 1: Overview of ShellfishNet from dataset construction to classification, retrieval, descriptive breadth and efficiency analyses, synthetic sensitivity tests, and open-model image captioning.

tion models can show domain-specific limitations in fine-grained recognition (Khan et al., 2025; Felemban et al., 2025; Yu et al., 2026). Such issues can restrict the utility of artificial intelligence technologies for marine ecological monitoring. Furthermore, complex and variable field conditions, including turbid water, dynamic lighting, challenging viewing angles, and complex backgrounds, can degrade image quality or obscure discriminative features, thereby increasing recognition difficulty (Akkaynak & Treibitz, 2018; Khan et al., 2023).

High-quality, domain-specific benchmarks remain scarce. Existing shell-recognition resources include datasets acquired under standardized protocols, such as paired frontal and lateral specimen views, which capture less of the variation in natural lighting, habitat backgrounds, viewing angles, and postures found in field imagery (Zhang et al., 2019). Broader aquatic resources provide greater scale and visual diversity. FishNet contains 94,532 images spanning 17,357 aquatic species, incorporates varied image sizes, resolutions, and illumination conditions, and supports prediction of functional ecological traits such as habitat and nutritional value (Khan et al., 2023). Its emphasis, however, remains on fish rather than benthic shellfish. The resulting gap in taxonomic scope and imaging conditions motivates a shellfish-specific benchmark for fine-grained recognition across varied habitats and acquisition settings (Zhang et al., 2019; Valverde et al., 2025; Khan et al., 2023). Together, these resources leave limited support for a fixed-label, group-disjoint mollusc benchmark unifying classification, retrieval, and controlled condition analysis.

To systematically evaluate the fine-grained recognition and description capabilities of existing vision and multimodal models across varied shellfish imaging conditions, we introduce a domain-specific benchmark. Fig. 1 presents an overview of ShellfishNet, from dataset construction to visual benchmark evaluation and image captioning. The benchmark targets visual variation relevant to shellfish identification across varied acquisition conditions and subject presentations. Comprising 40,130 images across 100 classes, ShellfishNet uses group-disjoint training, validation, and test partitions across 30,879 acquisition or observation groups. Based on the proposed dataset, we benchmark representative models for image classification and image-to-image retrieval and conduct a separate descriptive breadth scan of 80 backbones. The benchmark spans Convolutional Neural Networks (He et al., 2016), Vision Transformers (ViTs) (Dosovitskiy et al., 2021), hybrid and efficient architectures (Tu et al., 2022), self-supervised learning methods (Oquab et al., 2024), and vision-language encoders (Cherti et al., 2023). Furthermore, to extend the benchmark beyond categorical classification, we provide an additional subset with descriptive text captions. Whereas discrete taxonomic labels specify class identity, natural-language descriptions can additionally express visible

attributes and scene context. The subset supports comparison of 10 open image-captioning and vision-language models. Results show different classification and retrieval rankings and distinct responses to controlled transformations among fixed checkpoints with similar clean-image accuracy.

The main contributions of this work are as follows:

- **A provenance-aware, group-disjoint benchmark for fine-grained mollusc recognition.** We construct ShellfishNet, an image dataset comprising 40,130 images across 100 classes, with group-disjoint training, validation, and test partitions across 30,879 acquisition or observation groups. The benchmark targets visual variation relevant to shellfish identification across controlled and intertidal scenes, with variation in image source and subject presentation.
- **Detailed model evaluation.** We benchmark a diverse set of representative models on image classification and image-to-image retrieval using fixed data partitions, common evaluation metrics, and documented model-family training profiles. Furthermore, we assess the sensitivity of selected fixed checkpoints to controlled generic and marine-inspired synthetic transformations.
- **A human-anchored multimodal extension.** Beyond visual recognition, we use a subset with descriptive text annotations to explore the image-captioning capabilities of 10 open image-captioning and vision-language models. This exploration provides a common basis for comparing multimodal image description on shellfish imagery.

2 Related Work

Natural-world benchmarks such as CUB-200-2011 (Wah et al., 2011), iNaturalist (Van Horn et al., 2018), and Stanford Cars (Krause et al., 2013) established widely used settings for fine-grained visual recognition. Marine datasets extend these ideas to aquatic taxonomy and underwater imagery. FishNet contains 94,532 images spanning 17,357 aquatic species and supports recognition, detection, and functional-trait prediction (Khan et al., 2023). FathomNet aggregates expert-curated labels for marine animals and other underwater concepts (Katija et al., 2022). BenthicNet compiles more than 11.4 million seafloor images and curates a representative subset of 1.3 million images for model development (Lowe et al., 2025). These resources provide broad taxonomic or environmental coverage, but their tasks and annotations differ from those of a fixed mollusc recognition benchmark.

Recent marine benchmarks examine fine-grained categories, source variation, and multimodal supervision. ReefNet aggregates approximately 925,000 genus-level hard-coral annotations from 76 CoralNet sources and an additional Red Sea site. Its benchmark covers zero-shot recognition, cross-domain few-shot adaptation, within-source evaluation, and cross-source transfer (Felemban et al., 2025). FishNet++ evaluates multimodal large language models on fine-grained fish recognition. The best tested open-source model achieves less than 10% accuracy, and the resource adds descriptions, keypoints, and bounding boxes (Khan et al., 2025). ORCA provides 14,647 images from 478 marine species, 42,217 bounding boxes, and 22,321 expert-verified instance captions. It evaluates object detection, instance captioning, and visual grounding (Wong et al., 2026).

Fine-grained model diagnosis also extends beyond marine datasets. FG-BMK evaluates large vision–language models at dialogue and feature levels across common and specialized domains. It identifies persistent fine-grained recognition failures and reports DINOv2 among the stronger encoders in feature-level retrieval and classification (Yu et al., 2026). Together, FishNet++ and FG-BMK support direct comparison of language-aligned models and vision-only encoders in fine-grained evaluation.

Work focused on shells and molluscs remains more specialized. Zhang et al. (2019) compiled 59,244 paired frontal and lateral images from 29,622 specimens representing 7,894 shell species. The standardized two-view protocol provides controlled specimen imagery, while later recognition methods address high visual similarity and transfer learning in shellfish and fish classification (Zhang et al., 2023; Ergün, 2025). BackHome19K contains 19,051 natural-light images across 516 mollusc species and targets coast-level geographic provenance (Valverde et al., 2025). ShellfishNet complements these resources with group-disjoint splits for fine-grained mollusc classification and retrieval, a common protocol across model families, controlled transformation tests, and a 512-image caption

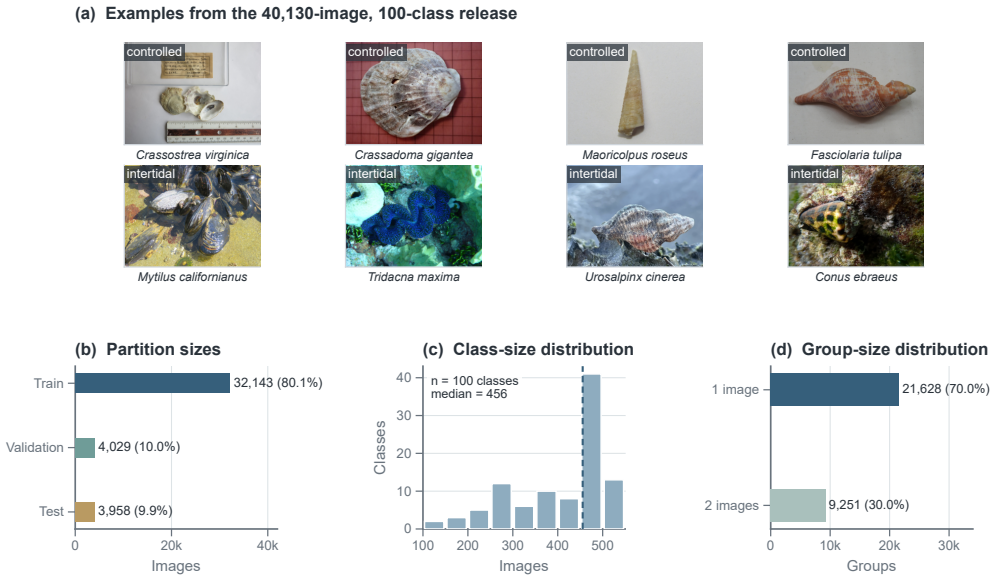


Figure 2: ShellfishNet dataset overview: (a) representative controlled and intertidal scenes; (b) partition sizes; (c) image counts across 100 classes; and (d) acquisition or observation group sizes.

subset. ShellfishNet emphasizes image-level taxonomic comparison. BackHome19K targets provenance, ReefNet studies coral genera and cross-source transfer, and ORCA evaluates instance-level detection and description. BenthicNet provides seafloor coverage at a much larger scale.

3 Benchmark Pipeline Design

ShellfishNet organizes four complementary evaluations around group-disjoint data partitions. Closed-set classification measures category prediction, while image-to-image retrieval tests whether the same classifiers organize same-class images in their learned representation spaces. Captioning extends the benchmark from class identity to visible attributes and scene context. Controlled transformations compare fixed classifiers under prescribed changes in noise, blur, brightness, and contrast, as well as marine-inspired Jerlov II and III attenuation and marine snow.

3.1 Dataset Construction and Split Design

ShellfishNet draws on publicly accessible biodiversity and natural-history repositories. Sources include iNaturalist (Van Horn et al., 2018), GBIF (Lane & Edwards, 2007), Wikimedia Commons (Vaidya et al., 2015), iDigBio (Singer et al., 2018), and the Encyclopedia of Life (Parr et al., 2014), among others. For each image, the release records the source platform, record and media references, URLs, attribution, and license. Images are distributed under CC0, CC BY, or CC BY-SA terms. The corresponding attribution and license links are retained as part of each image's provenance record. Fig. 2 shows representative scenes, partition sizes, class-size variation, and acquisition or observation group sizes.

Before inclusion, each record must have a traceable source URL and an eligible license. Images must have a shorter side of at least 512 pixels. We retain at most two images from one acquisition or observation group and at most 25 images per photographer within each class. An acquisition or observation group links images from the same capture event or source record. Group identifiers use an acquisition session, observation, or specimen identifier when available. Otherwise, they use the source platform and source record. Class labels are checked by a reviewer familiar with shellfish taxonomy, who verifies that the principal subject in each image matches its assigned class. Images that cannot be identified reliably, lack a clear principal subject, or contain conflicting subjects are

marked as ambiguous. Scientific names follow the accepted names in WoRMS (WoRMS Editorial Board, 2026); outdated names and synonyms are mapped to their accepted names. Ambiguous cases and nomenclatural inconsistencies are rechecked against the original source record, image content, and taxonomic database. Samples that cannot be resolved are excluded. Each released image is assigned to the resulting fixed 100-class label set with a class identifier and scientific name. The benchmark contains 40,130 images across these classes. Each group is assigned wholly to one partition. The resulting partitions contain 32,143 training images, 4,029 validation images, and 3,958 test images across 30,879 unique groups. No group is shared between partitions, and the test partition remains reserved for evaluation. Class sizes range from 132 to 500 images. We report class-aware metrics alongside overall accuracy.

The classification and retrieval evaluations use these fixed partitions, while the captioning extension draws its fixed evaluation subset exclusively from the test partition. The collection retains variation in image source and subject presentation, while the released scene field supports controlled and intertidal stratification.

3.2 Classification and Retrieval Benchmark

The primary visual benchmark covers closed-set classification over 100 shellfish classes and class-based image-to-image retrieval. We compare representative visual models from convolutional, transformer-based, hybrid and efficient, self-supervised, and vision-language model families. Retrieval uses the 3,958 test images as queries and the 32,143 training images as a gallery. Images from the same class are treated as relevant.

A descriptive breadth scan evaluates 80 backbones using one fixed random seed and 30 training epochs. Appendix Secs. B.4 and B.5 report the complete model-level and efficiency results and the per-class analysis, respectively.

3.3 Captioning Extension

The captioning extension adds descriptive output to categorical recognition. It uses a fixed subset of 512 test images that covers all 100 classes. Chat-based vision-language models receive a common English instruction, Florence-2 uses its native caption task prompt, and classic captioning models use their native image-to-text interfaces with deterministic generation. We compare 10 open image-captioning and vision-language models against a separate set of reference captions.

A human domain expert leads the reference-caption annotation and determines the final wording. GPT-5.6 Sol (OpenAI, 2026) assists with initial drafts, after which the expert inspects each image and revises every caption to describe visible shellfish attributes and scene context within the concise one-sentence format.

4 Experiments

We evaluate ShellfishNet through closed-set classification, image-to-image retrieval, captioning, and controlled transformations. We use fixed data partitions, specified input, and evaluation protocols.

4.1 Experimental Setup

All classification models use the same 100-class training, validation, and test partitions. During training, random resized cropping produces 224×224-pixel images. The crop scale ranges from 0.80 to 1.00, with aspect ratios between $3/4$ and $4/3$. We apply horizontal flipping with probability 0.5. Brightness, contrast, and saturation factors are sampled from $[0.90, 1.10]$. Validation and test images are resized so that their shorter side is 256 pixels and then center-cropped to 224×224, without test-time augmentation. All classification models except OpenCLIP ViT-B/16 (Cherti et al., 2023) and SigLIP2 Base patch16-224 (Tschannen et al., 2025) use ImageNet normalization. These two encoders use their released normalization statistics. Each model is trained for 100 epochs using cross-entropy loss without label smoothing. We use a cosine-annealing schedule updated after each epoch, with a minimum learning rate of zero. Validation Macro-F1 determines checkpoint selection. Training uses FP16 automatic mixed precision, while validation and test inference use FP32.

Table 1: Held-out classification and retrieval means for 23 visual models. Values are percentages, and column-best means are bold. Sample standard deviations for all seven metrics and seed-specific Top-1 and class-macro mAP@R results appear in Appendix Sec. B.2; mAP@R is computed as a class macro-average (Musgrave et al., 2020).

Model	Top-1	Top-5	Macro-F1	Balanced Acc.	R@1	R@5	mAP@R
ResNet-50	79.82	93.92	79.61	79.47	77.75	87.70	46.61
ResNet-101	80.99	94.72	80.87	80.54	79.55	88.25	50.81
DenseNet-121	77.67	92.71	77.30	77.23	74.63	86.64	39.08
EfficientNetV2-S	82.35	94.60	82.31	82.16	82.06	87.70	69.74
RegNetY-8GF	83.33	95.54	83.10	83.03	81.83	**89.96**	51.09
MobileNetV3-Large	74.91	91.38	74.37	74.33	72.56	85.67	33.42
ConvNeXt-Tiny	82.10	95.16	82.01	81.75	81.79	87.19	67.44
ConvNeXtV2-Tiny	78.20	93.69	77.80	77.52	77.77	84.04	67.98
ViT-B/16	75.59	91.44	75.61	75.29	75.08	80.67	65.77
DeiT III-Small/16	74.76	92.13	74.68	74.19	73.27	81.88	52.53
Swin-Tiny	81.67	94.76	81.46	81.25	81.15	88.63	54.30
SwinV2-Tiny	49.08	76.58	45.93	46.97	46.36	72.48	4.04
PVTv2-B2	79.63	94.00	79.32	79.08	79.32	86.33	56.62
MaxViT-Tiny	**84.39**	**95.62**	**84.40**	**84.26**	**84.26**	87.91	**77.83**
EVA-02-Tiny	70.87	88.79	70.44	70.17	69.18	77.72	56.22
EfficientFormerV2-S2	77.40	92.77	76.88	76.86	77.43	84.50	60.20
MAE ViT-B/16	72.97	90.98	72.85	72.35	71.46	80.76	49.03
iBOT ViT-B/16	79.59	93.91	79.61	79.22	78.78	87.49	47.39
DINOv2 ViT-B/14	82.71	94.87	82.75	82.31	82.95	87.79	69.66
OpenCLIP ViT-B/16	75.12	91.19	74.93	74.56	74.20	80.95	65.04
SigLIP2 Base patch16-224	78.58	92.73	78.36	77.94	78.15	84.46	64.37
MambaOut-Tiny	76.93	92.93	76.75	76.55	75.69	85.51	51.07
RepViT-M1.0	51.37	77.24	50.30	50.41	34.73	59.68	4.21

Model references. ResNet-50/101 (He et al., 2016); DenseNet-121 (Huang et al., 2017); EfficientNetV2-S (Tan & Le, 2021); RegNetY-8GF (Radosavovic et al., 2020); MobileNetV3-Large (Howard et al., 2019); ConvNeXt-Tiny (Liu et al., 2022b); ConvNeXtV2-Tiny (Woo et al., 2023); ViT-B/16 (Dosovitskiy et al., 2021); DeiT III-Small/16 (Touvron et al., 2022); Swin-Tiny (Liu et al., 2021b); SwinV2-Tiny (Liu et al., 2022a); PVTv2-B2 (Wang et al., 2022); MaxViT-Tiny (Tu et al., 2022); EVA-02-Tiny (Fang et al., 2024); EfficientFormerV2-S2 (Li et al., 2023b); MAE ViT-B/16 (He et al., 2022); iBOT ViT-B/16 (Zhou et al., 2022); DINOv2 ViT-B/14 (Oquab et al., 2024); OpenCLIP ViT-B/16 (Cherti et al., 2023); SigLIP2 Base patch16-224 (Tschannen et al., 2025); MambaOut-Tiny (Yu & Wang, 2025); RepViT-M1.0 (Wang et al., 2024).

Optimizer settings follow two fixed architecture-specific profiles. Sixteen models use AdamW with learning rate 3×10^{-5} and weight decay 0.05. The remaining seven models use SGD with learning rate 0.01, momentum 0.9, and weight decay 0.0001. We selected these profiles using validation performance before test evaluation. Every run has an effective batch size of 48, with gradient accumulation used where required by memory. Each model is trained with seeds 41, 42, and 43, yielding 69 formal runs. For each model, we aggregate held-out classification and retrieval results using the arithmetic mean and sample standard deviation over its runs. Table 1 reports the means for all seven measures, while Appendix Sec. B.2 reports sample standard deviations for all seven metrics and seed-specific Top-1 and class-macro mAP@R results.

For classification, Top-1 and Top-5 accuracy measure whether the true class is the highest-scoring prediction or appears among the five highest-scoring predictions, respectively. Macro-F1 and Balanced Accuracy average performance across classes and can reveal effects hidden by the overall image count. We report all four measures on the held-out test partition.

For image-to-image retrieval, the 3,958 test images serve as queries against a gallery of 32,143 training images. Images from the same class are relevant. For each classifier, retrieval uses the tensor supplied directly to its trained 100-class classification head. Query and gallery vectors are L2-normalized and ranked by cosine similarity, without additional retrieval training. We report Recall@1, Recall@5, and class-macro mAP@R (Musgrave et al., 2020). Here, R is the number of same-class gallery images for a query. Appendix Sec. C.1 reports other retrieval measures.

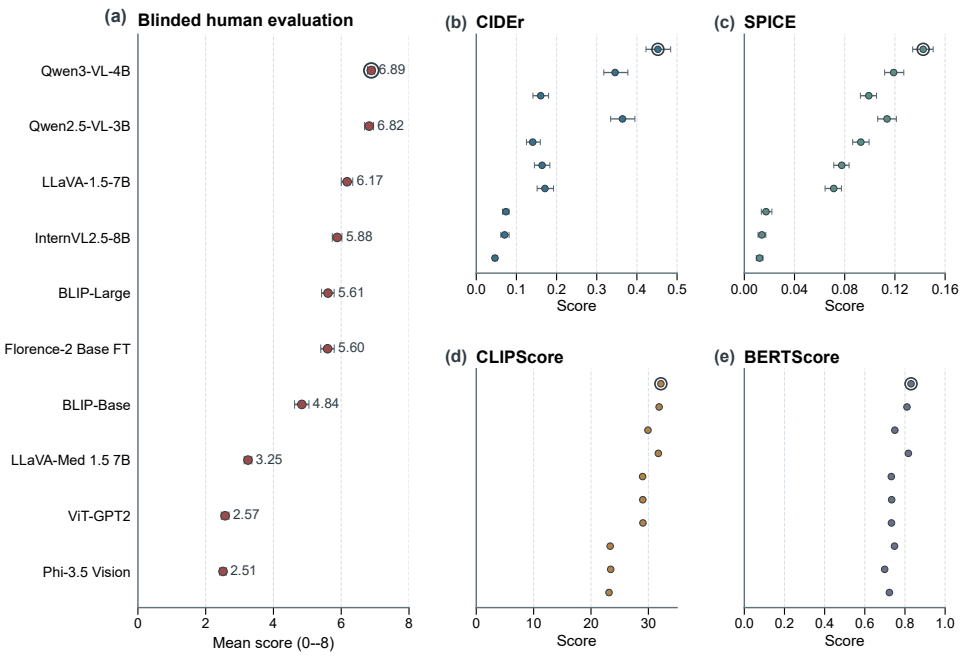


Figure 3: Human and automatic caption evaluation for ten open models on 512 images. Panel (a) shows blinded human overall scores, and panels (b)–(e) show CIDEr, SPICE, CLIPScore, and BERTScore. Horizontal intervals denote 95% bootstrap confidence intervals; hollow rings mark panel-best point estimates.

