ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Mask-Aware DINOv2 Foundation Features for Label-Efficient Environmental Microorganism Classification: A Proof-of-Concept Study on Environmental Microorganism Image Dataset Sixth Version

Mzhda Hiwa Hama[1], Chewan Jalal Mohammed[1], Abdulsalam Abdulla Ghadir[2]

[1]Department of Computer Science, College of Sciences, University of University of Sulaimani, Sulaymaniyah, Iraq, [2]Department of Software Engineering, Faculty of Engineering and Computer Science, Qaiwan International University, Sulaymaniyah, Iraq.

ABSTRACT

Environmental microorganism classification is difficult because curated microscopy datasets are small, visual appearance varies within classes, and expert annotation is costly. We examined mask-guided frozen DINOv2 representations on environmental microorganism image dataset sixth version (EMDS-6) using 840 images from 21 balanced classes. Twenty repeated partitions were stratified by class and grouped by exact image hash so that identical content could not cross training, validation, and test sets. The global DINOv2 class-token baseline and full hybrid were evaluated with 1–24 labeled training images per class; DINOv2-only, handcrafted-only, partial-fusion, full-fusion, and leave-one-block-out ablations were evaluated at 24 shots. Ground-truth masks were used for every mask-guided feature, so this remains an oracle spatial-guidance study rather than an end-to-end system. At 24 shots, the full hybrid linear support vector machine achieved 95.71% ± 1.23% accuracy and 95.67% ± 1.26% macro-F1, compared with 92.93% ± 1.82% macro-F1 for global DINOv2. The paired gain was 2.74 percentage points (95% bootstrap confidence interval, 1.86–3.61; exact signed-rank $P = 0.000019$; Holm-adjusted P across six label budgets = 0.000057). However, global-plus-crop features reached 95.33% ± 1.40% macro-F1 and did not differ significantly from the full hybrid after correction. Handcrafted features alone reached 64.00% ± 3.44%, and removing them from the hybrid did not reduce performance. The results support mask-guided frozen features within EMDS-6, but not deployment, segmentation robustness, supervised-model superiority, or external generalizability.

Index Terms: Environmental Microorganisms, DINOv2, Vision Transformers, Support Vector Machine, Environmental Microorganism Image Dataset Sixth Version-6

1. INTRODUCTION

Environmental microorganisms contribute to aquatic and terrestrial processes and can serve as ecological indicators, yet microscopic identification is slow and depends on specialist expertise. Automated image classification could support environmental monitoring, wastewater assessment, ecological studies, and high-throughput screening. The

Access this article online

DOI:10.21928/uhdjst.v10n2y2026.pp226-237 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Copyright © 2026 Hama, *et al*. This is an open access article distributed under the Creative Commons Attribution Non-Commercial No Derivatives License 4.0 (CC BY-NC-ND 4.0)

Corresponding author's e-mail:
Mzhda Hiwa Hama, Assistant Lecturer, Department of Computer Science, College of Science, University of University of Sulaimani, Sulaymaniyah, Iraq mzhd.hama@univsul.edu.iq

Received: 08-08-2026 Accepted: 22-09-2026 Published: 05-11-2026

task is challenging because images show substantial within-class variation, limited between-class contrast, uncertain boundaries, variable orientation, and similar morphology across classes, while curated labeled datasets remain small.

The environmental microorganism image dataset (EMDS) series provides public benchmarks for environmental microorganism image analysis. EMDS sixth version (EMDS-6) contains 21 classes, with 40 original images and 40 corresponding ground-truth masks per class for denoising, segmentation, feature extraction, classification, and detection research [1], [2]. In a prior comparison of 21 convolutional and vision-transformer architectures, the best reported direct test accuracy was 40.32% and increased to 45.71% with augmentation [3]. A later few-shot study reported 57.40% accuracy across the same number of classes [4]. Differences in data partitions, supervision, and evaluation prevent direct comparison, but these studies illustrate the difficulty of learning stable representations from EMDS-6.