For captioning, the fixed subset contains 512 test images and covers all 100 classes. Each image is provided to all ten open models in the same standardized 1024×1024 RGB format. Chat-based vision-language models use a common scientific image-captioning instruction. Florence-2 uses its native caption task prompt, while the classic captioning models use their native image-to-text interfaces. The common instruction requests one English sentence of at most 30 words describing the visible subject, salient attributes, and marine scene context. It permits a common name, genus, or species only when supported by visible evidence; otherwise, it requests observable characteristics without speculation. Decoding uses batch size one and at most 48 new tokens. Prompted models use temperature zero where supported, while the classic models use deterministic generation through their native interfaces. The reference captions are annotated and finalized by a human domain expert, with GPT-5.6 Sol assisting with initial drafts. We compare generated descriptions with these references using CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), CLIPScore (Hessel et al., 2021), and BERTScore (Zhang et al., 2020). We also report BLEU-4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), scientific-name token mention recall, and caption length ratio. Two human reviewers additionally score the generated captions with model identities hidden. Subject correctness, visual factuality, grounded informativeness, and language and format are each rated from 0 to 2, and their sum gives an overall score from 0 to 8. Every output receives one primary human score. A 200-output subset is also scored by both reviewers to measure agreement; these second scores are excluded from model means. Human-score intervals use 10,000 stratified paired-image bootstrap resamples; automatic-metric intervals use image-level bootstrap resampling.

4.2 Main Classification and Retrieval Benchmark

Mean Top-1 accuracy ranges from 49.08% for SwinV2-Tiny (Liu et al., 2022a) to 84.39% for MaxViT-Tiny (Tu et al., 2022). Mean Top-1 and Macro-F1 differ by less than one percentage point for 21 of the 23 models, showing close agreement between image-level and class-aware performance for most architectures. MaxViT-Tiny gives the highest mean on all four classification measures,

whereas SwinV2-Tiny gives the lowest. Several convolutional and hybrid models occupy the upper part of the table. The self-supervised encoders also vary in performance: DINOv2 ViT-B/14 (Oquab et al., 2024) is near the top, while MAE ViT-B/16 (He et al., 2022) is lower. Classification and retrieval rankings are not identical, with the largest differences appearing for class-macro mAP@R rather than Recall@1. MaxViT-Tiny leads mean R@1 and class-macro mAP@R, while RegNetY-8GF (Radosavovic et al., 2020) leads mean R@5. RepViT-M1.0 (Wang et al., 2024) has the lowest mean R@1, whereas SwinV2-Tiny has the lowest mean class-macro mAP@R.

Using the fixed seed-41 checkpoints, a scene-stratified analysis separates the test set into 949 controlled-scene and 3,009 intertidal-scene images. Controlled-scene Top-1 exceeds intertidal-scene Top-1 for 20 of the 23 models, and the largest absolute difference between the strata is 4.10 percentage points. Appendix Sec. C.5 reports these aggregate results and a taxon-matched recall analysis across three seeds for the 96 taxa represented in both strata. In the held-out benchmark, 20 of the 23 models have a Top-1 sample standard deviation below one percentage point; SwinV2-Tiny, iBOT ViT-B/16, and RepViT-M1.0 exceed that threshold. Dispersion is metric-dependent: the class-macro mAP@R sample standard deviation reaches 11.51 percentage points for iBOT ViT-B/16 and 7.40 points for PVTv2-B2. These standard deviations are descriptive summaries over three runs, not significance tests. In the separate taxonomy analysis, projecting the fixed held-out predictions to higher taxonomic ranks raises Top-1 for every model at both family and order levels. The mean gains over class-level accuracy are 4.05 and 8.27 percentage points, respectively. This pattern indicates that some fine-grained errors occur within the correct family or order. Appendix Sec. C.3 provides the complete results.

4.3 Captioning Benchmark

The captioning benchmark evaluates ten open image-captioning and vision-language models on the same 512 held-out images. The reference descriptions are the human expert annotations described above, prepared with draft assistance from GPT-5.6 Sol. Fig. 3 reports the blinded human overall score together with four automatic metrics, each with a 95% bootstrap interval. CIDEr measures TF-IDF-weighted n-gram agreement with the expert-finalized reference for each image, whereas SPICE compares semantic scene-graph content. CLIPScore measures image–text alignment, while BERTScore measures contextual textual similarity. Appendix Secs. D.2 and D.3 provide the human dimension scores, agreement analysis, and complete automatic metric matrix.

The mean human score is 6.89/8 for Qwen3-VL-4B (Bai et al., 2025a) (95% CI: 6.78–6.99), 6.82 for Qwen2.5-VL-3B (Bai et al., 2025b) (6.69–6.96), and 6.17 for LLaVA-1.5-7B (Liu et al., 2024) (6.00–6.34). LLaVA-Med 1.5 7B (Li et al., 2023a), ViT-GPT2 (NLP Connect, 2022; Dosovitskiy et al., 2021; Radford et al., 2019), and Phi-3.5 Vision (Abdin et al., 2024) receive lower overall means of 3.25, 2.57, and 2.51. Their mean language-and-format scores remain high (2.00, 1.85, and 1.89), while their subject-correctness means are 0.18, 0.09, and 0.03, respectively, separating well-formed output from correct visual identification. Qwen3-VL-4B also has the largest point estimate for each displayed automatic metric and leads BLEU-4, METEOR, ROUGE-L, and scientific-name token mention recall in the complete automatic evaluation. InternVL2.5-8B (Chen et al., 2024) and Qwen2.5-VL-3B form the next group on most reference-based and image–text measures, although their order varies by metric. Appendix Sec. D.5 contrasts the expert-annotated references with exact model outputs in five qualitative failure cases.

4.4 Targeted Synthetic Sensitivity Analyses

We evaluate two sets of fixed seed-41 classifiers on the same 3,958 held-out images, applying every transformation independently and without parameter updates. The generic set comprises ResNet-50 (He et al., 2016), ConvNeXtV2-Tiny (Woo et al., 2023), SwinV2-Tiny, MaxViT-Tiny, and DINOv2 ViT-B/14. It covers Gaussian noise, Gaussian blur, brightness, and contrast at three increasing levels. A separate marine-inspired set comprises ResNet-50, MobileNetV3-Large (Howard et al., 2019), EfficientFormerV2-S2 (Li et al., 2023b), DINOv2 ViT-B/14, and RepViT-M1.0. Its two water-attenuation conditions follow the CADDY implementation of Jerlov II and III water types (Gomez Chavez et al., 2019; Gomez Chavez, 2019), while the marine-snow condition adapts the PHISWID synthesis to the RGB benchmark inputs (Kaneko et al., 2026). Figure 4 summarizes Top-1 accuracy; Appendix Secs. E.1 and F.1 provide the exact settings and complete Top-1, Macro-

F1, and Balanced Accuracy results. Accordingly, the analysis characterizes checkpoint-level rather than seed-averaged sensitivity.

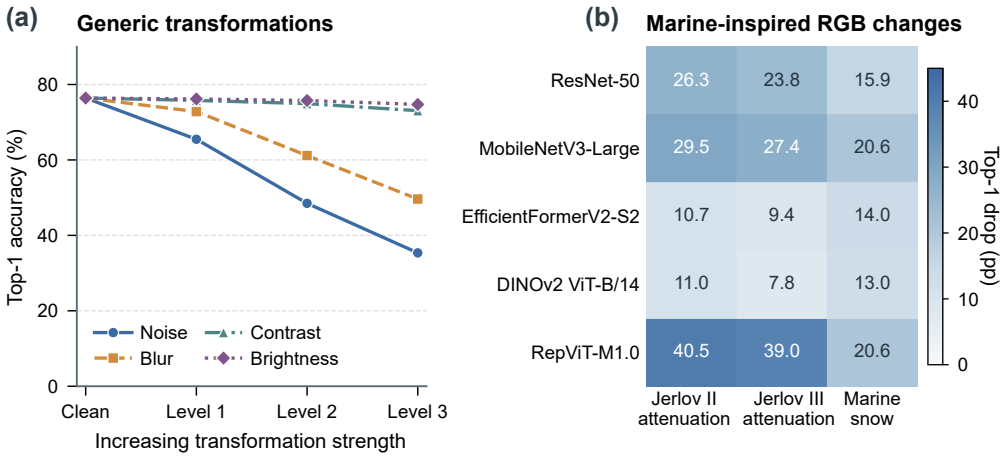


Figure 4: Top-1 sensitivity to controlled transformations on 3,958 held-out images. Panel (a) shows mean accuracy for five fixed checkpoints under Gaussian noise, Gaussian blur, brightness, and contrast at three levels. Panel (b) shows per-checkpoint decreases from clean accuracy for a separate five-checkpoint set under Jerlov II and III attenuation and marine snow; cell values are percentage-point decreases. Cell values are rounded to one decimal place; narrative averages use the underlying unrounded values.

In Figure 4(a), the generic five-model mean is 76.41% on clean images. At level 3, it decreases to 35.37% under Gaussian noise and 49.64% under Gaussian blur, compared with 73.06% under contrast change and 74.71% under brightness change. Panel (b) expresses each marine-transformation result as a decrease from the corresponding model's clean Top-1. The five-model mean decreases are 23.59 percentage points for Jerlov II attenuation, 16.82 for marine snow, and 21.46 for Jerlov III attenuation. The model rows show a wider separation: RepViT-M1.0 loses 40.5 and 39.0 points under Jerlov II and III attenuation, whereas the three decreases for DINOv2 ViT-B/14 range from 7.8 to 13.0 points. Across these fixed transformations, sensitivity varies across appearance changes and evaluated fixed checkpoints.

5 Conclusions and Future Work

This paper introduces ShellfishNet, a benchmark of 40,130 images across 100 shellfish classes with provenance metadata and group-disjoint splits. Across representative visual models, image-level and class-aware classification measures generally agree, whereas classification and image-to-image retrieval produce different rankings. Post-hoc taxonomic projection shows that some exact-class errors fall within the correct family or order, indicating that species-level accuracy alone does not reveal the taxonomic proximity of all errors. Selected fixed checkpoints spanning different clean held-out accuracies can still exhibit distinct sensitivity to controlled generic and marine-inspired synthetic transformations. In the captioning study, human scores and automatic metrics assess different properties and yield metric-dependent model rankings. Overall, within ShellfishNet and the evaluated transformations, no single classification score captures all measured aspects of model performance, motivating task- and condition-aware comparison. We will release the dataset, splits, and evaluation protocols to support reproducible research on shellfish recognition and related marine vision tasks. Future work will extend ShellfishNet toward broader cross-dataset transfer evaluation, validation under real underwater conditions and on externally collected degraded imagery, broader geographic and temporal coverage, and larger-scale multi-reviewer evaluation of generated captions.

AI USE STATEMENT

Generative AI tools assisted with polishing the English manuscript. For the captioning subset, these tools also assisted with initial reference-caption drafts; a human domain expert led the annotation, inspected every image, revised the drafts, and determined the final wording. All AI-assisted text was verified against the relevant data sources or experimental outputs. The authors take full responsibility for the final manuscript, data, figures, code, and conclusions.

ETHICS STATEMENT

This work uses publicly accessible biodiversity and natural-history images rather than data collected from human participants or through animal experimentation. The released records retain source, attribution, and license information, and the included images are distributed under the licenses recorded for their sources, including CC0, CC BY, and CC BY-SA.

REPRODUCIBILITY STATEMENT

The main paper specifies the source platforms, inclusion and taxonomy checks, group-disjoint partitions, preprocessing, training, inference, and evaluation protocols. Appendix Sec. A documents the complete class inventory, partition statistics, and released record schema. Appendix Secs. B.2, B.3, and B.4 provide the complete held-out results, the specialized fine-grained baseline, and the 80-backbone scan configuration, while Sec. C reports complementary retrieval, few-shot prototype readout, taxonomy-aware, family-auxiliary, scene-stratified, shared-taxon transfer, and zero-shot analyses. Appendix Sec. D gives the model-specific captioning interfaces, complete metric matrix, and qualitative cases. Sec. E gives the complete transformation-by-level results, and Sec. F documents the marine-degradation and class-frequency diagnostics. The dataset manifest, provenance fields, experiment outputs, and accompanying implementation materials will be released with the paper to support independent reconstruction of the benchmark.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6723–6732, 2018. doi: 10.1109/CVPR.2018.00703. URL https://doi.org/10.1109/CVPR.2018.00703.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pp. 382–

398, 2016. doi: 10.1007/978-3-319-46454-1_24. URL https://doi.org/10.1007/978-3-319-46454-1_24.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005. URL https://aclanthology.org/W05-0909/.

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. EfficientViT: Lightweight multi-scale attention for high-resolution dense prediction. In *IEEE/CVF International Conference on Computer Vision*, pp. 17302–17313, 2023. doi: 10.1109/ICCV51070.2023.01587. URL https://openaccess.thecvf.com/content/ICCV2023/html/Cai_EfficientViT_Lightweight_Multi-Scale_Attention_for_High-Resolution_Dense_Prediction_ICCV_2023_paper.html.

Chengpeng Chen, Zichao Guo, Haien Zeng, Pengfei Xiong, and Jian Dong. RepGhost: A hardware-efficient ghost module via re-parameterization, 2022. URL https://arxiv.org/abs/2211.06088.

Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arXiv.1707.01629. URL https://arxiv.org/abs/1707.01629.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. URL https://arxiv.org/abs/2412.05271.

Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 569–578, 2021. doi: 10.1109/ICCV48922.2021.00063. URL https://doi.org/10.1109/ICCV48922.2021.00063.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language–image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276. URL https://doi.org/10.1109/CVPR52729.2023.00276.

Po-Yung Chou. FGVC-HERBS. Software, GitHub repository, 2023. URL https://github.com/chou141253/FGVC-HERBS/tree/f3deb47895f1ee5ce8058f28c44b1b6b35057407. Commit f3deb47895f1ee5ce8058f28c44b1b6b35057407.

Po-Yung Chou, Yu-Yung Kao, and Cheng-Hung Lin. Fine-grained visual classification with high-temperature refinement and background suppression, 2023. URL https://arxiv.org/abs/2303.06442.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Advances in Neural Information Processing Systems*, 2021. doi: 10.48550/arXiv.2104.13840. URL https://arxiv.org/abs/2104.13840.

Cheng Cui, Tingquan Gao, Shengyu Wei, Yuning Du, Ruoyu Guo, Shuilong Dong, Bin Lu, Ying Zhou, Xueying Lv, Qiwen Liu, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. PP-LCNet: A lightweight CPU convolutional neural network, 2021. URL https://arxiv.org/abs/2109.15099.

Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Bichen Wu, Zijian He, Zhen Wei, Kan Chen, Yuandong Tian, Matthew Yu, Peter Vajda, and Joseph E. Gonzalez. FBNetV3: Joint architecture–recipe search using predictor pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16271–16280, 2021a. doi: 10.1109/CVPR46437.2021.01601. URL https://doi.org/10.1109/CVPR46437.2021.01601.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems*, 2021b. doi: 10.48550/arXiv.2106.04803. URL https://arxiv.org/abs/2106.04803.

Stéphane d'Ascoli, Hugo Touvron, Matthew L. Leavitt, Ari S. Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, volume 139, pp. 2286–2296, 2021. URL https://proceedings.mlr.press/v139/d-ascoli21a.html.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. DaViT: Dual attention vision transformers. In *European Conference on Computer Vision*, pp. 74–92, 2022. doi: 10.1007/978-3-031-20053-3_5. URL https://doi.org/10.1007/978-3-031-20053-3_5.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. RepVGG: Making VGG-style convnets great again. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13728–13737, 2021. doi: 10.1109/CVPR46437.2021.01352. URL https://doi.org/10.1109/CVPR46437.2021.01352.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity–uncertainty trade-off. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–45, 2020. doi: 10.1109/CVPRW50498.2020.00010. URL https://doi.org/10.1109/CVPRW50498.2020.00010.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. XCiT: Cross-covariance image transformers. In *Advances in Neural Information Processing Systems*, 2021. doi: 10.48550/arXiv.2106.09681. URL https://arxiv.org/abs/2106.09681.

Ebru Ergün. SwinFishNet: A Swin Transformer-based approach for automatic fish species classification using transfer learning. *PLOS ONE*, 20(5):e0322711, 2025. doi: 10.1371/journal.pone.0322711.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. doi: 10.1016/j.imavis.2024.105171. URL https://doi.org/10.1016/j.imavis.2024.105171.

Abdulwahab Felemban, Yahia Battach, Faizan Farooq Khan, Yuqian Fu, Xuhui Liu, Yesmeen M. Khattab, Yousef A. Radwan, Xiang Li, Fabio Marchese, Sara Beery, Burton H. Jones, Francesca Benzoni, and Mohamed Elhoseiny. ReefNet: A large-scale dataset and benchmark for fine-grained coral reef recognition, 2025. URL https://arxiv.org/abs/2510.16822.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2021. doi: 10.1109/TPAMI.2019.2938758. URL https://doi.org/10.1109/TPAMI.2019.2938758.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. doi: 10.48550/arXiv.1705.08500. URL https://arxiv.org/abs/1705.08500.

Arturo Gomez Chavez. Underwater image distortions. Software, GitHub repository, 2019. URL https://github.com/jacobs-robotics/underwater-image-distortions/tree/6182c41323272ca0cc1344b379e782a0d6bc4416. Commit 6182c41323272ca0cc1344b379e782a0d6bc4416.

Arturo Gomez Chavez, Andrea Ranieri, Davide Chiarella, Enrica Zereik, Anja Babić, and Andreas Birk. CADDY underwater stereo-vision dataset for human–robot interaction (HRI) in the context of diver activities. *Journal of Marine Science and Engineering*, 7(1):16, 2019. doi: 10.3390/jmse7010016. URL https://doi.org/10.3390/jmse7010016.

Jianyang Gu, Samuel Stevens, Elizabeth G. Campolongo, Matthew J. Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasila Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 102778–102811. Curran Associates, Inc., 2025. doi: 10.52202/085713-3436. URL https://papers.nips.cc/paper_files/paper/2025/hash/94da80cbfe870c1db958c88a8a27018c-Abstract-Conference.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017. doi: 10.48550/arXiv.1706.04599. URL https://arxiv.org/abs/1706.04599.

Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 732–741, 2021. doi: 10.1109/CVPR46437.2021.00079. URL https://doi.org/10.1109/CVPR46437.2021.00079.

Ali Hatamizadeh and Jan Kautz. MambaVision: A hybrid mamba–transformer vision backbone. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25261–25270, 2025. doi: 10.1109/CVPR52734.2025.02352. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hatamizadeh_MambaVision_A_Hybrid_Mamba-Transformer_Vision_Backbone_CVPR_2025_paper.html.

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2206.09959. URL https://arxiv.org/abs/2206.09959.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Deevesh A. Hemraj, Melanie J. Bishop, Boze Hancock, Jay J. Minuti, Ruth H. Thurstan, Philine S. E. Zu Ermgassen, and Bayden D. Russell. Oyster reef restoration fails to recoup global historic ecosystem losses despite substantial biodiversity gain. *Science Advances*, 8(47):eabp8747, 2022. doi: 10.1126/sciadv.abp8747. URL https://doi.org/10.1126/sciadv.abp8747.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 11916–11925, 2021. doi: 10.1109/ICCV48922.2021.01172. URL https://doi.org/10.1109/ICCV48922.2021.01172.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc V. Le. Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019. doi: 10.1109/ICCV.2019.00140. URL https://doi.org/10.1109/ICCV.2019.00140.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243. URL https://doi.org/10.1109/CVPR.2017.243.

Reina Kaneko, Takumi Ueda, Hiroshi Higashi, and Yuichi Tanaka. PHISWID: Physics-inspired underwater image data set synthesized from RGB-D images. *APSIPA Transactions on Signal and Information Processing*, 15(1):1–25, 2026. doi: 10.1108/ATSIP-02-2026-001. URL https://doi.org/10.1108/ATSIP-02-2026-001.

Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, and Katherine L. C. Bell. Fathomnet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1):15914, 2022. doi: 10.1038/s41598-022-19939-2.

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*, pp. 20496–20506, 2023. doi: 10.1109/ICCV51070.2023.01874. URL https://openaccess.thecvf.com/content/ICCV2023/html/Khan_FishNet_A_Large-scale_Dataset_and_Benchmark_for_Fish_Recognition_Detection_ICCV_2023_paper.html.

Faizan Farooq Khan, Yousef Radwan, Eslam Abdelrahman, Abdulwahab Felemban, Aymen Mir, Nico K. Michiels, Andrew J. Temple, Michael L. Berumen, and Mohamed Elhoseiny. FishNet++: Analyzing the capabilities of multimodal large language models in marine biology, 2025. URL https://arxiv.org/abs/2509.25564.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Meredith A. Lane and James L. Edwards. The global biodiversity information facility (gbif). In *Biodiversity Databases*, Systematics Association Special Volumes, pp. 1–4. CRC Press, 2007. doi: 10.1201/9781439832547.ch1.

Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and GPU-computation efficient backbone network for real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 752–760, 2019. doi: 10.1109/CVPRW.2019.00103. URL https://doi.org/10.1109/CVPRW.2019.00103.

Franck Lejzerowicz, Philippe Esling, Loïc Pillet, Thomas A. Wilding, Kenneth D. Black, and Jan Pawlowski. High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports*, 5(1):13932, 2015. doi: 10.1038/srep13932. URL https://doi.org/10.1038/srep13932.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems 36*, pp. 28541–28564, 2023a. doi: 10.52202/075280-1240. URL https://doi.org/10.52202/075280-1240.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping language–image pre-training for unified vision–language understanding and generation. In *International Conference on Machine Learning*, 2022a. doi: 10.48550/arXiv.2201.12086. URL https://arxiv.org/abs/2201.12086.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4794–4804, 2022b. doi: 10.1109/CVPR52688.2022.00476. URL https://doi.org/10.1109/CVPR52688.2022.00476.

Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. EfficientFormer: Vision transformers at MobileNet speed. In *Advances in Neural Information Processing Systems*, pp. 12934–12949, 2022c. doi: 10.52202/068431-0940. URL https://doi.org/10.52202/068431-0940.

Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for MobileNet size and speed. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16843–16854, 2023b. doi: 10.1109/ICCV51070.2023.01549. URL https://doi.org/10.1109/ICCV51070.2023.01549.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004. URL https://aclanthology.org/W04-1013/.

Ming Lin, Hesen Chen, Xiuyu Sun, Qi Qian, Hao Li, and Rong Jin. Neural architecture design for GPU-efficient networks, 2020. URL https://arxiv.org/abs/2006.14090.

Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay attention to MLPs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9204–9215. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/4cc05b35c2f937c5bd9e7d41d3686fff-Abstract.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. EfficientViT: Memory efficient vision transformer with cascaded group attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14430, 2023. doi: 10.1109/CVPR52729.2023.01386. URL https://doi.org/10.1109/CVPR52729.2023.01386.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021b. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.org/10.1109/ICCV48922.2021.00986.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11999–12009, 2022a. doi: 10.1109/CVPR52688.2022.01170. URL https://doi.org/10.1109/CVPR52688.2022.01170.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022b. doi: 10.1109/CVPR52688.2022.01167. URL https://doi.org/10.1109/CVPR52688.2022.01167.

Scott C. Lowe, Benjamin Misiuk, Isaac Xu, Shakhboz Abdulazizov, Amit R. Baroi, Alex C. Bastos, Merlin Best, Vicki Ferrini, Ariell Friedman, Deborah Hart, Ove Hoegh-Guldberg, Daniel Ierodiaconou, Julia Mackin-McLaughlin, Kathryn Markey, Pedro S. Menandro, Jacquomo Monk, Shreya Nemani, John O'Brien, Elizabeth Oh, Luba Y. Reshitnyk, Katleen Robert, Chris M. Roelfsema, Jessica A. Sameoto, Alexandre C. G. Schimel, Jordan A. Thomson, Brittany R. Wilson, Melisa C. Wong, Craig J. Brown, and Thomas Trappenberg. BenthicNet: A global compilation of seafloor images for deep learning applications. *Scientific Data*, 12(1):230, 2025. doi: 10.1038/s41597-025-04491-1. URL https://doi.org/10.1038/s41597-025-04491-1.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision – ECCV 2020*, pp. 681–699, 2020. doi: 10.1007/978-3-030-58595-2_41. URL https://doi.org/10.1007/978-3-030-58595-2_41.

Niv Nayman, Yonathan Aflalo, Asaf Noy, and Lihi Zelnik-Manor. HardCoRe-NAS: Hard constrained differentiable neural architecture search. In *International Conference on Machine Learning*, volume 139, pp. 7979–7990, 2021. doi: 10.48550/arXiv.2102.11646. URL https://proceedings.mlr.press/v139/nayman21a.html.

NLP Connect. vit-gpt2-image-captioning model card. Hugging Face, 2022. URL https://huggingface.co/nlpconnect/vit-gpt2-image-captioning.

OpenAI. GPT-5.6: Frontier Intelligence That Scales with Your Ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed: 2026-09-02.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Cynthia S. Parr, Nathan Wilson, Patrick Leary, Katja Schulz, Kristen Lans, Lisa Walley, Jennifer Hammock, Anthony Goddard, Jeremy Rice, Marie Studer, Jeffrey Holmes, and Robert Corrigan, Jr. The encyclopedia of life v2: Providing global access to knowledge about life on earth. *Biodiversity Data Journal*, 2:e1079, 2014. doi: 10.3897/bdj.2.e1079.

Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, Vaibhav Aggarwal, Tenghui Zhu, Daniele Moro, and Andrew Howard. MobileNetV4: Universal models for the mobile ecosystem. In *European Conference on Computer Vision*, pp. 78–96, 2024. doi: 10.1007/978-3-031-73661-2_5. URL https://doi.org/10.1007/978-3-031-73661-2_5.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10425–10433, 2020. doi: 10.1109/CVPR42600.2020.01044. URL https://doi.org/10.1109/CVPR42600.2020.01044.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2306.00989. URL https://arxiv.org/abs/2306.00989.

Randal A. Singer, Kevin J. Love, and Lawrence M. Page. A survey of digitized data from u.s. fish collections in the idigbio data aggregator. *PLOS ONE*, 13(12):e0207636, 2018. doi: 10.1371/journal.pone.0207636.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 2019. doi: 10.48550/arXiv.1905.11946. URL https://arxiv.org/abs/1905.11946.

Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10096–10106. PMLR, 2021. URL https://proceedings.mlr.press/v139/tan21a.html.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. MnasNet: Platform-aware neural architecture search for mobile. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2815–2823, 2019. doi: 10.1109/CVPR.2019.00293. URL https://doi.org/10.1109/CVPR.2019.00293.

Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Chao Xu, and Yunhe Wang. GhostNetV2: Enhance cheap operation with long-range attention. In *Advances in Neural Information Processing Systems*, pp. 9969–9982, 2022. doi: 10.52202/068431-0724. URL https://doi.org/10.52202/068431-0724.

Yuki Tatsunami and Masato Taki. Sequencer: Deep LSTM for image classification. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2205.01972. URL https://arxiv.org/abs/2205.01972.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems*, 2021. doi: 10.48550/arXiv.2105.01601. URL https://arxiv.org/abs/2105.01601.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, 2021. doi: 10.48550/arXiv.2012.12877. URL https://proceedings.mlr.press/v139/touvron21a.html.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *Computer Vision – ECCV 2022*, pp. 516–533, 2022. doi: 10.1007/978-3-031-20053-3_30. URL https://doi.org/10.1007/978-3-031-20053-3_30.

Asher Trockman and J. Zico Kolter. Patches are all you need? *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=rAnB7JSMXL.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *Computer Vision – ECCV 2022*, pp. 459–479, 2022. doi: 10.1007/978-3-031-20053-3_27. URL https://doi.org/10.1007/978-3-031-20053-3_27.

Gaurav Vaidya, Dimitris Kontokostas, Magnus Knuth, Jens Lehmann, and Sebastian Hellmann. Dbpedia commons: Structured multimedia metadata from the wikimedia commons. In *The Semantic Web - ISWC 2015*, Lecture Notes in Computer Science, pp. 281–289. Springer International Publishing, 2015. doi: 10.1007/978-3-319-25010-6_17.

Alexander Valverde, Luis Fernando Solano Coto, and André Montoya Fernández. Back home: A computer vision solution to seashell identification for ecological restoration. In *2025 IEEE/CVF International Conference on Computer Vision Workshops*, pp. 5159–5168, 2025. doi: 10.1109/ICCVW69036.2025.00537. URL https://doi.org/10.1109/ICCVW69036.2025.00537.

Andrew van der Schatte Olivier, Laurence Jones, Lewis Le Vay, Michael Christie, James Wilson, and Shelagh K. Malham. A global review of the ecosystem services provided by bivalve aquaculture. *Reviews in Aquaculture*, 12(1):3–25, 2020. doi: 10.1111/raq.12301. URL https://doi.org/10.1111/raq.12301. First published online in 2018.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018. doi: 10.1109/CVPR.2018.00914.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. MobileOne: An improved one millisecond mobile backbone. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7907–7917, 2023. doi: 10.1109/CVPR52729.2023.00764. URL https://doi.org/10.1109/CVPR52729.2023.00764.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. RepViT: Revisiting mobile CNN from ViT perspective. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15909–15920, 2024. doi: 10.1109/CVPR52733.2024.01506. URL https://doi.org/10.1109/CVPR52733.2024.01506.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021. doi: 10.1109/TPAMI.2020.2983686. URL https://doi.org/10.1109/TPAMI.2020.2983686.

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. doi: 10.1007/s41095-022-0274-8. URL https://doi.org/10.1007/s41095-022-0274-8.

Ben G. Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3): 533–545, 2018. doi: 10.1111/1365-2656.12780. URL https://doi.org/10.1111/1365-2656.12780. First published online in 2017.

Yuk-Kwan Wong, Haixin Liang, Zeyu Ma, Yiwei Chen, Ziqiang Zheng, Rinaldi Gotama, Pascal Sebastian, Lauren D. Sparks, and Sai-Kit Yeung. ORCA: Object recognition and comprehension for archiving marine species. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1597–1609, 2026. doi: 10.1109/WACV61042.2026.00161. URL https://openaccess.thecvf.com/content/WACV2026/html/Wong_ORCA_Object_Recognition_and_Comprehension_for_Archiving_Marine_Species_WACV_2026_paper.html.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, 2023. doi: 10.1109/CVPR52729.2023.01548. URL https://doi.org/10.1109/CVPR52729.2023.01548.

WoRMS Editorial Board. World register of marine species. Flanders Marine Institute (VLIZ), 2026. URL https://doi.org/10.14284/170. Accessed: 2026-08-21.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, pp. 68–85, 2022. doi: 10.1007/978-3-031-19803-8_5. URL https://doi.org/10.1007/978-3-031-19803-8_5.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024. doi: 10.1109/CVPR52733.2024.00461. URL https://doi.org/10.1109/CVPR52733.2024.00461.

Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2203.11926. URL https://arxiv.org/abs/2203.11926.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2403–2412, 2018. doi: 10.1109/CVPR.2018.00255. URL https://doi.org/10.1109/CVPR.2018.00255.

Hong-Tao Yu, Chen-Wei Xie, Yuxin Peng, Serge Belongie, and Xiu-Shen Wei. Benchmarking large vision-language models on fine-grained image tasks: From evaluation to diagnosis, 2026. URL https://arxiv.org/abs/2606.19053.

Weihao Yu and Xinchao Wang. MambaOut: Do we really need mamba for vision? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4484–4496, 2025. doi: 10.1109/CVPR52734.2025.00423. URL https://doi.org/10.1109/CVPR52734.2025.00423.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. MetaFormer is actually what you need for vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10819, 2022. doi: 10.1109/CVPR52688.2022.01055. URL https://doi.org/10.1109/CVPR52688.2022.01055.

Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. MetaFormer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):896–912, 2024a. doi: 10.1109/TPAMI.2023.3329173. URL https://doi.org/10.1109/TPAMI.2023.3329173.

Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. InceptionNeXt: When inception meets ConvNeXt. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5672–5683, 2024b. doi: 10.1109/CVPR52733.2024.00542. URL https://doi.org/10.1109/CVPR52733.2024.00542.

Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. VOLO: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (5):6575–6586, 2023. doi: 10.1109/TPAMI.2022.3206108. URL https://doi.org/10.1109/TPAMI.2022.3206108.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-attention networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2735–2745, 2022. doi: 10.1109/CVPRW56347.2022.00309. URL https://doi.org/10.1109/CVPRW56347.2022.00309.

Qi Zhang, Jianhang Zhou, Jing He, Xiaodong Cun, Shaoning Zeng, and Bob Zhang. A shell dataset, for shell features extraction and recognition. *Scientific Data*, 6(1):226, 2019. doi: 10.1038/s41597-019-0230-3. URL https://doi.org/10.1038/s41597-019-0230-3.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Yang Zhang, Jun Yue, Aihuan Song, Shixiang Jia, and Zhenbo Li. A high-similarity shellfish recognition method based on convolutional neural network. *Information Processing in Agriculture*, 10 (2):149–163, 2023. doi: 10.1016/j.inpa.2022.05.009.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.07832.

Appendix

This appendix provides detailed dataset specifications and extended analyses of the classification, retrieval, captioning, and synthetic-sensitivity experiments. It includes the complete class inventory and partition statistics, three-seed held-out results, the 80-backbone descriptive scan, and complementary recognition and retrieval analyses. It also reports a specialized fine-grained baseline, a family-auxiliary comparison, and a shared-taxon transfer study. Further results cover the complete captioning evaluation, frozen zero-shot sensitivity checks, transformation-by-level results, marine-inspired RGB transformation tests, and a class-frequency analysis.

A Benchmark Specification and Data Partitions

A.1 Dataset Scope and Complete 100-Class Inventory

The benchmark contains 40,130 images from 100 shellfish classes. The collection consolidates records from public biodiversity and media repositories. Sources include iNaturalist (Van Horn et al., 2018), GBIF (Lane & Edwards, 2007), Wikimedia Commons (Vaidya et al., 2015), iDigBio (Singer et al., 2018), and the Encyclopedia of Life (Parr et al., 2014), among others. Table 2 lists the complete taxonomic inventory and the image allocation for each class.

The 100 classes were selected from a broader pool of marine shellfish taxa through taxonomic, image-availability, and visual-discriminability screening. Each retained species required an accepted species-level name and consistent identifiers in WoRMS (WoRMS Editorial Board, 2026), GBIF, and iNaturalist, together with sufficient rights-cleared public imagery to support quality filtering and cross-source deduplication. Candidate taxa also had to exhibit diagnostic external features in ordinary photographs. We excluded taxa whose separation depended mainly on anatomy, microscopy, genetic evidence, or geographic context, as well as candidates with unresolved name collisions or severe visual overlap with an existing class. The resulting inventory spans Bivalvia, Gastropoda, and Polyplacophora and defines the fixed 100-class label space used in the classification and retrieval experiments.

Table 2: Complete class inventory and image allocation. Scientific names follow the released taxonomy table.

Scientific name	Group	Total	Train	Val.	Test
Crassostrea virginica	Oyster	500	400	51	49
Ostrea edulis	Oyster	500	401	50	49
Mytilus californianus	Mussel	500	401	50	49
Geukensia demissa	Mussel	500	401	50	49
Perna canaliculus	Mussel	500	400	51	49
Argopecten irradians	Scallop	478	382	48	48
Crassadoma gigantea	Scallop	486	390	49	47
Tridacna maxima	Giant_Clam	481	385	48	48
Mercenaria mercenaria	Venus_Clam	479	383	48	48
Ruditapes philippinarum	Venus_Clam	491	394	48	49
Mya arenaria	Soft_Clam	486	389	49	48
Spisula solidissima	Surf_Clam	482	387	48	47
Donax variabilis	Coquina	465	372	47	46
Cerastoderma edule	Cockle	482	386	49	47
Dinocardium robustum	Cockle	455	364	46	45
Ensis leei	Razor_Clam	492	394	49	49
Anomia simplex	Jingle	462	370	46	46
Arca noae	Ark_Clam	374	299	38	37
Venus verrucosa	Venus_Clam	426	341	43	42
Cyrtopleura costata	Angelwing	451	361	45	45
Nuttallia obscurata	Tellin	490	392	49	49
Tivela stultorum	Pismo_Clam	466	374	47	45
Haliotis iris	Abalone	477	381	48	48
Patella vulgata	True_Limpet	476	381	48	47

Continued on next page

Table 2 (continued)

Scientific name	Group	Total	Train	Val.	Test
Lottia gigantea	True_Limpet	489	391	49	49
Diodora aspera	Keyhole_Limpet	440	352	45	43
Crepidula fornicata	Slipper_Snail	488	390	49	49
Littorina littorea	Periwinkle	477	381	48	48
Littoraria irrorata	Periwinkle	352	283	34	35
Bembicium nanum	Periwinkle	374	299	38	37
Tegula funebralis	Top_Snail	481	386	48	47
Calliostoma ligatum	Top_Snail	464	372	46	46
Megastraea undosa	Turban	485	389	49	47
Lunella undulata	Turban	416	334	42	40
Neverita duplicata	Moon_Snail	451	361	45	45
Aliger gigas	Conch	445	357	45	43
Strombus pugilis	Conch	425	340	43	42
Californiconus californicus	Cone	379	303	38	38
Hexaplex trunculus	Murex	464	371	47	46
Dicathais orbita	Murex	473	379	47	47
Nucella lapillus	Dogwhelk	470	376	47	47
Buccinum undatum	Whelk	479	383	48	48
Busycon carica	Whelk	469	375	47	47
Busycotypus canaliculatus	Whelk	412	330	41	41
Urosalpinx cinerea	Oyster_Drill	318	254	32	32
Ilyanassa obsoleta	Mudsnail	473	379	47	47
Tritia reticulata	Nassa	456	366	46	44
Oliva sayana	Olive	451	361	45	45
Callianax biplicata	Olive	459	367	47	45
Cabestana spengleri	Triton	458	366	46	46
Mopalia muscosa	Chiton	492	394	49	49
Cryptochiton stelleri	Chiton	487	389	49	49
Tonicella lineata	Chiton	488	390	49	49
Acanthopleura granulata	Chiton	483	387	49	47
Sypharochiton pelliserpentis	Chiton	496	398	50	48
Lunarca ovalis	Ark_Clam	500	400	50	50
Acanthocardia tuberculata	Cockle	500	400	50	50
Raeta plicatella	Duck_Clam	382	307	38	37
Crassula aequilatera	Triangle_Clam	371	298	37	36
Atrina zelandica	Pen_Shell	266	213	27	26
Pecten maximus	Scallop	370	297	36	37
Plicatula gibbosa	Kitten_Paw	253	204	25	24
Maoricolpus roseus	Turret_Snail	500	401	50	49
Nerita tessellata	Nerite	299	239	30	30
Alcithoe arabica	Volute	266	213	28	25
Katharina tunicata	Chiton	442	355	44	43
Sinistrofulgur sinistrum	Whelk	434	347	45	42
Megathura crenulata	Keyhole_Limpet	332	266	34	32
Scutus breviculus	Shield_Shell	296	237	30	29
Janthina janthina	Violet_Snail	354	283	36	35
Charonia lampas	Triton	271	218	27	26
Cypraea tigris	Cowrie	267	213	27	27
Melongena corona	Crown_Conch	351	281	35	35
Neoterebra dislocata	Auger	301	241	30	30
Cittarium pica	Top_Snail	276	221	28	27
Conus ebraeus	Cone	239	191	25	23
Cyphoma gibbosum	Ovulid	253	203	25	25
Noetia ponderosa	Ark_Clam	500	401	50	49
Mimachlamys varia	Scallop	491	393	49	49
Stramonita haemastoma	Rock_Shell	463	372	46	45
Monetaria annulus	Cowrie	317	253	32	32
Phorcus lineatus	Top_Snail	500	400	51	49
Austrovenus stutchburyi	Venus_Clam	500	400	51	49
Spondylus gaederopus	Thorny_Oyster	145	116	16	13
Clinocardium nuttallii	Cockle	500	401	50	49

Continued on next page

Table 2 (continued)

Scientific name	Group	Total	Train	Val.	Test
Mactra stultorum	Trough_Clam	500	401	50	49
Ceratostoma foliatum	Murex	320	256	32	32
Siliqua patula	Razor_Clam	352	283	34	35
Fissurella volcano	Keyhole_Limpet	348	278	36	34
Semicassis granulata	Helmet	201	162	19	20
Monoplex parthenopeus	Triton	226	182	22	22
Rapana venosa	Rapa_Whelk	267	214	26	27
Lutraria lutraria	Otter_Clam	251	202	24	25
Arctica islandica	Quahog	261	209	26	26
Sinum perspectivum	Baby_Ear	238	191	24	23
Bulla gouldiana	Bubble	152	123	15	14
Telescopium telescopium	Mangrove_Snail	156	126	16	14
Lambis lambis	Spider_Conch	132	106	13	13
Aporrhais pespelecani	Pelican_Foot	205	164	21	20
Fasciolaria tulipa	Tulip_Shell	189	151	19	19

A.2 Released Record Schema

The released metadata supports taxonomic interpretation, source tracing, rights checking, scene-based analysis, and reconstruction of the benchmark partitions. Table 3 summarizes the corresponding field groups. The public release retains source, attribution, and license fields for each image.