Vision transformers represent an image as a sequence of patches and transfer visual knowledge learned at scale [5]. DINOv2 extends this idea through self-supervised pre-training and provides both global and patch-level representations without task-specific fine-tuning [6]. Frozen DINOv2 features are attractive for small scientific datasets because only a lightweight downstream classifier must be fitted. However, general-purpose pre-training does not identify which pixels belong to a microorganism rather than background, debris, or acquisition artifacts.

Spatial annotations offer several ways to guide a frozen DINOv2 representation. Foreground patch tokens can be pooled, background context can be summarized separately, organism-centered crops can be formed, and morphology and texture descriptors can be computed from the segmented region. Hu moments describe aspects of contour [7], local binary patterns summarize local texture [8], and gray-level co-occurrence statistics characterize intensity organization [9]. These cues may complement global DINOv2 features when classes differ in fine morphology rather than broad semantic appearance.

The conceptual framework separates three information sources: Global image context from the DINOv2 class token, spatially localized organism appearance from mask-guided patch and crop features, and explicit morphology and texture from handcrafted descriptors. If these sources are complementary, their concatenation should improve mean performance relative to a global DINOv2 baseline. If they are redundant, the simpler representation should perform similarly. Because the available masks are expert ground truth, this framework tests representational complementarity under ideal spatial guidance rather than real-world segmentation robustness.

We, therefore, evaluated a frozen DINOv2 feature extractor for EMDS-6, with particular attention to spatial guidance from the supplied masks. The global DINOv2 class-token baseline and full hybrid were compared across six label budgets under 20 exact-content-group-aware partitions. At the primary 24-shot condition, the analysis separately tested handcrafted descriptors, global-plus-crop and global-plus-mask partial combinations, individual DINOv2 blocks, the full hybrid, and six leave-one-block-out variants. The primary question was whether full-hybrid macro-F1 exceeded the global DINOv2 baseline after duplicate-safe partitioning; the ablation question was whether the additional blocks supplied measurable incremental value. Because test-time ground-truth masks and one dataset were used, the study remains an oracle proof-of-concept rather than a deployable recognition system.

2. LITERATURE REVIEW

2.1. Automated Recognition of Environmental Microorganisms

Environmental microorganisms participate in aquatic ecosystems, nutrient cycling, and wastewater processes, and their identification can inform assessments of ecosystem and water quality. Conventional microscopy remains time-consuming, requires experienced taxonomists, and is subject to observer variability. These constraints motivate automated image-classification methods.

Environmental microorganism classification is fine-grained: Images may have low contrast, complex backgrounds, illumination variation, and large differences in organism shape, size, and orientation, while different classes can look similar. Earlier systems combined handcrafted shape, texture, and statistical descriptors with classifiers such as support vector machines (SVMs) and random forests. Such methods are computationally modest and interpretable, but their performance depends on segmentation quality and descriptor design [7]–[9].

2.2. Deep Learning and Environmental Microorganism Datasets

Deep learning can learn representations directly from images and has improved microorganism recognition. Transfer

learning is commonly used when only small labeled datasets are available. Liang *et al.* obtained useful results through careful convolutional-network optimization, although their setting still required substantial augmentation and tuning [10].

EMDS-6 is a public benchmark containing 840 original images from 21 balanced classes and a corresponding binary mask for each image [1], [2]. Its modest size and the visual similarity of several classes make classification difficult. Comparative work with convolutional networks, vision transformers, and few-shot learning has likewise reported the challenge of learning transferable features on this dataset [3], [4].

2.3. Self-Supervised Vision Transformers

Vision transformers use self-attention to model long-range relationships between image patches and have become strong alternatives to convolutional networks. Fully supervised transformers, however, generally require large labeled datasets and can overfit in small microscopy studies.

Self-supervised learning reduces reliance on manual class labels during pre-training. DINO and DINOv2 produce transferable global and patch-level representations for downstream tasks with limited labeled data [6], [11], [12]. Related biomedical-imaging studies have shown that self-supervised features can work with comparatively simple classifiers, an appealing property when microscopy annotations are scarce [13], [14].