Table 3: Field groups in the released image-level metadata.

Field group	Purpose
Taxonomic identity	Benchmark class and scientific name used to connect each image with the 100-class inventory.
Source provenance	Public source platform and source-record references retained for traceability.
Rights	License, attribution, and rights-verification information associated with each image.
File integrity	File checksum and image dimensions used to verify the released image representation.
Scene context	Controlled or intertidal scene label used in the scene-stratified analysis.
Grouping	Acquisition or observation grouping used to keep related images in one partition.
Partition	Training, validation, or test assignment used by all benchmark tasks.

Controlled and intertidal are the only scene labels used for quantitative stratification; the marine transformations are separate synthetic diagnostics. Ecological observation, aquaculture, and post-harvest use are application contexts rather than separate metadata strata. Class-label verification was conducted by one reviewer familiar with shellfish taxonomy; independent second-expert agreement was not measured.

The release audit connects the candidate-image deduplication records, the selection manifest, and the released asset table. Strict duplicate detection identified 216 clusters containing 442 of the 57,171 candidate images and 234 duplicate pairs. We removed 226 noncanonical cluster members. An additional 754 pairs covering 878 image paths were retained for review. After accounting for seven paths shared with the strict clusters, this review removed 871 further paths, yielding 1,097 unique exclusions. Every exclusion is absent from the selected and released records. All 40,130 selected records map one-to-one to the released manifest through the stored checksum and source-identity fields.

Partitioning assigns every image from a recorded acquisition or observation group to one split. The grouping key uses a session, observation, or specimen identifier when available; otherwise, it uses the source platform and source record. Group disjointness applies to this key; photographer identity, geographic location, and source platform are not partition keys. Cross-platform identity resolution is limited to the available identifiers.

Table 4: Deduplication, release-linkage, and partition audit.

Audit quantity	Result
Candidate image inventory	57,171 images
Strict duplicate clusters	216 clusters; 442 members; 234 pairs
Noncanonical members removed from strict clusters	226 paths
Additional reviewed paths removed	871 paths
Excluded paths absent from the release	1,097 / 1,097
Selected-to-released record linkage	40,130 / 40,130
Acquisition or observation groups	30,879 groups; at most 2 images per group
Groups spanning more than one split	0

B Extended Classification Results and Breadth Scan

B.1 Training Configuration Registry

The training registry covers all combinations of 23 models and three seeds (41, 42, and 43), yielding 69 model–seed runs. The accompanying machine-readable registry records model construction, public initialization checkpoint and revision, input normalization, optimizer schedule, batch construction, and the 100-class output head. All runs use 100 epochs, cross-entropy without label smoothing, and per-epoch cosine annealing to zero. Each run has an effective batch size of 48, and validation Macro-F1 determines checkpoint selection. Training uses AMP FP16, whereas validation and test inference use FP32. Table 6 summarizes the optimizer assignments, and Table 7 lists the public model initializations.

The two optimizer recipes were selected using validation-only anchor experiments. MobileNetV3-Large served as the SGD anchor, and ViT-B/16 served as the AdamW anchor. For each anchor, six learning-rate and weight-decay candidates were evaluated for 30 epochs with seed 41. Validation Macro-F1 selected the recipe. The selected recipes were then assigned by the predefined model-family policy and held fixed for the formal 100-epoch runs. After the policy was fixed, there was no per-model search or test-set access. Table 5 reports the validation-only candidate grid. All reported benchmark values are outcomes of the evaluated fixed model configurations; they are not estimates of architecture-specific optima or capacity upper bounds.

Table 5: Validation-only optimizer anchor search with seed 41 and a 30-epoch horizon. Validation Macro-F1 values are percentages; bold marks the selected recipe for each anchor.

Anchor and optimizer	LR	WD	Val. Macro-F1
MobileNetV3-Large (SGD)			
	3×10^{-3}	10^{-4}	72.46
	3×10^{-3}	5×10^{-4}	73.56
	10^{-2}	10^{-4}	**74.97**
	10^{-2}	5×10^{-4}	73.97
	3×10^{-2}	10^{-4}	72.86
	3×10^{-2}	5×10^{-4}	69.95
ViT-B/16 (AdamW)			
	10^{-5}	0.01	69.83
	10^{-5}	0.05	69.85
	3×10^{-5}	0.01	73.27
	3×10^{-5}	0.05	**73.85**
	10^{-4}	0.01	73.20
	10^{-4}	0.05	72.92

Table 6: Training configuration registry for the 69 benchmark runs. Actual batch and gradient accumulation yield an effective batch size of 48 in every row.

Models	Optimizer	Runs	LR	WD	Batch × accum.
ResNet-50; ResNet-101; DenseNet-121; EfficientNetV2-S; RegNetY-8GF; MobileNetV3-Large; RepViT-M1.0	SGD ($m = 0.9$)	21	10^{-2}	10^{-4}	48×1
ConvNeXt-Tiny; ConvNeXtV2-Tiny; ViT-B/16; DeiT III-Small/16; Swin-Tiny; SwinV2-Tiny; PVTv2-B2; MaxViT-Tiny; EVA-02-Tiny; EfficientFormerV2-S2; MAE ViT-B/16; iBOT ViT-B/16; DINOv2 ViT-B/14; OpenCLIP ViT-B/16; SigLIP2 Base patch16-224	AdamW	45	3×10^{-5}	0.05	48×1
MambaOut-Tiny	AdamW	3	3×10^{-5}	0.05	24×2

Model references. ResNet-50/101 (He et al., 2016); DenseNet-121 (Huang et al., 2017); EfficientNetV2-S (Tan & Le, 2021); RegNetY-8GF (Radosavovic et al., 2020); MobileNetV3-Large (Howard et al., 2019); ConvNeXt-Tiny (Liu et al., 2022b); ConvNeXtV2-Tiny (Woo et al., 2023); ViT-B/16 (Dosovitskiy et al., 2021); DeiT III-Small/16 (Touvron et al., 2022); Swin-Tiny (Liu et al., 2021b); SwinV2-Tiny (Liu et al., 2022a); PVTv2-B2 (Wang et al., 2022); MaxViT-Tiny (Tu et al., 2022); EVA-02-Tiny (Fang et al., 2024); EfficientFormerV2-S2 (Li et al., 2023b); MAE ViT-B/16 (He et al., 2022); iBOT ViT-B/16 (Zhou et al., 2022); DINOv2 ViT-B/14 (Oquab et al., 2024); OpenCLIP ViT-B/16 (Cherti et al., 2023); SigLIP2 Base patch16-224 (Tschannen et al., 2025); MambaOut-Tiny (Yu & Wang, 2025); RepViT-M1.0 (Wang et al., 2024).

Table 7: Public implementation identifiers, initialization revisions, and retrieval dimensions for the 23-model benchmark.

Model	Implementation identifier	Public initialization revision	Retrieval dim.
ResNet-50	`torchvision.models.resnet50`	`torchvision==0.16.2:ResNet50_Weights.IMAGENET1K_V2`	2048
ResNet-101	`torchvision.models.resnet101`	`torchvision==0.16.2:ResNet101_Weights.IMAGENET1K_V2`	2048
DenseNet-121	`torchvision.models.densenet121`	`torchvision==0.16.2:DenseNet121_Weights.IMAGENET1K_V1`	1024
EfficientNetV2-S	`torchvision.models.efficientnet_v2_s`	`torchvision==0.16.2:EfficientNet_V2_S_Weights.IMAGENET1K_V1`	1280
RegNetY-8GF	`torchvision.models.regnet_y_8gf`	`torchvision==0.16.2:RegNet_Y_8GF_Weights.IMAGENET1K_V2`	2016
MobileNetV3-Large	`torchvision.models.mobilenet_v3_large`	`torchvision==0.16.2:MobileNet_V3_Large_Weights.IMAGENET1K_V2`	1280
ConvNeXt-Tiny	`torchvision.models.convnext_tiny`	`torchvision==0.16.2:ConvNeXt_Tiny_Weights.IMAGENET1K_V1`	768
ConvNeXtV2-Tiny	`convnextv2_tiny.fcmae_ft_in1k`	`508caf1bc24f`	768
ViT-B/16	`torchvision.models.vit_b_16`	`torchvision==0.16.2:ViT_B_16_Weights.IMAGENET1K_V1`	768
DeiT III-Small/16	`deit3_small_patch16_224.fb_in1k`	`f3f500f06a4e`	384
Swin-Tiny	`torchvision.models.swin_t`	`torchvision==0.16.2:Swin_T_Weights.IMAGENET1K_V1`	768
SwinV2-Tiny	`torchvision.models.swin_v2_t`	`torchvision==0.16.2:Swin_V2_T_Weights.IMAGENET1K_V1`	768

Continued on next page

Table 7 (continued)

Model	Implementation identifier	Public initialization revision	Retrieval dim.
PVTv2-B2	`pvt_v2_b2.in1k`	`8dca9e485d6a`	512
MaxViT-Tiny	`torchvision.models.maxvit_t`	`torchvision==0.16.2:MaxVit_T_Weights.IMAGENET1K_V1`	512
EVA-02-Tiny	`eva02_tiny_patch14_224.mim_in22k`	`e0a5e8afdf15`	192
EfficientFormerV2-S2	`efficientformerv2_s2.snap_dist_in1k`	`1c56a7635500`	288
MAE ViT-B/16	`vit_base_patch16_224.mae`	`a64dffd137c0`	768
iBOT ViT-B/16	`vit_base_patch16_224.ibot_teacher`	`da316d82636a`	768
DINOv2 ViT-B/14	`vit_base_patch14_dinov2.lvd142m`	`4685c99dabff`	768
OpenCLIP ViT-B/16	`open-clip-torch==2.24.0 ViT-B-16::openai`	`openai:5806e77cd80f`	512
SigLIP2 Base patch16-224	`google/siglip2-base-patch16-224`	`75de2d55ec2d`	768
MambaOut-Tiny	`mambaout_tiny.in1k`	`d080941663a6`	2304
RepViT-M1.0	`repvit_m1_0`	`v1.0:repvit_m1_0_distill_450e_timm.pth`	448

Git-style hashes in Table 7 are abbreviated to 12 characters; the machine-readable registry retains the complete revisions. Complete image-level manifests are unavailable for several public pretraining corpora, leaving pretraining overlap unresolved for some encoders. The reported results describe performance on the fixed ShellfishNet evaluation. Section C.7 gives a provider-level analysis for BioCLIP 2.

B.2 Three-Seed Held-Out Results for the 23-Model Benchmark

Each of the 23 models was trained with random seeds 41, 42, and 43 under the same model-specific configuration, yielding 69 runs. For each model and metric, Tables 8 and 9 report the arithmetic mean and sample standard deviation across the three held-out evaluations. Figure 5 retains the seed-specific Top-1 and class-macro mAP@R values so that classification and retrieval dispersion can be inspected directly.

Table 8: Complete three-seed held-out classification results. Values are mean ± sample SD over seeds 41, 42, and 43, in percentage points; column-best means are bold.

Model	Top-1	Top-5	Macro-F1	Balanced Acc.
ResNet-50	79.82±0.05	93.92±0.08	79.61±0.12	79.47±0.14
ResNet-101	80.99±0.10	94.72±0.20	80.87±0.12	80.54±0.10
DenseNet-121	77.67±0.47	92.71±0.28	77.30±0.26	77.23±0.27
EfficientNetV2-S	82.35±0.29	94.60±0.05	82.31±0.40	82.16±0.42
RegNetY-8GF	83.33±0.47	95.54±0.15	83.10±0.50	83.03±0.47
MobileNetV3-Large	74.91±0.87	91.38±0.26	74.37±0.91	74.33±0.99
ConvNeXt-Tiny	82.10±0.42	95.16±0.19	82.01±0.41	81.75±0.31
ConvNeXtV2-Tiny	78.20±0.32	93.69±0.13	77.80±0.50	77.52±0.46
ViT-B/16	75.59±0.73	91.44±0.34	75.61±0.80	75.29±0.91
DeiT III-Small/16	74.76±0.81	92.13±0.10	74.68±0.92	74.19±0.94
Swin-Tiny	81.67±0.43	94.76±0.27	81.46±0.57	81.25±0.62
SwinV2-Tiny	49.08±6.64	76.58±5.28	45.93±6.82	46.97±7.05
PVTv2-B2	79.63±0.85	94.00±0.47	79.32±1.04	79.08±1.03
MaxViT-Tiny	**84.39±0.22**	**95.62±0.11**	**84.40±0.29**	**84.26±0.44**
EVA-02-Tiny	70.87±0.28	88.79±0.16	70.44±0.25	70.17±0.29
EfficientFormerV2-S2	77.40±0.31	92.77±0.24	76.88±0.38	76.86±0.31
MAE ViT-B/16	72.97±0.33	90.98±0.28	72.85±0.21	72.35±0.30
iBOT ViT-B/16	79.59±1.40	93.91±0.66	79.61±1.55	79.22±1.43
DINOv2 ViT-B/14	82.71±0.94	94.87±0.09	82.75±1.00	82.31±1.01
OpenCLIP ViT-B/16	75.12±0.13	91.19±0.03	74.93±0.24	74.56±0.11
SigLIP2 Base patch16-224	78.58±0.32	92.73±0.12	78.36±0.67	77.94±0.71
MambaOut-Tiny	76.93±0.38	92.93±0.23	76.75±0.42	76.55±0.37
RepViT-M1.0	51.37±2.63	77.24±1.85	50.30±2.87	50.41±2.96

Model references. ResNet-50/101 (He et al., 2016); DenseNet-121 (Huang et al., 2017); EfficientNetV2-S (Tan & Le, 2021); RegNetY-8GF (Radosavovic et al., 2020); MobileNetV3-Large (Howard et al., 2019); ConvNeXt-Tiny (Liu et al., 2022b); ConvNeXtV2-Tiny (Woo et al., 2023); ViT-B/16 (Dosovitskiy et al., 2021); DeiT III-Small/16 (Touvron et al., 2022); Swin-Tiny (Liu et al., 2021b); SwinV2-Tiny (Liu et al., 2022a); PVTv2-B2 (Wang et al., 2022); MaxViT-Tiny (Tu et al., 2022); EVA-02-Tiny (Fang et al., 2024); EfficientFormerV2-S2 (Li et al., 2023b); MAE ViT-B/16 (He et al., 2022); iBOT ViT-B/16 (Zhou et al., 2022); DINOv2 ViT-B/14 (Oquab et al., 2024); OpenCLIP ViT-B/16 (Cherti et al., 2023); SigLIP2 Base patch16-224 (Tschannen et al., 2025); MambaOut-Tiny (Yu & Wang, 2025); RepViT-M1.0 (Wang et al., 2024).

Table 9: Complete three-seed held-out retrieval results. Values are mean ± sample SD over seeds 41, 42, and 43, in percentage points; column-best means are bold. mAP@R denotes the class-macro average (Musgrave et al., 2020).

Model	R@1	R@5	mAP@R
ResNet-50	77.75±0.23	87.70±0.42	46.61±0.93
ResNet-101	79.55±0.40	88.25±0.24	50.81±0.29
DenseNet-121	74.63±0.25	86.64±0.98	39.08±2.89
EfficientNetV2-S	82.06±0.37	87.70±0.03	69.74±0.35
RegNetY-8GF	81.83±0.42	**89.96±0.15**	51.09±0.87
MobileNetV3-Large	72.56±0.78	85.67±0.54	33.42±0.13
ConvNeXt-Tiny	81.79±0.28	87.19±0.24	67.44±0.51
ConvNeXtV2-Tiny	77.77±0.40	84.04±0.63	67.98±0.30
ViT-B/16	75.08±0.88	80.67±0.78	65.77±0.56
DeiT III-Small/16	73.27±0.58	81.88±0.27	52.53±3.73
Swin-Tiny	81.15±0.59	88.63±0.28	54.30±0.35
SwinV2-Tiny	46.36±1.68	72.48±1.73	4.04±0.44
PVTv2-B2	79.32±1.31	86.33±0.65	56.62±7.40
MaxViT-Tiny	**84.26±0.66**	87.91±0.31	**77.83±0.59**
EVA-02-Tiny	69.18±0.34	77.72±0.22	56.22±1.02
EfficientFormerV2-S2	77.43±0.86	84.50±0.35	60.20±0.63
MAE ViT-B/16	71.46±0.17	80.76±0.43	49.03±0.74
iBOT ViT-B/16	78.78±0.91	87.49±2.41	47.39±11.51
DINOv2 ViT-B/14	82.95±0.79	87.79±0.54	69.66±0.63
OpenCLIP ViT-B/16	74.20±0.33	80.95±0.32	65.04±0.16
SigLIP2 Base patch16-224	78.15±0.25	84.46±0.23	64.37±1.43
MambaOut-Tiny	75.69±0.45	85.51±0.21	51.07±2.18
RepViT-M1.0	34.73±2.30	59.68±1.92	4.21±0.76

Model references. ResNet-50/101 (He et al., 2016); DenseNet-121 (Huang et al., 2017); EfficientNetV2-S (Tan & Le, 2021); RegNetY-8GF (Radosavovic et al., 2020); MobileNetV3-Large (Howard et al., 2019); ConvNeXt-Tiny (Liu et al., 2022b); ConvNeXtV2-Tiny (Woo et al., 2023); ViT-B/16 (Dosovitskiy et al., 2021); DeiT III-Small/16 (Touvron et al., 2022); Swin-Tiny (Liu et al., 2021b); SwinV2-Tiny (Liu et al., 2022a); PVTv2-B2 (Wang et al., 2022); MaxViT-Tiny (Tu et al., 2022); EVA-02-Tiny (Fang et al., 2024); EfficientFormerV2-S2 (Li et al., 2023b); MAE ViT-B/16 (He et al., 2022); iBOT ViT-B/16 (Zhou et al., 2022); DINOv2 ViT-B/14 (Oquab et al., 2024); OpenCLIP ViT-B/16 (Cherti et al., 2023); SigLIP2 Base patch16-224 (Tschannen et al., 2025); MambaOut-Tiny (Yu & Wang, 2025); RepViT-M1.0 (Wang et al., 2024).

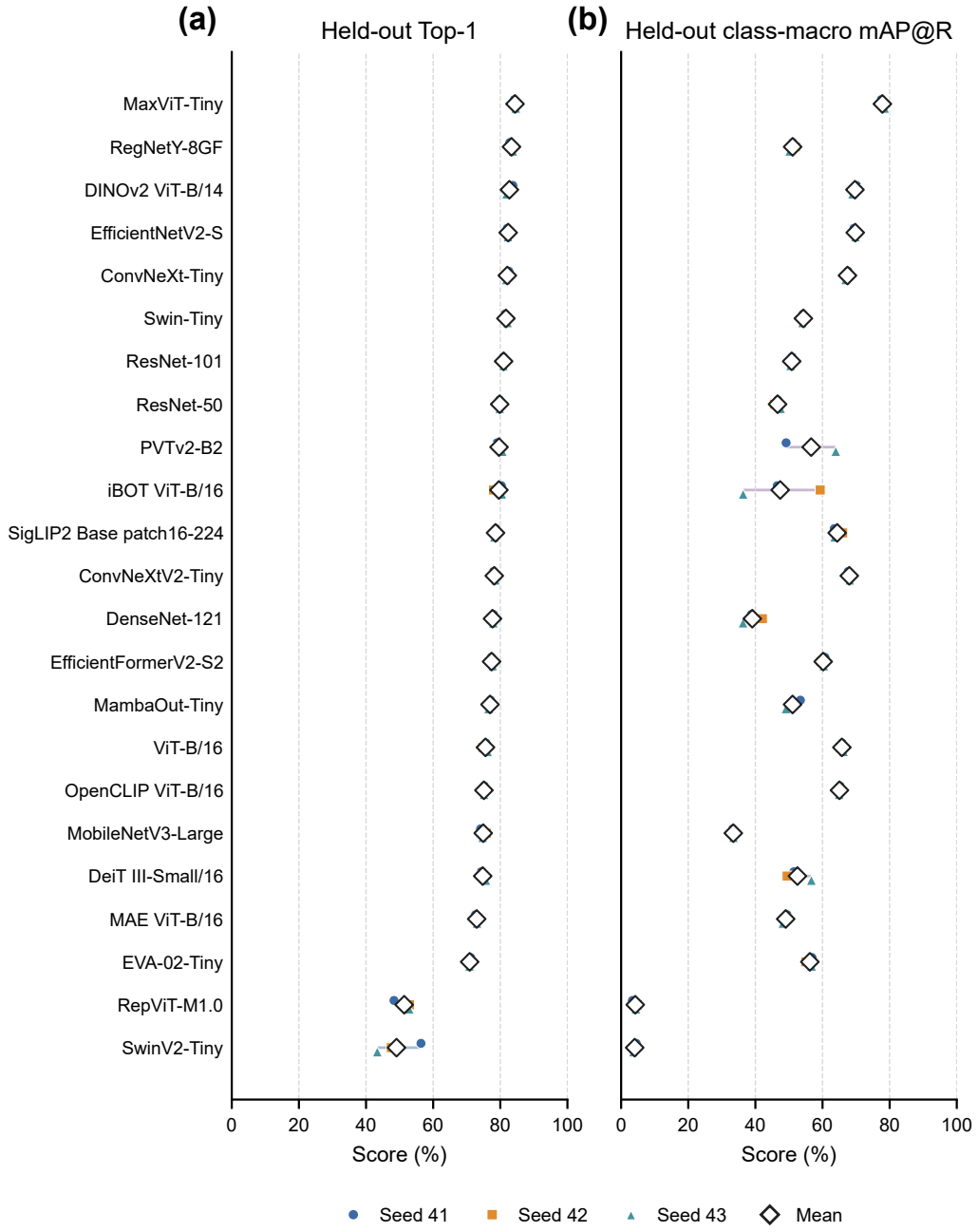


Figure 5: Seed-specific held-out Top-1 and class-macro mAP@R results for 23 models. Circles, squares, and triangles denote seeds 41, 42, and 43; horizontal segments span the observed minimum and maximum, and hollow diamonds mark arithmetic means. Rows are ordered by mean Top-1. Both panels use the full 0–100 percentage scale.

B.3 Specialized Fine-Grained Baseline

We evaluate HERBS with a Swin-L/384 backbone as a specialized fine-grained baseline. The model uses the released selector, feature-pyramid, and combiner design. It is trained for 80 epochs with SGD-Nesterov, an effective batch size of 32, a maximum learning rate of 5×10^{-4}, and weight decay of 3×10^{-4}. Validation Macro-F1 selects the epoch-56 checkpoint, which is evaluated on

the held-out test partition in FP32. Table 10 reports 91.66% Top-1 accuracy and 91.93% Macro-F1. This single-run result remains separate from the three-run, 224-pixel benchmark ranking.

Table 10: Held-out classification results for the specialized fine-grained baseline. All metrics are percentages from one run.

Model	Runs	Input	Epochs	Top-1	Top-5	Macro-F1	Bal. acc.
HERBS (Swin-L/384)	1	384	80	91.66	98.23	91.93	91.71

Model references. HERBS (Chou et al., 2023); released implementation (Chou, 2023); Swin Transformer (Liu et al., 2021b).

B.4 Complete 80-Backbone Descriptive Scan

The breadth scan evaluates 80 backbones on the same 100-class task. All runs use seed 41, 30 epochs, 224×224 inputs, ImageNet normalization, and an effective batch size of 48. Optimization uses SGD with an initial learning rate of 0.01, weight decay of 0.0005, and cosine annealing. Figure 6 summarizes the model-size tradeoff and displays the complete Top-1 ordering across all 80 backbones. Validation Macro-F1 selects the checkpoint within the 30-epoch horizon. The accompanying machine-readable registry reports each backbone's implementation identifier, pretrained initialization, micro-batch size, gradient-accumulation setting, and selected epoch. The scan describes architectural breadth rather than seed-stable rankings. Parameter counts support direct cross-model comparison, while timing and throughput retain their run-specific hardware context in Table 11.

Table 11: Computational measurements for the 80-backbone descriptive scan. Paired entries report parameter count/peak memory, training throughput/duration, and test throughput/duration; timing and throughput reflect the hardware used for each run.

Model	Parameters Peak memory (M / MiB)	Training rate / time (images/s / min)	Testing rate / time (images/s / s)
DaViT-Tiny	27.67 / 2856	244.9 / 65.6	555.2 / 7.13
DLA-34	15.28 / 1482	239.8 / 67.0	575.1 / 6.88
HRNet-W18	19.45 / 2506	138.7 / 115.9	520.8 / 7.60
ConvMixer-768/32	20.42 / 5908	62.8 / 256.0	124.3 / 31.84
ResNeSt-26D	15.23 / 2716	235.2 / 68.3	553.7 / 7.15
FBNetV3-B	6.81 / 2830	244.4 / 65.8	531.0 / 7.45
ConvNeXt-Small	49.53 / 2144	275.8 / 58.3	686.5 / 5.77
Sequencer2D-L	53.95 / 3268	110.0 / 146.2	377.3 / 10.49
XCiT-S12/8	25.87 / 6698	101.0 / 159.1	179.2 / 22.09
Res2Net-50	23.86 / 2366	244.9 / 65.6	559.9 / 7.07
EfficientNet-B0	4.14 / 3090	242.8 / 66.2	560.7 / 7.06
PoolFormer-S12	11.45 / 2624	236.0 / 68.1	540.7 / 7.32
GCViT-XTiny	19.52 / 3322	198.0 / 81.2	514.5 / 7.69
PiT-Tiny	4.64 / 848	243.5 / 66.0	574.4 / 6.89
PiT-Small	23.00 / 1292	235.7 / 68.2	545.3 / 7.26
RegNetY-1.6GF	10.40 / 2314	233.8 / 68.7	548.8 / 7.21
gMLP-S16	19.19 / 2870	245.8 / 65.4	554.9 / 7.13
RegNetY-8GF (torchvision)	37.36 / 2126	314.1 / 51.2	658.2 / 6.01
ReXNet-1.0	3.64 / 3264	241.2 / 66.6	555.8 / 7.12
MLP-Mixer-B/16	59.19 / 1862	316.1 / 50.8	666.1 / 5.94
GhostNetV2-1.0	5.00 / 1982	244.7 / 65.7	572.2 / 6.92
RepGhostNet-1.0	2.92 / 1410	231.9 / 69.3	545.1 / 7.26
MnasNet-Small	0.88 / 974	246.9 / 65.1	580.6 / 6.82
MambaOut-Femto	6.27 / 2302	244.6 / 65.7	550.6 / 7.19
MambaVision-T	31.22 / 1604	302.9 / 53.1	650.0 / 6.09
MobileOne-S1	3.67 / 2444	243.6 / 66.0	552.7 / 7.16
PP-LCNet-1.0	1.80 / 1292	231.6 / 69.4	548.4 / 7.22

Continued on next page

Table 11 (continued)

Model	Parameters Peak memory (M / MiB)	Training rate / time (images/s / min)	Testing rate / time (images/s / s)
RepGhostNet-0.5	1.16 / 814	247.2 / 65.0	562.5 / 7.04
MobileNetV4-Conv-S	2.62 / 988	244.8 / 65.7	567.7 / 6.97
HardCoRe-NAS-A	4.11 / 1850	243.2 / 66.1	582.7 / 6.79
EfficientFormer-L1	11.48 / 2332	233.7 / 68.8	553.8 / 7.15
EfficientViT-M0	2.18 / 658	246.3 / 65.2	583.9 / 6.78
EfficientFormerV2-S1	5.78 / 2786	244.9 / 65.6	556.2 / 7.12
InceptionNeXt-Tiny	25.98 / 2446	245.6 / 65.4	555.7 / 7.12
ConvFormer-S18	24.93 / 4284	231.2 / 69.5	508.7 / 7.78
EfficientViT-B1	7.66 / 2616	246.5 / 65.2	569.4 / 6.95
VOLO-D1	25.94 / 2596	238.6 / 67.4	455.0 / 8.70
FocalNet-Small	49.20 / 2296	193.7 / 83.0	463.9 / 8.53
TinyViT-11M	10.59 / 3732	244.4 / 65.8	561.1 / 7.05
TinyViT-21M	20.68 / 3268	231.7 / 69.4	544.5 / 7.27
PVTv2-B0	3.44 / 1844	245.9 / 65.4	549.2 / 7.21
ESE-VoVNet-19b-DW	5.62 / 2970	245.1 / 65.6	562.7 / 7.03
Swin-Small	48.91 / 2398	159.2 / 100.9	656.9 / 6.03
Twins-PCPVT-S	23.64 / 2740	244.6 / 65.7	545.6 / 7.25
Twins-SVT-S	23.60 / 2268	232.6 / 69.1	537.9 / 7.36
Visformer-Small	39.53 / 1232	312.9 / 51.4	656.1 / 6.03
MViTv2-Tiny	23.48 / 4540	207.7 / 77.4	441.6 / 8.96
MViTv2-Small	34.18 / 5426	140.2 / 114.6	331.4 / 11.94
GCViT-Tiny	27.76 / 4568	127.1 / 126.4	374.3 / 10.57
CoAtNet-1	41.03 / 3362	306.7 / 52.4	651.2 / 6.08
Hiera-Tiny	27.22 / 2940	242.0 / 66.4	556.0 / 7.12
Hiera-Small	34.32 / 3272	245.3 / 65.5	494.2 / 8.01
DeiT-Small/16	21.70 / 1420	243.5 / 66.0	575.1 / 6.88
ViT-Small/16	21.70 / 1420	245.5 / 65.5	575.2 / 6.88
ConViT-Tiny	5.54 / 2506	238.2 / 67.5	543.3 / 7.28
CAFormer-S18	24.50 / 3924	247.0 / 65.1	522.5 / 7.58
ViT-Tiny/16	5.54 / 1150	243.7 / 65.9	574.1 / 6.89
ResNet-50	23.71 / 2090	244.8 / 65.7	561.7 / 7.05
ResNet-101	42.71 / 1676	298.2 / 53.9	639.7 / 6.19
DenseNet-121	7.06 / 3930	243.6 / 66.0	557.9 / 7.09
EfficientNetV2-S	20.31 / 2656	242.0 / 66.4	543.8 / 7.28
RegNetY-8GF (timm)	37.57 / 2262	303.4 / 53.0	633.1 / 6.25
MobileNetV3-Large	4.33 / 1774	241.5 / 66.6	551.5 / 7.18
ConvNeXt-Tiny	27.90 / 2412	241.9 / 66.4	560.7 / 7.06
ConvNeXtV2-Tiny	27.94 / 3780	219.2 / 73.3	502.2 / 7.88
ViT-B/16	85.88 / 1986	290.0 / 55.4	575.6 / 6.88
DeiT III-Small/16	21.71 / 1518	245.1 / 65.6	563.2 / 7.03
Swin-Tiny	27.60 / 2928	241.8 / 66.5	556.9 / 7.11
RegNetY-3.2GF	18.07 / 2112	240.8 / 66.7	541.5 / 7.31
PVTv2-B2	24.90 / 3228	233.6 / 68.8	523.3 / 7.56
MaxViT-Tiny	30.46 / 3296	78.4 / 205.1	335.1 / 11.81
EVA-02-Tiny	5.52 / 1700	245.3 / 65.5	580.7 / 6.82
EfficientFormerV2-S2	12.19 / 4028	243.8 / 65.9	450.4 / 8.79
MAE ViT-B/16	85.88 / 1944	222.7 / 72.2	267.9 / 14.77
DPN-68b	11.86 / 3020	244.3 / 65.8	534.9 / 7.40
GeRNet-S	6.44 / 1060	232.7 / 69.1	544.8 / 7.27
RepVGG-A1	12.94 / 1654	242.7 / 66.2	551.8 / 7.17
SigLIP2 Base patch16-224	92.96 / 2064	183.1 / 87.8	310.9 / 12.73
MambaOut-Tiny	24.47 / 3458	243.4 / 66.0	539.5 / 7.34
RepViT-M1.0	6.49 / 2728	234.6 / 68.5	548.4 / 7.22

Model references. DaViT-Tiny (Ding et al., 2022); DLA-34 (Yu et al., 2018); HRNet-W18 (Wang et al., 2021); ConvMixer-768/32 (Trockman & Kolter, 2023); ResNeSt-26D (Zhang et al., 2022); FBNetV3-B (Dai et al., 2021a); ConvNeXt-Small/Tiny (Liu et al., 2022b); Sequencer2D-L (Tatsunami & Taki, 2022); XCiT-S12/8 (El-Nouby et al., 2021); Res2Net-50 (Gao et al., 2021); EfficientNet-B0 (Tan & Le, 2019);

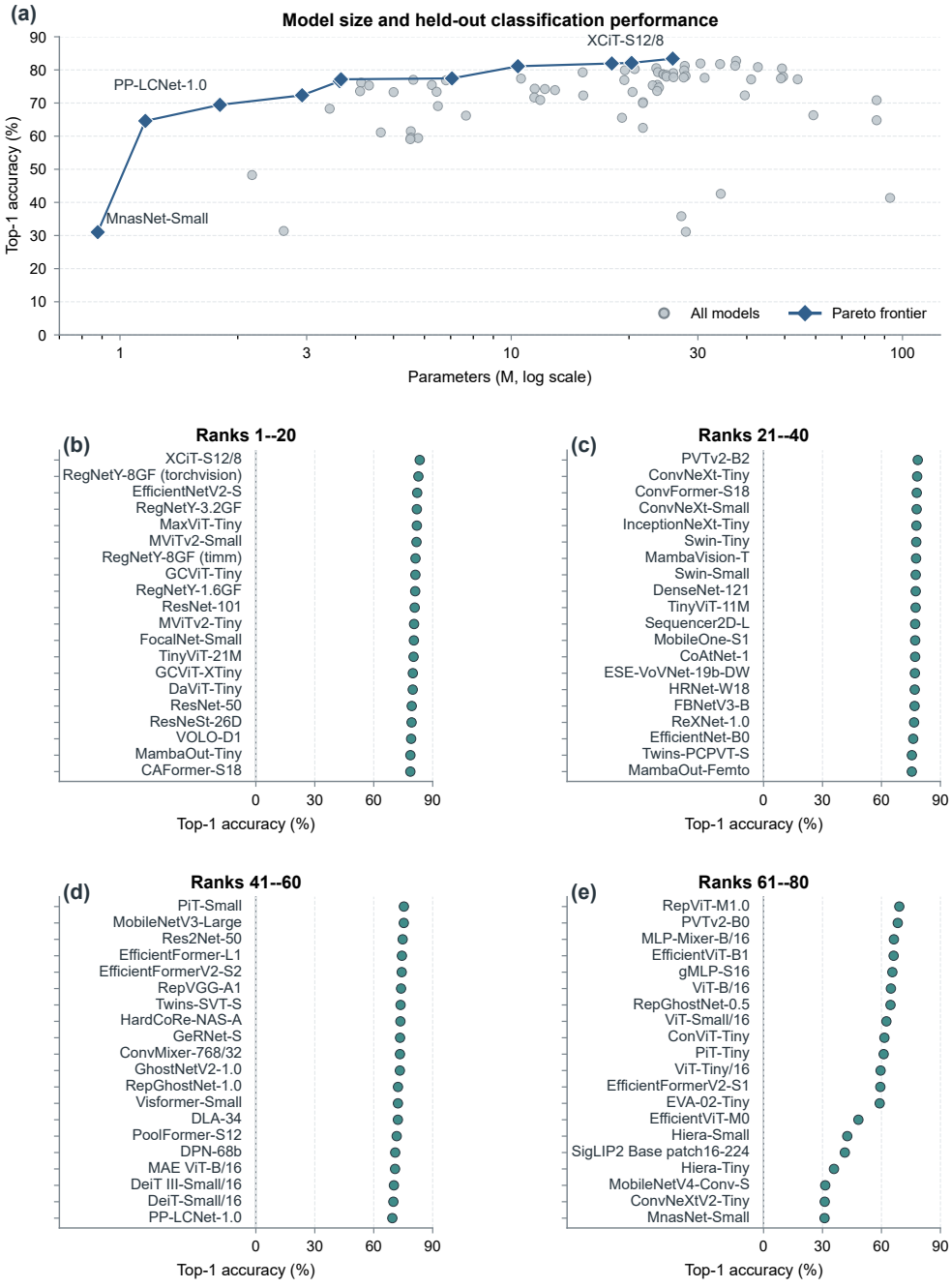


Figure 6: Complete 80-backbone descriptive scan. Panel (a) relates parameter count to held-out Top-1 accuracy and marks the Pareto frontier. Panels (b)–(e) show the observed Top-1 ordering for all 80 models in groups of 20. Each model uses one seed and a 30-epoch training horizon.

PoolFormer-S12 (Yu et al., 2022); GCViT-XTiny/Tiny (Hatamizadeh et al., 2023); PiT-Tiny/Small (Heo et al., 2021); RegNetY-1.6GF/3.2GF/8GF (Radosavovic et al., 2020); gMLP-S16 (Liu et al., 2021a); ReXNet-1.0 (Han et al., 2021); MLP-Mixer-B/16 (Tolstikhin et al., 2021); GhostNetV2-1.0 (Tang et al., 2022); RepGhostNet-0.5/1.0 (Chen et al., 2022); MnasNet-Small (Tan et al., 2019); MambaOut-Femto/Tiny (Yu & Wang, 2025); MambaVision-T (Hatamizadeh & Kautz, 2025); MobileOne-S1 (Vasu et al., 2023); PP-LCNet-1.0 (Cui et al., 2021); MobileNetV4-Conv-S (Qin et al., 2024); HardCoRe-NAS-A (Nayman et al., 2021); EfficientFormer-L1 (Li et al., 2022c); EfficientViT-M0 (Cai et al., 2023); EfficientFormerV2-S1/S2 (Li et al., 2023b); InceptionNeXt-Tiny (Yu et al., 2024b); ConvFormer-S18/CAFormer-S18 (Yu et al., 2024a); EfficientViT-B1 (Liu et al., 2023); VOLO-D1 (Yuan et al., 2023); FocalNet-Small (Yang et al., 2022); TinyViT-11M/21M (Wu et al., 2022); PVTv2-B0/B2 (Wang et al., 2022); ESE-VoVNet-19b-DW (Lee et al., 2019); Swin-Small/Tiny (Liu et al., 2021b); Twins-PCPVT-S/SVT-S (Chu et al., 2021); Visformer-Small (Chen et al., 2021); MViTv2-Tiny/Small (Li et al., 2022b); CoAtNet-1 (Dai et al., 2021b); Hiera-Tiny/Small (Ryali et al., 2023); DeiT-Small/16 (Touvron et al., 2021); ViT-Tiny/16, ViT-Small/16, and ViT-B/16 (Dosovitskiy et al., 2021); ConViT-Tiny (d'Ascoli et al., 2021); ResNet-50/101 (He et al., 2016); DenseNet-121 (Huang et al., 2017); EfficientNetV2-S (Tan & Le, 2021); MobileNetV3-Large (Howard et al., 2019); ConvNeXtV2-Tiny (Woo et al., 2023); DeiT III-Small/16 (Touvron et al., 2022); MaxViT-Tiny (Tu et al., 2022); EVA-02-Tiny (Fang et al., 2024); MAE ViT-B/16 (He et al., 2022); DPN-68b (Chen et al., 2017); GeRNet-S (Lin et al., 2020); RepVGG-A1 (Ding et al., 2021); SigLIP2 Base patch16-224 (Tschannen et al., 2025); RepViT-M1.0 (Wang et al., 2024).

B.5 Per-Class Behavior Across the Breadth Scan

Figure 7 aggregates the 8,000 model–class records into a class-centered view. It relates test support to median class F1, summarizes the distribution of class-level median Recall and F1, and retains the median and interquartile F1 range for all 100 classes.

B.6 Calibration and Inference Profiling

Two five-model analyses complement accuracy-based comparisons. Calibration uses held-out classification predictions and 15 equal-width confidence bins. Expected calibration error (ECE) and maximum calibration error (MCE) follow the standard bin-based definitions (Guo et al., 2017). Selective prediction uses the risk–coverage formulation (Geifman & El-Yaniv, 2017), and the area under this curve (AURC) provides a scalar summary (Ding et al., 2020). Tables 12 and 13 report these two views separately.

Inference profiling uses the seed-41, 100-class checkpoint for each model; RepViT-M1.0 uses timm's fused inference graph. Measurements were collected on one NVIDIA GeForce RTX 4080 SUPER in FP32 with 224×224 RGB tensors. Each timing block used five warm-up iterations followed by 20 synchronized iterations. Table 14 reports parameter count, FLOPs, and peak CUDA memory; Table 15 reports batch-one and batch-32 timing. MobileNetV3-Large provides the highest batch-32 throughput in this profile, while DINOv2 ViT-B/14 has the largest operation count and memory footprint.