2.4. Spatial Guidance and Hybrid Feature Learning

A global image representation also encodes background regions that may be irrelevant to microorganism identity. Binary masks can provide spatial guidance by focusing feature extraction on biologically relevant regions and reducing background influence. Prior work in microscopy and related image-analysis settings has reported benefits from combining spatial information with learned representations [14]–[16].

Handcrafted morphology and texture descriptors remain relevant because microorganism classes can differ in shape, contour, and internal texture. Studies in microorganism, plankton, and biomedical image analysis have reported that learned and handcrafted features can be complementary [15]–[18].

2.5. Research Gap

Environmental microorganism classification has been studied with convolutional networks, vision transformers, transfer learning, few-shot learning, spatial guidance, and handcrafted descriptors. What remains less clear is whether global self-supervised features, mask-guided dense features, crop features, and classical descriptors provide complementary information under the same repeated-partition protocol, and whether any observed gain is stable enough to support a strong claim.

To examine this question, we combined global and mask-guided DINOv2 features with handcrafted morphology and texture descriptors while keeping the DINOv2 backbone frozen. We compared the resulting representation with global, crop-only, and mask-pooled alternatives across multiple label budgets on EMDS-6. The contribution is deliberately framed as a proof-of-concept evaluation of spatially guided frozen features; it is not presented as external validation or as evidence of operational robustness.

3. MATERIALS AND METHODS

3.1. Study Design

A retrospective computational benchmark was conducted using the public EMDS-6 dataset; no new biological specimens were collected. The global DINOv2 baseline and full hybrid were evaluated under 20 repeated class-stratified, exact-content-group-aware partitions and six per-class training budgets. The 24-shot primary analysis additionally included a classical-only model, two partial combinations, four individual learned-feature conditions, and six leave-one-block-out variants. Model selection used training data only. The validation set was retained, but calibration was not repeated in the corrected analysis. The test set remained untouched until final evaluation within each repetition.

3.2. Dataset and Class Structure

EMDS-6 was obtained from Figshare (Version 1; DOI: 10.6084/m9.figshare.17125025.v1) [2]. The dataset consists of 840 original images and 840 corresponding binary masks, evenly distributed across 21 classes, with 40 image–mask pairs assigned to each class [1]. The class codes and their corresponding source dataset names are listed in Appendix A. The original terminology provided with EMDS-6 was preserved, including source spellings such as *Siprostomum*, *K. Quadrala*, *Gymlyano*, and *Stylongchia*.

3.3. Data Indexing, Pairing, and Integrity Checks

All image files were recursively indexed, and each original image was paired with its binary mask using the class directory and image identifier. This produced 840 image-mask pairs with no unmatched originals. Each pair received a permanent sample identifier, while MD5 hashes defined exact-content groups. Automated checks verified identifier uniqueness,

sample-label alignment, finite features, and mutually exclusive training, validation, and test partitions at both sample and content-group levels.

Exact duplicate screening identified one duplicated image pair in class 20 (sample identifiers 764 and 766). The earlier sample-level five-seed split placed the pair across training and test in two repetitions. In the corrected analysis, the hash was treated as a grouping key: both files were assigned to the same partition in every one of the 20 repetitions. The final reported classification and ablation results are therefore free of exact-content crossing between training, validation, and test sets.

3.4. Repeated Partitions and Label Budgets

For each of 20 seeds (11, 22., 220), exact-content groups were shuffled within class and allocated as intact groups to eight test images and eight validation images; the remaining 24 images formed the training pool. A deterministic exact-sum selection retained the 24/8/8 sample count even for the two-file duplicate group. For each label budget, intact content groups totaling k images per class were sampled from the training pool. Every method used the same partition and labeled training subset within a seed and budget.