Table 12: Calibration measures for five seed-41 checkpoints on held-out data. Top-1, confidence, ECE, and MCE are percentages.

Model	Top-1	Conf.	NLL	ECE	MCE
ResNet-50	79.76	89.42	0.910	9.65	29.06
ConvNeXtV2-Tiny	77.97	95.99	1.736	18.07	51.58
SwinV2-Tiny	56.37	29.23	2.025	27.13	46.50
MaxViT-Tiny	84.21	95.27	0.878	11.06	39.28
DINOv2 ViT-B/14	83.73	95.42	1.085	11.69	38.88

Model references. ResNet-50 (He et al., 2016); ConvNeXtV2-Tiny (Woo et al., 2023); SwinV2-Tiny (Liu et al., 2022a); MaxViT-Tiny (Tu et al., 2022); DINOv2 ViT-B/14 (Oquab et al., 2024).

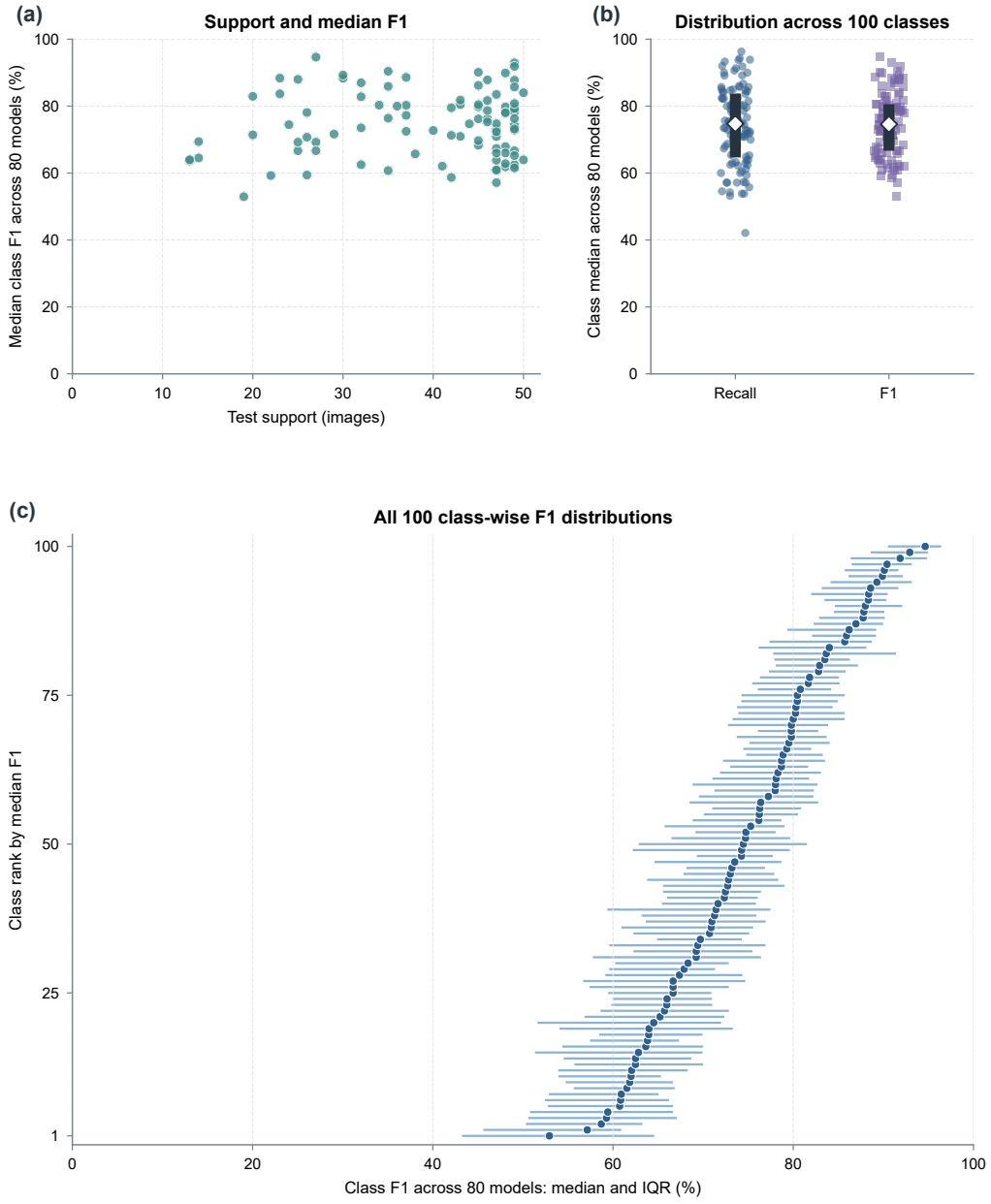


Figure 7: Per-class behavior across the 80-backbone scan. Panel (a) relates test support to median class F1. Panel (b) shows the distributions of all 100 class-level median Recall and F1 values; thick intervals denote IQRs and hollow diamonds denote medians. Panel (c) shows median F1 and IQR for every class, ranked by median F1.

Table 13: Selective-prediction measures for the same five seed-41 checkpoints. Risk and accuracy at 95% coverage are percentages. AURC is reported as a raw, unitless area.

Model	Risk@95	Acc.@95	AURC
ResNet-50	17.28	82.72	0.047
ConvNeXtV2-Tiny	19.17	80.83	0.058
SwinV2-Tiny	41.53	58.47	0.205
MaxViT-Tiny	12.79	87.21	0.033
DINOv2 ViT-B/14	13.24	86.76	0.033

Model references. ResNet-50 (He et al., 2016); ConvNeXtV2-Tiny (Woo et al., 2023); SwinV2-Tiny (Liu et al., 2022a); MaxViT-Tiny (Tu et al., 2022); DINOv2 ViT-B/14 (Oquab et al., 2024).

Table 14: Model complexity and peak CUDA memory for FP32 inference on an RTX 4080 SUPER with 224×224 RGB inputs. RepViT-M1.0 uses timm's fused inference graph in this profile, whereas the 80-backbone scan retains its two-head distillation graph.

Model	Params (M)	FLOPs (G)	Peak MB
ResNet-50	23.71	8.17	453.44
MobileNetV3-Large	4.33	0.43	263.69
EfficientFormerV2-S2	12.19	2.50	298.19
DINOv2 ViT-B/14	85.80	43.90	642.96
RepViT-M1.0	6.41	2.21	246.75

Model references. ResNet-50 (He et al., 2016); MobileNetV3-Large (Howard et al., 2019); EfficientFormerV2-S2 (Li et al., 2023b); DINOv2 ViT-B/14 (Oquab et al., 2024); RepViT-M1.0 (Wang et al., 2024).

Table 15: Synchronized FP32 inference timing on an RTX 4080 SUPER. Values are medians across three repetitions, each averaged over 20 synchronized timed iterations after five warm-up iterations.

Model	B1 lat. (ms)	B1 img/s	B32 lat. (ms)	B32 img/s
ResNet-50	6.03	165.8	17.29	1850.6
MobileNetV3-Large	6.31	158.4	6.59	4857.2
EfficientFormerV2-S2	22.38	44.7	28.47	1123.8
DINOv2 ViT-B/14	5.57	179.4	66.13	483.9
RepViT-M1.0	7.05	141.8	7.06	4534.9

Model references. ResNet-50 (He et al., 2016); MobileNetV3-Large (Howard et al., 2019); EfficientFormerV2-S2 (Li et al., 2023b); DINOv2 ViT-B/14 (Oquab et al., 2024); RepViT-M1.0 (Wang et al., 2024).

C Complementary Recognition and Retrieval Analyses

C.1 Complementary Retrieval Measures

The retrieval protocol uses the 3,958 test images as queries and the 32,143 training images as the gallery. For every classifier, the retrieval vector is the tensor supplied directly to the trained 100-class head, captured before the head operation and then L2-normalized. Table 9 reports three-seed R@1, R@5, and class-macro mAP@R. Panels (a)–(d) of Figure 8 complement those aggregate results with ordinary mAP@R (Musgrave et al., 2020) and pairwise discrimination measures from a separate fixed held-out diagnostic. For each query, a deterministic image-ID-based procedure fixes 10 same-class positive and 10 different-class negative gallery images, yielding $3{,}958 \times 20 = 79{,}160$ pairs with equal positive and negative prevalence. This diagnostic is neither the full query–gallery Cartesian product nor a hard-negative sample. The pairwise panels use axis limits of 65–100, with labeled ticks beginning at 70.

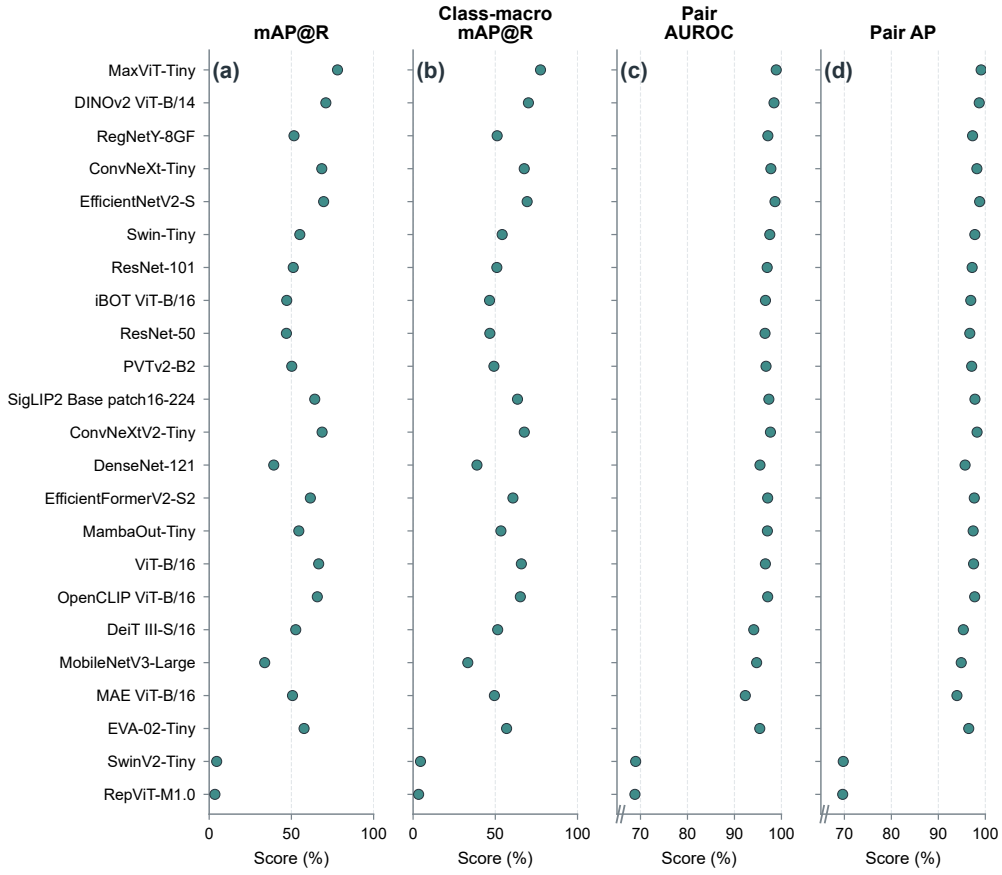


Figure 8: Complementary held-out retrieval results for 23 classifiers. Panels (a)–(d) report mAP@R, class-macro mAP@R, Pair AUROC, and Pair AP. The pairwise measures use 79,160 fixed query–gallery comparisons per classifier and axis limits of 65–100, with labeled ticks beginning at 70.

C.2 Few-Shot Prototype Readout

We compare five fully trained seed-41 checkpoints using the same frozen support episodes. The set contains ResNet-50, ConvNeXtV2-Tiny, SwinV2-Tiny, DINOv2 ViT-B/14, and SigLIP2 Base patch16-224. Each representation is learned during its full 100-epoch supervised run; the few-shot stage constructs class prototypes without updating the encoder.

For each checkpoint, we evaluate 1, 5, 10, and 20 support images per class over 90 fixed episodes. Support images come from distinct training groups, and every episode queries the fixed 3,958-image held-out test set. We average the L2-normalized support embeddings within each class, normalize the resulting prototype, and assign queries by cosine similarity. Figure 9 reports the episode means and sample SDs for Accuracy and Macro-F1.

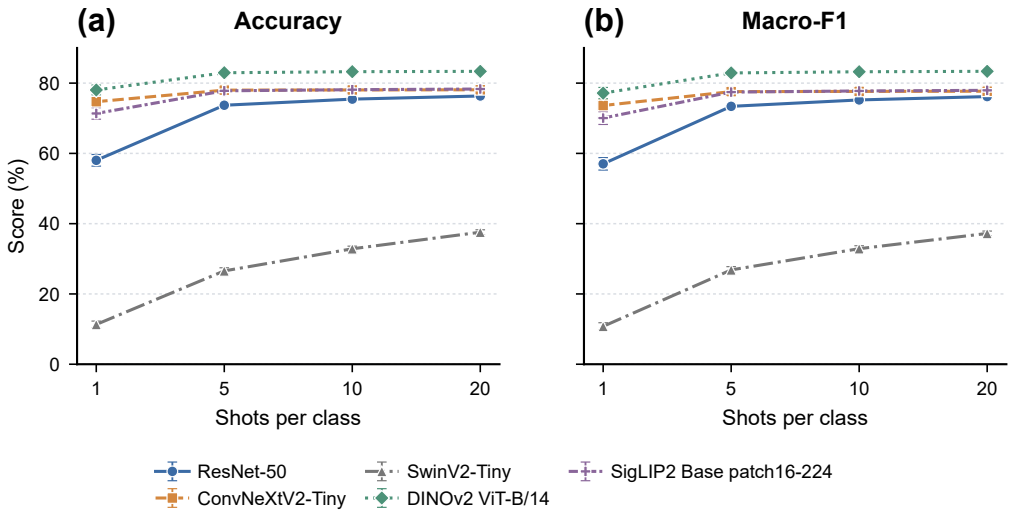


Figure 9: Few-shot prototype readout for five fully trained fixed checkpoints. Panels (a) and (b) report Accuracy and Macro-F1 across 1, 5, 10, and 20 support images per class. Points show means over 90 frozen episodes, and error bars denote episode-level sample SD. All checkpoints use seed 41, and every episode queries the same 3,958-image held-out test set.

DINOv2 records the highest mean at every shot count for both metrics. At 20 shots, it reaches 83.35% Accuracy and 83.36% Macro-F1. SwinV2-Tiny shows the largest increase, gaining 26.25 and 26.44 percentage points from one to 20 shots.

C.3 Taxonomy-Aware Recognition

Top-1 class predictions are projected to genus, family, and order. This separates exact-class errors from errors within a nearby taxonomic group. Figure 10 aligns Top-1, Macro-F1, and Balanced Accuracy across class, genus, family, and order for all 23 models, making the retained higher-rank information directly comparable across metrics.

C.4 Family-Auxiliary Supervision

Taxonomic projection measures information retained after species prediction. A separate paired comparison tests whether family supervision changes recognition performance. We add a 55-way family head to the ResNet-50 benchmark configuration while retaining the 100-way species head as the primary output. The species and family losses receive equal weight, and both heads use the same augmented image. Each paired run follows the benchmark training schedule and selects its checkpoint by validation species Macro-F1.

Table 16 compares the three-run held-out results. The family head reaches 82.11±0.46 accuracy, 81.10±0.71 Macro-F1, and 80.31±0.78 Balanced Accuracy on its 55-way output. Under the tested objective, family supervision does not improve species classification: Top-1, Macro-F1, and Balanced Accuracy decrease by 1.34, 1.22, and 1.20 percentage points. Retrieval changes direction across metrics. R@1 and R@5 decrease by 1.02 and 1.35 points, whereas class-macro mAP@R increases by 1.16 points.

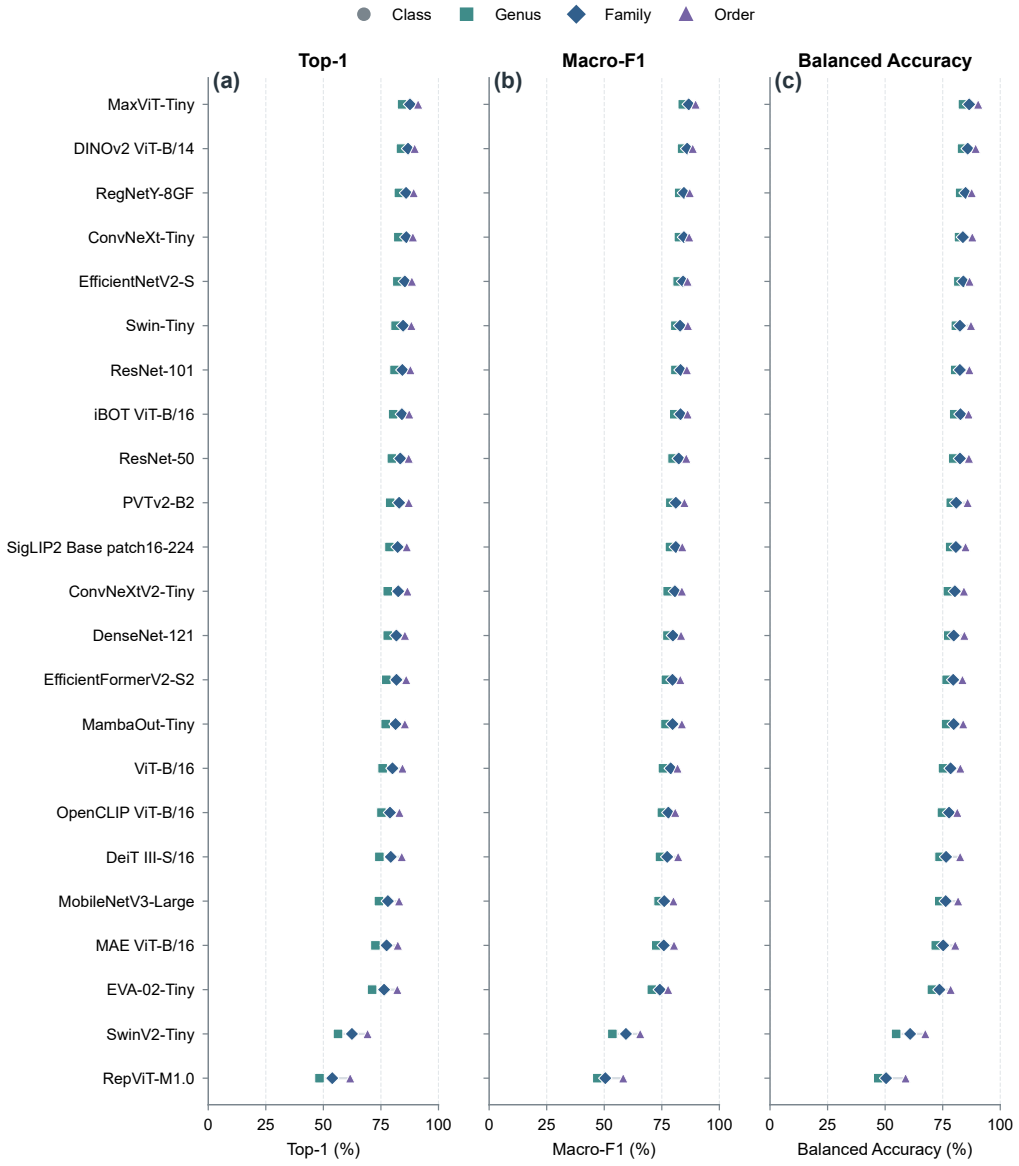


Figure 10: Taxonomy-aware recognition for 23 models. Panels report Top-1, Macro-F1, and Balanced Accuracy after projecting the same class predictions to class, genus, family, and order. Connecting segments trace the four taxonomic levels for each model.

Table 16: Three-run held-out results with and without family-auxiliary supervision. Values are mean ± sample SD in percentage points; paired changes are auxiliary minus baseline. Retrieval mAP@R is class-macro mAP@R. Paired changes are computed from the unrounded seed-level results before display rounding.

Setting	Top-1	Top-5	Macro-F1	Bal. acc.	R@1	R@5	mAP@R
ResNet-50	79.82±0.05	93.92±0.08	79.61±0.12	79.47±0.14	77.75±0.23	87.70±0.42	46.61±0.93
+ family auxiliary head	78.48±0.55	93.41±0.23	78.39±0.68	78.27±0.70	76.73±0.57	86.36±0.50	47.77±0.16
Paired change	−1.34	−0.51	−1.22	−1.20	−1.02	−1.35	+1.16

Model reference. ResNet-50 (He et al., 2016).