3.5. Image Preprocessing and Orientation Averaging

Original images and masks were resized to 224 × 224 pixels using aspect-ratio-preserving letterboxing. RGB images were normalized using the DINOv2 image mean and standard deviation. To reduce sensitivity to microorganism orientation, features were averaged across five deterministic views: the unmodified image, horizontal reflection, and rotations of 90°, 180°, and 270°. The same geometric transformation was applied to the image and mask. A second organism-centered view was constructed by cropping the bounding box of the foreground mask with 12% proportional padding, followed by letterboxing.

3.6. Frozen DINOv2 Feature Extraction

The DINOv2-small model (facebook/dinov2-small; hidden dimension 384) was loaded through the Hugging Face Transformers implementation and kept frozen [6]. For each transformed whole image, the class token and all dense patch tokens were extracted. The class token represented global image content. Patch tokens were averaged globally and pooled spatially using the resized binary mask. Foreground pooling was calculated as the weighted mean of patch tokens inside the mask, while background pooling used the complementary weights. A foreground-background contrast vector was formed by subtracting the background representation from the foreground representation and applying L2 normalization. A separate class-token representation was extracted from the foreground bounding-box crop.

3.7. Handcrafted Morphology and Texture Descriptors

Forty-seven interpretable descriptors were computed from the image and largest connected component of the binary mask. Shape features included area fraction, normalized perimeter, circularity, eccentricity, solidity, extent, bounding-box aspect ratio, normalized major and minor axis lengths, Euler number, and seven log-transformed Hu moment invariants [7]. Intensity descriptors included the foreground mean, standard deviation, 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles, entropy, skewness, and kurtosis. Texture was represented using a 10-bin uniform local binary pattern histogram [8] and gray-level co-occurrence statistics – contrast, dissimilarity, homogeneity, energy, and correlation – summarized by their means and standard deviations across two distances and four angles [9].

3.8. Feature Families and Classification Models

The cached feature bank provided global class tokens, crop class tokens, global patch means, mask-pooled foreground features, background features, foreground-background contrast, and 47 handcrafted morphology and texture descriptors. The primary ablation compared global DINOv2 alone (384 dimensions), handcrafted descriptors alone (47), global-plus-crop (768), global-plus-mask (1,152), and the full hybrid (1,967). Crop-only, patch-mean, mask-pool, and contrast-only conditions were also fitted. Six leave-one-block-out variants removed each constituent of the full hybrid in turn. Orientation averaging had already been applied during feature extraction and could not be separated from the saved bank.

For all non-prototype models, features were standardized and reduced using principal component analysis retaining 95% variance with whitening. Linear SVM regularization C was selected from 0.003 to 10; logistic-regression C from 0.01 to 30; and RBF-SVM C and gamma from predefined grids. Hyperparameters were selected by three-fold stratified inner cross-validation using macro-F1. When one-shot training precluded stratified inner cross-validation, default hyperparameters were used. Linear SVMs used class-balanced weights. The support-vector formulation follows the maximum-margin framework introduced by Cortes and Vapnik [19].

3.9. Future Validation of Calibration and Prediction Sets

Probability calibration and conformal prediction were not recomputed under the corrected 20-seed exact-content-

group-aware protocol. As explained in Section 4.6, they therefore provide no empirical support for the revised conclusions. Temperature scaling [20] and adaptive prediction sets [21] remain candidate methods for future validation. This work should follow the definition of a fixed operational pipeline, including predicted masks, and use duplicate-safe partitions and independent external data. Evaluation should report calibration error, negative log-likelihood, empirical coverage, prediction-set size, and singleton rate. No validated uncertainty estimates are claimed here.

3.10. Evaluation Metrics and Statistical Analysis

The prespecified primary endpoint was test macro-F1 in the 24-shot condition; accuracy, balanced accuracy, weighted F1, and Matthews correlation coefficient were secondary classification metrics. Results are reported as mean ± standard deviation across 20 seeds. At every label budget, the full hybrid and global DINOv2 baseline were compared with a two-sided exact sign-permutation Wilcoxon signed-rank test that enumerated all sign assignments; Holm correction covered the six budgets [22]. A paired bootstrap with 50,000 resamples estimated the 95% confidence interval for the primary mean difference. Four core ablation comparisons and six leave-one-block-out comparisons received separate Holm corrections within their analysis families.