C.5 Scene-Stratified Recognition

The held-out partition contains 949 controlled-scene images and 3,009 intertidal-scene images. Figure 11 compares the two strata for the fixed seed-41 checkpoint of each model using Top-1, Macro-F1, and Balanced Accuracy. Paired points make the direction and magnitude of each within-model difference visible across the three metrics. The strata differ in subject presentation, background, class composition, and image source.

We also compare the 96 taxa represented in both strata. Controlled support ranges from 1 to 31 images per taxon, while intertidal support ranges from 4 to 48. For each model–seed run, recall is computed for every taxon and scene and macro-averaged across taxa. Table 17 reports the mean and sample SD over seeds 41, 42, and 43. This comparison fixes the taxon set and class weighting while retaining the observed images in each stratum.

Table 17: Taxon-matched scene analysis over the 96 taxa represented in both strata, comprising 949 controlled images and 2,874 intertidal images. Recall is computed for each taxon and scene, macro-averaged across taxa within each seed, and reported as mean ± sample SD over seeds 41, 42, and 43. Δ is controlled minus intertidal within each seed and is summarized across seeds in percentage points.

Model	Controlled	Intertidal	Δ
ResNet-50	77.64 ± 1.28	78.84 ± 0.68	-1.19 ± 1.63
ResNet-101	78.59 ± 0.44	79.85 ± 0.37	-1.26 ± 0.43
DenseNet-121	74.44 ± 1.85	76.80 ± 0.40	-2.36 ± 1.47
EfficientNetV2-S	80.68 ± 2.15	81.86 ± 0.63	-1.17 ± 2.73
RegNetY-8GF	80.05 ± 0.56	83.18 ± 0.41	-3.13 ± 0.84
MobileNetV3-Large	72.75 ± 2.15	73.61 ± 0.57	-0.86 ± 1.59
ConvNeXt-Tiny	80.36 ± 0.11	80.70 ± 0.36	-0.34 ± 0.29
ConvNeXtV2-Tiny	76.38 ± 1.97	77.02 ± 0.47	-0.65 ± 2.21
ViT-B/16	74.09 ± 1.26	74.37 ± 0.89	-0.28 ± 0.37
DeiT III-Small/16	72.79 ± 2.24	73.03 ± 1.67	-0.24 ± 1.40
Swin-Tiny	79.64 ± 1.38	80.93 ± 0.67	-1.29 ± 0.82
SwinV2-Tiny	39.58 ± 7.47	46.05 ± 6.79	-6.47 ± 1.62
PVTv2-B2	76.52 ± 0.87	78.25 ± 1.38	-1.73 ± 0.51
MaxViT-Tiny	81.22 ± 0.71	84.50 ± 0.45	-3.28 ± 0.27
EVA-02-Tiny	67.42 ± 1.21	69.89 ± 0.30	-2.47 ± 1.13
EfficientFormerV2-S2	75.57 ± 1.92	76.55 ± 0.39	-0.98 ± 2.02
MAE ViT-B/16	69.32 ± 0.48	71.49 ± 0.15	-2.17 ± 0.61
iBOT ViT-B/16	78.44 ± 1.17	78.42 ± 1.56	0.02 ± 0.59
DINOv2 ViT-B/14	79.51 ± 0.51	81.93 ± 0.98	-2.42 ± 0.66
OpenCLIP ViT-B/16	71.77 ± 2.11	73.46 ± 0.45	-1.70 ± 2.56
SigLIP2 Base patch16-224	75.39 ± 1.84	77.24 ± 0.54	-1.85 ± 1.68
MambaOut-Tiny	74.42 ± 2.03	76.18 ± 0.55	-1.76 ± 2.58
RepViT-M1.0	47.73 ± 2.04	49.11 ± 3.55	-1.38 ± 2.08

Across the three-seed means, controlled-scene recall has a lower point estimate for 22 of the 23 models; iBOT ViT-B/16 differs by 0.02 percentage points in the opposite direction.

C.6 Shared-Taxon Transfer to BackHome19K

We evaluate the three ResNet-50 benchmark checkpoints on taxonomically matched BackHome19K images without further training. Scientific binomials are parsed mechanically from filenames and matched by case-normalized exact equality. This procedure yields 227 images from eight shared taxa. Cross-dataset perceptual-hash screening removes three images from the shared set, leaving



Figure 11: Recognition results stratified by scene type for the fixed seed-41 checkpoint of each of the 23 models. Panels report Top-1, Macro-F1, and Balanced Accuracy, with paired points connecting the controlled ($n = 949$) and intertidal ($n = 3{,}009$) results for each model.

224 images from eight taxa. None of the removed images belongs to the official test subset, which contains 25 images from seven shared taxa.

Table 18 reports the official-test results in two prediction spaces. The original output retains all 100 ShellfishNet classes. The restricted output masks nonshared classes before ranking the seven shared taxa. On this subset, restriction raises mean Top-1 accuracy from 68.00% to 96.00% and Top-5 accuracy from 88.00% to 100.00%. Because the official subset contains only 25 images from seven shared taxa, this experiment is treated as a limited shared-taxon diagnostic. The restricted result measures discrimination after nonshared classes are masked and is not directly comparable with the original 100-class result.

Table 18: Frozen ResNet-50 transfer to the 25-image BackHome19K official-test subset. Values are mean ± sample SD over three checkpoints, in percentage points.

Prediction space	Top-1	Top-5	Macro-F1	Bal. acc.
Original 100-class output	68.00±4.00	88.00±0.00	71.77±3.78	63.89±5.50
Restricted to seven shared taxa	96.00±0.00	100.00±0.00	86.94±7.07	88.89±5.50

Dataset and model references. BackHome19K (Valverde et al., 2025); ResNet-50 (He et al., 2016).

C.7 Frozen Zero-Shot Classification and Sensitivity

Table 19 records frozen zero-shot classification with four scientific-name templates. OpenCLIP ViT-B/16 and SigLIP2 Base patch16-224 use their released pretrained representations. The OpenCLIP evaluation uses `open-clip-torch==2.24.0`. We also report BioCLIP 2 as an exploratory zero-shot baseline. Its Top-1 accuracy is 87.97%, above the highest 23-model fine-tuned mean of 84.39%, but this is not a leakage-controlled or like-for-like comparison. BioCLIP 2 was trained on TreeOfLife-200M, a 214-million-image biological dataset, with additional LAION-2B experience replay (Gu et al., 2025). The ShellfishNet test partition is drawn mainly from iNaturalist and also includes GBIF, Wikimedia Commons, iDigBio, and the Encyclopedia of Life. TreeOfLife-200M documentation lists GBIF and the Encyclopedia of Life among its core providers, so provider-level source overlap is confirmed for part of the evaluation data. The released BioCLIP 2 checkpoint does not provide a complete training-image manifest or source identifiers, so exact image-level and species-level overlap cannot be measured here; we neither assert nor exclude such overlap. The result is therefore kept outside the fine-tuned ranking and is interpreted only as an exploratory, overlap-sensitive diagnostic rather than evidence of strong external generalization. All three evaluations use the same fixed test partition without training or test-time augmentation. Each row comes from one deterministic frozen evaluation, so no seed dispersion is estimated.

Table 19: Frozen zero-shot classification on the 3,958-image test partition. All values are percentages from one evaluation with four scientific-name templates. BioCLIP 2 is an exploratory baseline.

Encoder	Top-1	Top-5	Macro-F1	Bal. acc.
OpenCLIP ViT-B/16	2.98	10.11	1.44	2.88
SigLIP2 Base patch16-224	0.71	4.95	0.25	1.08
BioCLIP 2 (ViT-L/14)	87.97	97.98	87.65	87.54

Model references. OpenCLIP ViT-B/16 (Cherti et al., 2023); SigLIP2 Base patch16-224 (Tschannen et al., 2025); BioCLIP 2 (Gu et al., 2025).

The sensitivity study keeps model weights and test images fixed. It evaluates 19 conditions per model. A 3×3 matrix crosses scientific, common, and combined class names with bare, photo, and four-template prompts. The bare condition uses `{name}`, and the photo condition uses `a photo of {name}.` The four-template condition averages `a photo of {name}.`, `a close-up photo of {name}.`, `a natural history image of {name}.`, and `a shellfish specimen of {name}.` Ten other checks vary tokenizer formatting, input normalization, feature normalization, or logit scale. Across the name–prompt matrix, OpenCLIP Top-1

ranges from 2.75% to 11.17%. SigLIP2 Base patch16-224 ranges from 0.56% to 1.19%. Figure 12 shows the full matrix and the remaining checks. The released-setting values stay as the benchmark baselines; the other conditions serve as sensitivity probes.

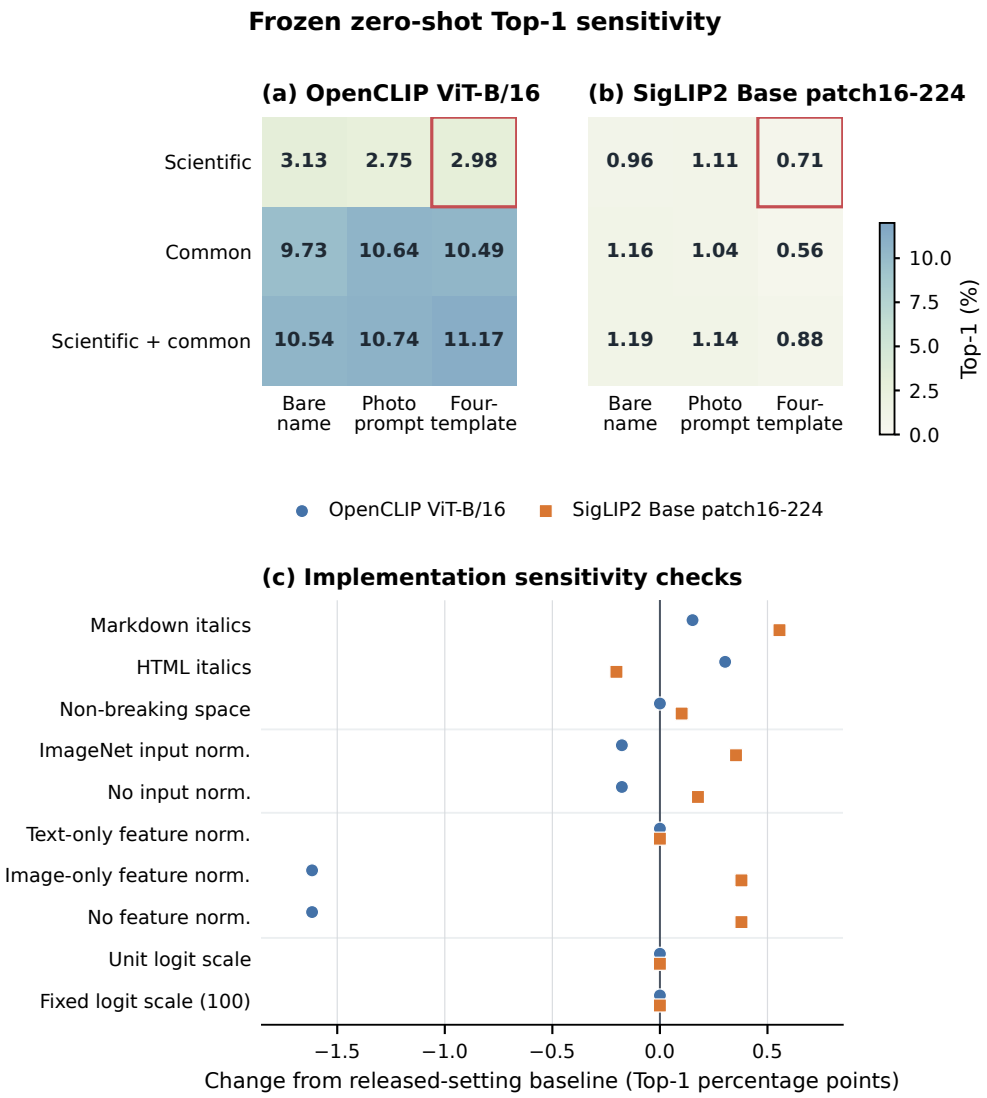


Figure 12: Frozen zero-shot sensitivity on the fixed 3,958-image test partition. Panels (a) and (b) show Top-1 for the nine name–prompt combinations. Both heatmaps use the same 0–12% color scale, and exact values are printed in each cell. The outlined cells corresponding to scientific names and the four-template prompt mark the released-setting baselines. Panel (c) reports Top-1 changes under tokenizer, input-normalization, feature-normalization, and logit-scale checks. All non-baseline conditions are diagnostic probes rather than selected benchmark entries.

D Complete Captioning Evaluation

D.1 Models and Common Captioning Protocol

The captioning benchmark uses a fixed subset of 512 test images covering all 100 classes and produces 5,120 model outputs. The ten open models span classic image-captioning systems and more recent vision–language models, as listed in Table 20. Every model receives the same standardized

1024×1024 RGB image. Each source image is converted to RGB, resized with Lanczos interpolation while preserving its aspect ratio, centered on a white 1024×1024 canvas, and saved as a quality-95 JPEG before model-specific processing. Chat-based vision–language models receive a common English instruction. Florence-2 uses its native <CAPTION> task token, while the classic captioning models use their native deterministic image-to-text interfaces. The resulting scores compare complete inference pipelines, so architecture and task-interface effects remain combined.

Table 20: Open models in the 512-image captioning evaluation.

Model	Family	Scale	Inference interface
ViT-GPT2	Classic captioning	< 1B	Single-image deterministic captioning
BLIP-Large	Classic captioning	< 1B	Single-image deterministic captioning
Florence-2 Base FT	Compact foundation model	< 1B	Single-image prompted captioning
LLaVA-Med 1.5 7B	Biology-specialized vision–language model	7B	Single-image chat inference
Qwen2.5-VL-3B	General vision–language model	3B	Single-image chat inference
Qwen3-VL-4B	General vision–language model	4B	Single-image chat inference
LLaVA-1.5-7B	General vision–language model	7B	Single-image chat inference
BLIP-Base	Classic captioning	< 1B	Single-image deterministic captioning
InternVL2.5-8B	General vision–language model	8B	Single-image chat inference
Phi-3.5 Vision	Compact vision–language model	4B	Single-image chat inference

Model references. ViT-GPT2 (NLP Connect, 2022; Dosovitskiy et al., 2021; Radford et al., 2019); BLIP-Large and BLIP-Base (Li et al., 2022a); Florence-2 Base FT (Xiao et al., 2024); LLaVA-Med 1.5 7B (Li et al., 2023a); Qwen2.5-VL-3B (Bai et al., 2025b); Qwen3-VL-4B (Bai et al., 2025a); LLaVA-1.5-7B (Liu et al., 2024); InternVL2.5-8B (Chen et al., 2024); Phi-3.5 Vision (Abdin et al., 2024).

Table 21 gives the common instruction used by chat-based vision–language models. Decoding uses batch size 1 and at most 48 new tokens. Prompted models use deterministic generation with `do_sample=False` and temperature 0.0 where supported. Classic captioners use deterministic native generation with `no_repeat_ngram_size=3`; model-specific end-of-sequence handling remains native. A human domain expert annotates and finalizes the reference captions, using GPT-5.6 Sol to assist with initial drafts. The expert inspects each image and revises every caption to follow the same one-sentence format.

Table 21: Common instruction used by chat-based vision–language models, shown as its constituent requirements.

Component	Instruction
Role	You are a scientific image captioner for a marine biology image set.
Task	Write one English sentence describing the image.
Visible content	Mention the main subject, visible attributes, and marine scene context.
Taxonomic wording	If a confident common name, genus, or species name fits the image, include it once.
Style	Prefer concrete nouns and visible traits over guesses.
Conflict handling	Ignore any cue that conflicts with the image.
Length	Keep the caption within 30 words.
Output format	Return a single sentence only, with no bullets, line breaks, or explanation.

Table 22: Operational anchors for the four blinded human-evaluation criteria.

Criterion	Score 0	Score 1	Score 2
M1: Subject correctness	The main subject is wrong, replaced by an unrelated object, or omitted.	The subject is related but overly broad or only partly identified, or the fine-grained name is incompatible with the fixed class.	The main subject is correct, and any taxonomic name is compatible with the fixed class.
M2: Visual factuality	A major count or logical error, a fabricated object, or multiple unsupported statements change the main meaning.	The core description is correct but contains one minor or local visual error.	Claims about color, form, texture, count, arrangement, state, behavior, and scene are supported by the image.
M3: Grounded informativeness	No usable image information is provided, or the content is mostly incorrect.	The subject is correct, but the description is clearly generic or omits salient information.	The subject is correct and at least one discriminative, image-supported attribute or scene detail is given; a complete, accurate short sentence also qualifies for a simple image.
M4: Language and format	The output is empty or unusable, contains multi-sentence explanation or prompt leakage, or seriously violates the length or format requirements.	The output is understandable but has a minor grammar, punctuation, completeness, or format issue.	The output is a clear, grammatically complete English sentence of 1–30 words, with no list, explanation, line break, or prompt residue.

D.2 Blinded Human Evaluation

Two human reviewers evaluated the generated captions with model identities hidden. They scored subject correctness (M1), visual factuality (M2), grounded informativeness (M3), and language and format (M4), assigning 0, 1, or 2 for each criterion. The four ratings sum to an overall score from 0 to 8. A fixed seeded permutation assigns 256 primary images to each reviewer. Ten images are sampled from each half for reciprocal overlap scoring, yielding the 20-image, 200-output agreement set. Model identities appear as anonymous slots, and outputs are shown in randomized order. Every model output received one primary score, giving 512 scores per model and 5,120 primary scores in total. Both reviewers independently scored these 200 outputs for agreement analysis; the second scores do not enter the model means. Confidence intervals use 10,000 stratified paired-image percentile bootstrap resamples, with the primary reviewer assignment defining the strata.

Table 23: Blinded human evaluation on the 512-image captioning subset. Overall scores are primary-score means with 95% bootstrap confidence intervals. M1–M4 are mean criterion scores on a 0–2 scale; higher values are better, and column-best means are bold.

Model	Overall / 8 (95% CI)	M1	M2	M3	M4
Qwen3-VL-4B	**6.89 (6.78–6.99)**	1.488	**1.566**	**1.840**	1.992
Qwen2.5-VL-3B	6.82 (6.69–6.96)	**1.549**	1.535	1.744	1.996
LLaVA-1.5-7B	6.17 (6.00–6.34)	1.354	1.479	1.447	1.895
InternVL2.5-8B	5.88 (5.74–6.02)	1.109	1.188	1.604	1.980
BLIP-Large	5.61 (5.42–5.79)	1.203	1.254	1.340	1.812
Florence-2 Base FT	5.60 (5.40–5.79)	1.076	1.301	1.291	1.930
BLIP-Base	4.84 (4.62–5.05)	0.943	1.084	1.074	1.736
LLaVA-Med 1.5 7B	3.25 (3.13–3.36)	0.176	0.562	0.510	**2.000**
ViT-GPT2	2.57 (2.46–2.69)	0.088	0.252	0.385	1.848
Phi-3.5 Vision	2.51 (2.41–2.61)	0.029	0.230	0.357	1.895

Model references. ViT-GPT2 (NLP Connect, 2022; Dosovitskiy et al., 2021; Radford et al., 2019); BLIP-Large and BLIP-Base (Li et al., 2022a); Florence-2 Base FT (Xiao et al., 2024); LLaVA-Med 1.5 7B (Li et al., 2023a); Qwen2.5-VL-3B (Bai et al., 2025b); Qwen3-VL-4B (Bai et al., 2025a); LLaVA-1.5-7B (Liu et al., 2024); InternVL2.5-8B (Chen et al., 2024); Phi-3.5 Vision (Abdin et al., 2024).

The agreement set indicates that the two reviewers most often assign exactly the same language-and-format score (92%) and subject-correctness score (77%). Exact agreement is 66% for visual factuality and 63% for grounded informativeness. Table 24 also reports quadratic-weighted κ, which accounts for the distance between ratings on the three-point scale. Grounded informativeness has the lowest exact agreement and a weighted κ of 0.538, indicating greater reviewer dependence for this criterion. The 200-caption overlap set measures agreement, while model means retain one primary rating per output.

Table 24: Inter-reviewer agreement on the 200 captions scored by both reviewers. These overlap scores are used only for agreement analysis.

Dimension	Criterion	Exact agreement	Weighted κ
M1	Subject correctness	77%	0.838
M2	Visual factuality	66%	0.729
M3	Grounded informativeness	63%	0.538
M4	Language and format	92%	0.538

For M4, both reviewers' ratings are strongly concentrated at score 2, so the 92% exact agreement coexists with a lower chance-corrected weighted κ.

D.3 Complete Automatic Metric Matrix

All nine automatic measures are computed for every model. CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), BLEU-4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Lin, 2004) compare generated text with the expert-annotated reference descriptions. BERTScore (Zhang et al., 2020) measures contextual textual similarity. CLIPScore (Hessel et al., 2021) measures image–text alignment. Scientific-name token mention recall records whether the final two normalized tokens of the class scientific name both occur in the generated caption. Caption length ratio compares generated and reference lengths. Figure 13 reports all nine measures for all ten models, with an image-level 95% bootstrap interval for every estimate.

CIDEr uses a single-reference TF-IDF n-gram cosine score. SPICE uses the `Spice` class from `pycocoevalcap.spice`. BERTScore reports English F1 from `distilbert-base-uncased` without baseline rescaling. CLIPScore uses `openai/clip-vit-base-patch32` and reports the cosine similarity between L2-normalized image and text embeddings, clipped at zero and multiplied by 100.

D.4 Cross-Metric Observations

Table 23 separates content judgments from output form. In particular, the three models with the lowest overall human scores retain mean language-and-format scores of 1.85–2.00 but have subject-correctness means of 0.03–0.18. Figure 13 preserves all nine automatic measures as separate scores. Reference-overlap metrics emphasize lexical and propositional agreement with the expert annotations. CLIPScore emphasizes image–text compatibility, while scientific-name token mention recall measures spontaneous full-binomial mention under a prompt that also permits common or genus names; caption length ratio describes output form. These complete-pipeline comparisons combine visual content, output length and format, and interface-specific instruction following. Ranking differences across the human and automatic measures reflect the different properties that they assess.

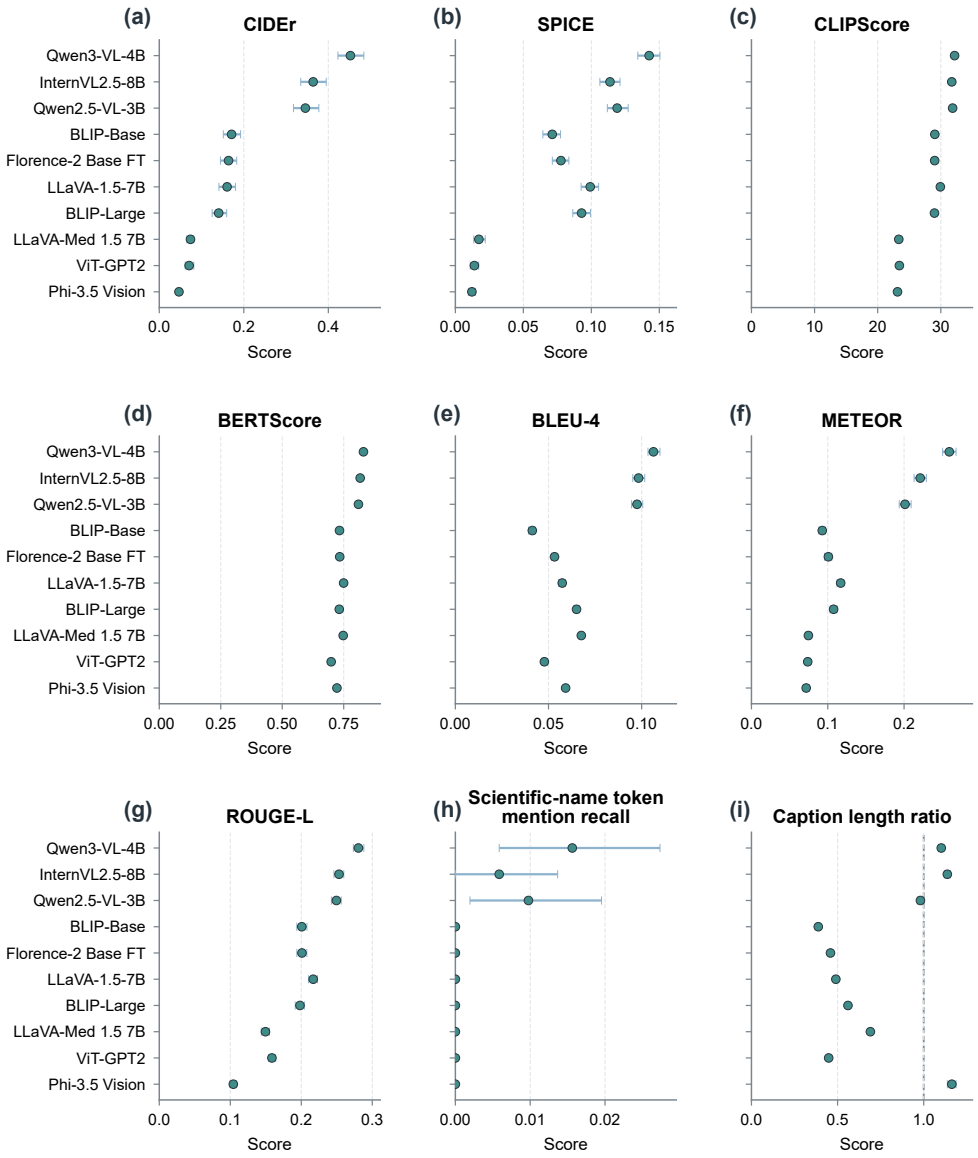


Figure 13: Complete automatic caption metrics for ten open models on 512 images. Panels report CIDEr, SPICE, CLIPScore, BERTScore, BLEU-4, METEOR, ROUGE-L, scientific-name token mention recall, and caption length ratio. Points show model estimates, and horizontal bars show image-level 95% bootstrap intervals; the dashed line marks a length ratio of 1.0.

D.5 Qualitative Failure Cases

Aggregate metrics summarize performance over the fixed subset but do not show the content of individual errors. Figures 14–18 compare five expert-annotated reference captions with the exact outputs produced by five evaluated models for images where the disagreement is visually clear. The examples include substitution of unrelated objects, confusion between a shell and another object category, omission of the main subject, and unsupported visual uncertainty.



Ground Truth (Expert Annotation):
A bay scallop (*Argopecten irradians*) shell, gray-brown and fan-shaped with deep radial ribs, is held above a rocky intertidal shore.

ViT-GPT2 (Failure Case):
a person holding a small bird on a beach

Figure 14: Qualitative comparison for *Argopecten irradians*. ViT-GPT2 replaces the clearly visible hand-held fan-shaped shell with an unrelated bird.



Ground Truth (Expert Annotation):
A pale, spiny *Aporrhais pespelecani* shell rests on fine sandy sediment in a shallow underwater marine scene.

BLIP-Large (Failure Case):
there is a starfish that is laying on the sand

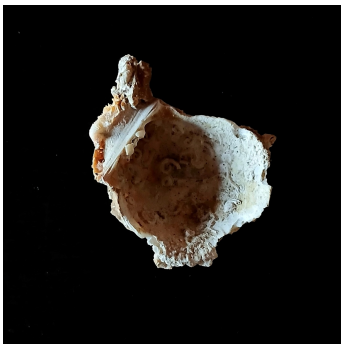
Figure 15: Qualitative comparison for *Aporrhais pespelecani*. BLIP-Large mistakes the prominent spiny shell for a starfish.



Ground Truth (Expert Annotation):
A pale cluster of common whelk (*Buccinum undatum*) egg capsules rests amid seaweed and driftwood on rounded beach stones at the tide line.

Florence-2 Base FT (Failure Case):
A pile of rocks and twigs on the ground.

Figure 16: Qualitative comparison for *Buccinum undatum* egg capsules. Florence-2 Base FT describes the surrounding stones and twigs but omits the central pale cluster.



Ground Truth (Expert Annotation):
A weathered, irregular *Spondylus gaederopus* shell, with jagged projections and chalky white-to-brown texture, is shown as an isolated marine bivalve specimen against a dark background.

LLaVA-1.5-7B (Failure Case):
A rock with a hole in it sits on a black surface.

Figure 17: Qualitative comparison for *Spondylus gaederopus*. LLaVA-1.5-7B mistakes the isolated jagged shell for a rock.



Ground Truth (Expert Annotation):
A spiny, algae-darkened crown conch (*Melongena corona*) rests in shallow seawater on a pale sandy surface scattered with fine marine debris.

Phi-3.5 Vision (Failure Case):
The image shows a blurred and indistinct marine scene, possibly a painting, with no clear subject or details to identify.

Figure 18: Qualitative comparison for *Melongena corona*. Phi-3.5 Vision reports a blurred, indeterminate scene despite the clearly visible spiny shell.

E Extended Synthetic Sensitivity Analysis

E.1 Complete Transformation-by-Level Results

The generic sensitivity analysis independently applies Gaussian noise, Gaussian blur, brightness changes, and contrast changes to the same 3,958 held-out images. All evaluations use fixed seed-41 checkpoints. Each transformation is applied independently to the cached 224×224 RGB image before model normalization. Gaussian noise uses a deterministic image-specific draw in the $[0, 1]$ pixel range and is clipped to that range. Gaussian blur uses Pillow's `GaussianBlur` operator with the reported radius. Brightness and contrast use Pillow's `ImageEnhance` operators. Gaussian noise uses standard deviations 0.04, 0.08, and 0.12. Gaussian blur uses radii 0.8, 1.4, and 2.0. Brightness factors are 0.80, 0.65, and 0.50; contrast factors are 0.75, 0.55, and 0.35. Main-text Figure 4(a) summarizes the five-checkpoint mean Top-1 trends. Figure 19 reports Top-1, Macro-F1, and Balanced Accuracy for every checkpoint, transformation, and level.

The generic and marine-inspired panels use different targeted five-checkpoint sets. We interpret each average within its own panel and use the analyses to characterize the evaluated fixed checkpoints.

Across the five selected fixed checkpoints, the strongest noise and blur settings produce larger average changes than the strongest brightness and contrast settings. At level 3, mean Top-1 is 35.37% for Gaussian noise and 49.64% for Gaussian blur. The corresponding values are 74.71% for brightness and 73.06% for contrast. Macro-F1 and Balanced Accuracy follow the same transformation ordering at the strongest settings, as shown in Figure 19. The ordering of the evaluated checkpoints changes with transformation strength, especially under Gaussian noise. Milder brightness and contrast changes preserve more of the clean-image ordering. The comparison distinguishes fixed checkpoints with similar held-out accuracy but different sensitivity profiles under the specified synthetic changes.

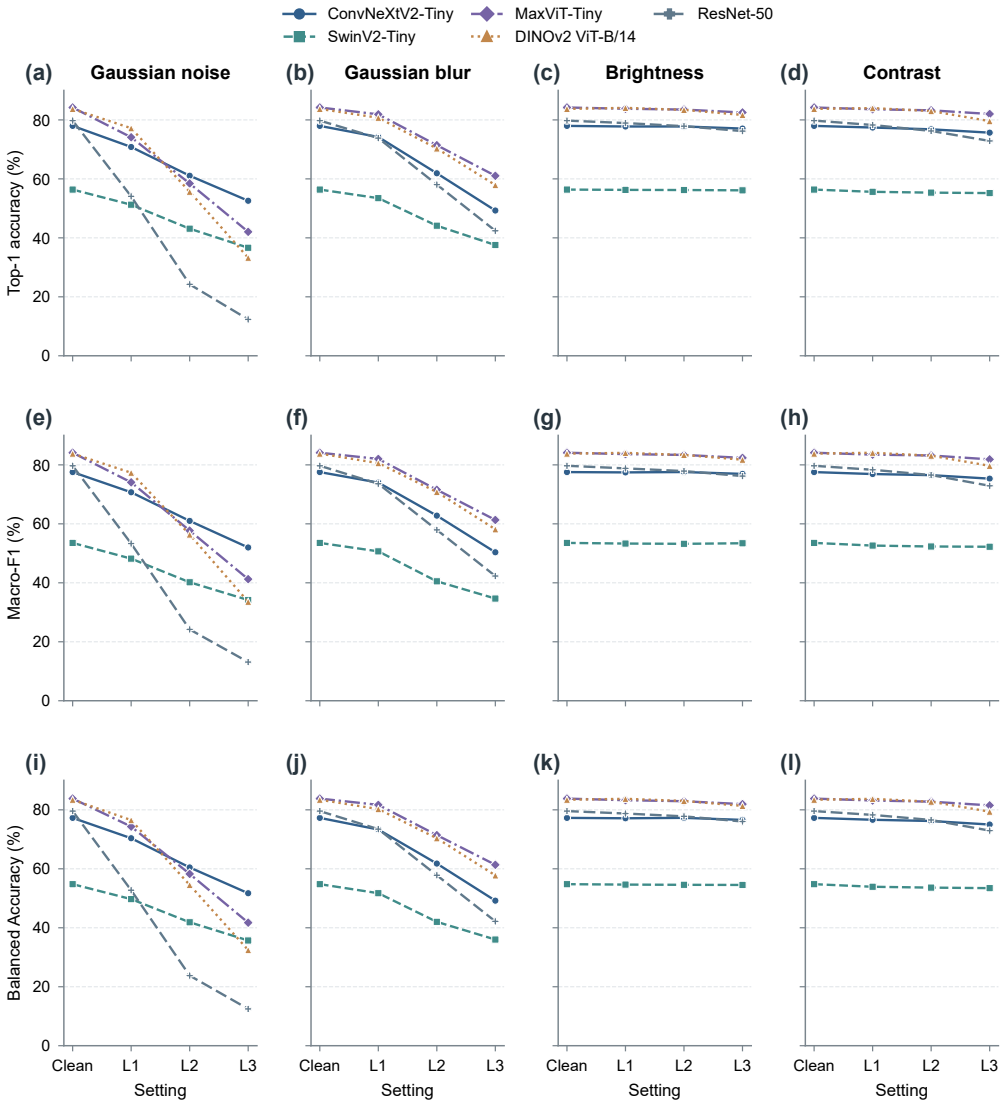


Figure 19: Complete controlled synthetic sensitivity results for five fixed checkpoints on 3,958 held-out images. Rows report Top-1, Macro-F1, and Balanced Accuracy; columns report Gaussian noise, Gaussian blur, brightness, and contrast. Each panel includes clean inputs and the three prescribed levels, with consistent checkpoint encodings across panels.

F Additional Frozen-Test Diagnostics

F.1 Marine-Inspired RGB Transformations

This analysis applies three fixed RGB transformations to the same 3,958 held-out images. Five fixed seed-41 checkpoints are evaluated without retraining or test-time augmentation. The two water-attenuation conditions use Jerlov II and Jerlov III coefficients from the released CADDY alpha-blend implementation (Gomez Chavez et al., 2019; Gomez Chavez, 2019). For channel c, the attenuation mask is $255\exp(-K_{d,c}\cdot 5)$. The transformation uses an equal blend of the input and mask. This empirical additive RGB color-cast and contrast-compression stress test is adapted from the CADDY utility; it is not a depth-resolved radiative-transfer simulation. The Jerlov II BGR coefficients are $(0.0619, 0.0998, 0.580)$; Jerlov III uses $(0.117, 0.145, 0.616)$. Under this formula, the larger Jerlov II mask transmission terms yield a brighter additive mask; the observed Jerlov II–III ordering is therefore not interpreted as a physical severity ranking of water types. The marine-snow condition adapts the PHISWID marine-snow synthesis to the RGB-only benchmark inputs (Kaneko et al., 2026).

The five-checkpoint mean Top-1 is 72.66% on clean images. It is 49.07% under Jerlov II attenuation, 55.84% under marine snow, and 51.20% under Jerlov III attenuation. Main-text Figure 4(b) reports the corresponding clean-to-transformed decreases for each checkpoint, while Figure 20(a) retains the absolute Top-1 results.

F.2 Class-Frequency Imbalance

Training support ranges from 106 to 401 images per class. The maximum-to-minimum ratio is 3.783, and the training-support Gini coefficient is 0.1395. ShellfishNet therefore has moderate class-frequency imbalance rather than an extreme long-tail distribution. We define support-ranked strata before inspecting test metrics. Tail contains the lowest 33 classes, Medium the next 33, and Head the highest 34. These groups account for 22.00%, 36.37%, and 41.62% of the training images.

The performance summary covers all 69 formal model–seed runs and 6,900 per-class records. Mean per-class Recall is 74.31% for Tail, 74.77% for Medium, and 77.48% for Head. Mean F1 is 75.62%, 75.01%, and 76.51%, respectively. Figure 20 reports the support strata and both metrics. No re-sampling, reweighting, augmentation, or retraining is used in this analysis. The stratum differences describe an association with training support; they are not a causal estimate of sample-count effects.

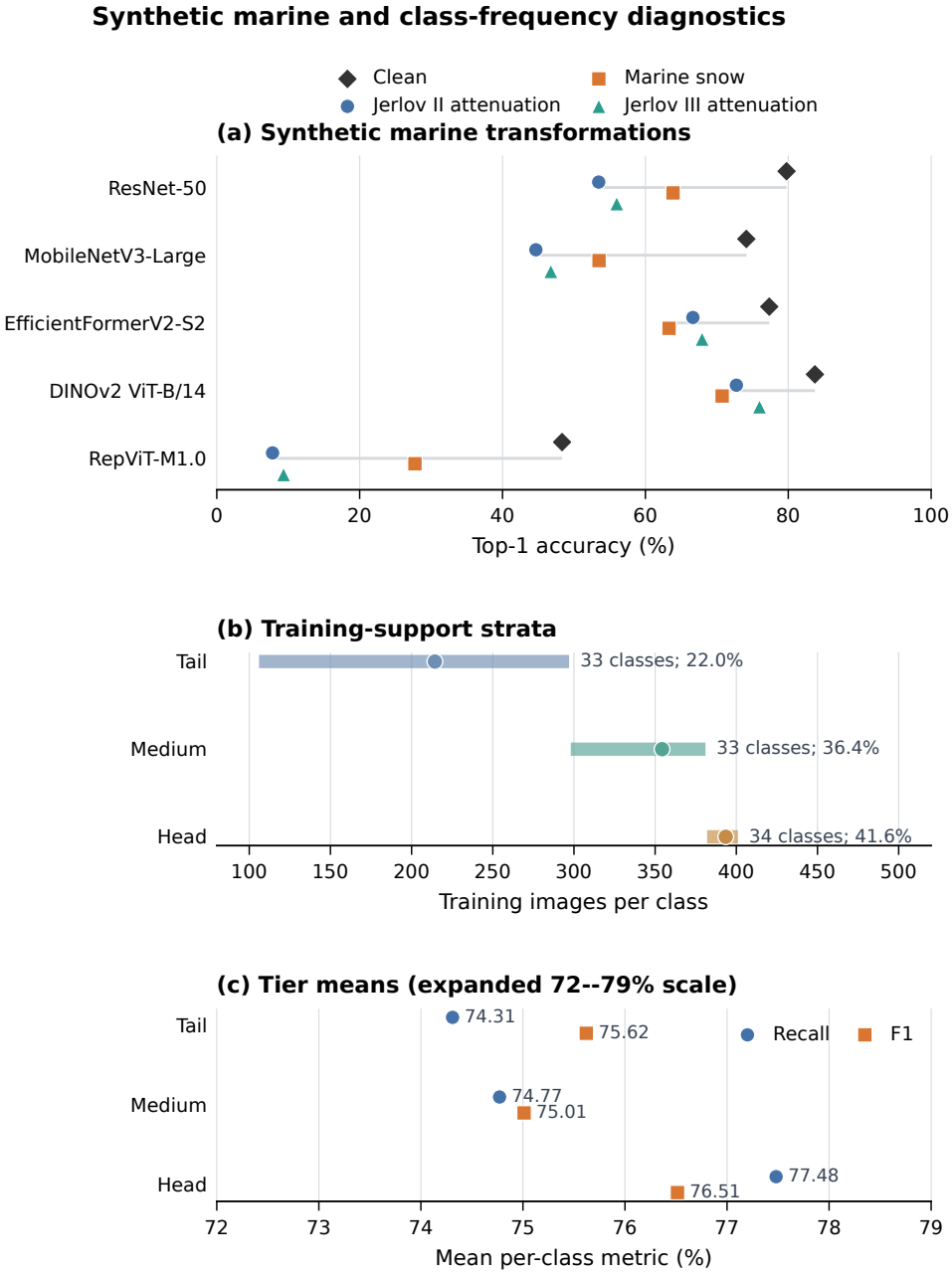


Figure 20: Marine-inspired RGB transformations and class-frequency analysis. Panel (a) reports absolute Top-1 accuracy for clean images and three fixed transformations, using five fixed seed-41 checkpoints on the same 3,958-image test partition. Panel (b) shows the training-support range and mean for each support-ranked stratum; labels give class counts and training-image shares. Panel (c) reports mean per-class Recall and F1 across all 69 model–seed runs on an expanded 72–79% scale.