3.11. Computational Environment

DINOv2 feature extraction was run with Python 3.12.13, PyTorch 2.11.0, CUDA 12.8, and an NVIDIA Tesla T4 GPU. The frozen feature bank was reanalyzed with Python 3.12.13 and scikit-learn 1.8.0 in a nine-vCPU environment. Automated checks verified sample-label alignment, finite values, exact-content-group partition integrity, unique result keys, and exact agreement between saved predictions and recomputed accuracy.

The executed notebook recorded 4 min 19 s to extract the ground-truth-mask feature bank in 35 batches on the Tesla T4. Stored as float 32 values, the 840 × 1,967 hybrid matrix requires approximately 6.30 MiB (7.68 KiB/image), excluding the DINOv2 parameters and intermediate tensors. In the corrected 24-shot rerun, training-only hyperparameter selection and refitting of the full hybrid took a median of

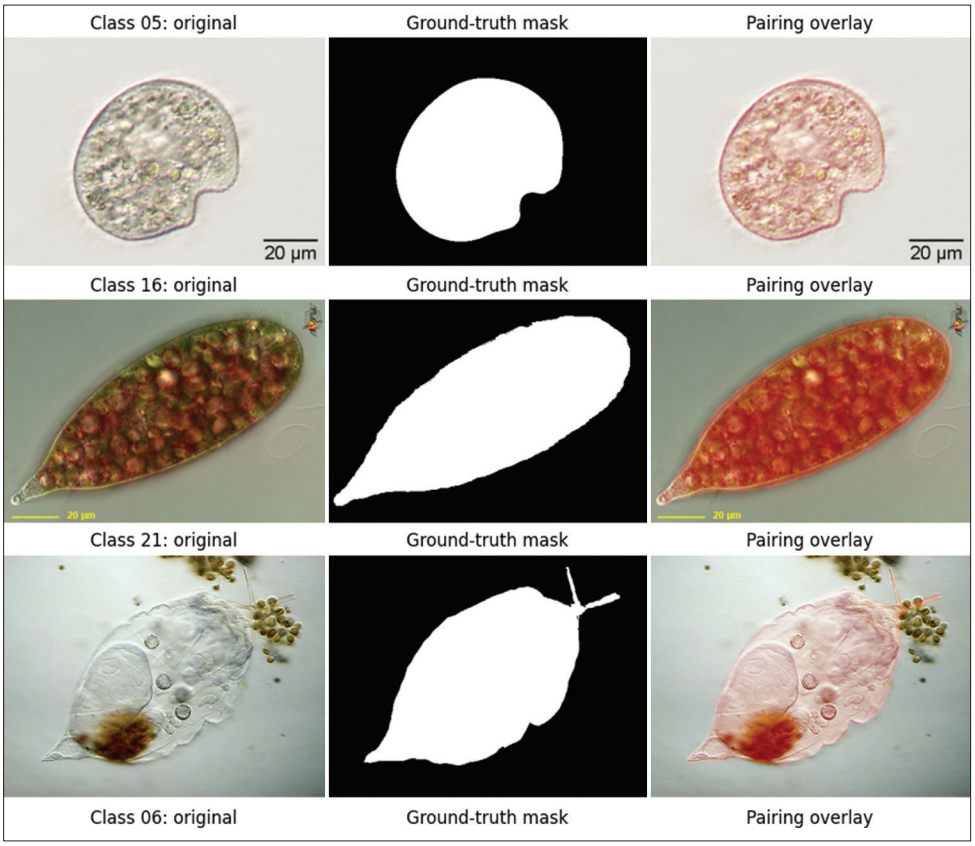


Fig. 1. Representative image–mask pairs used for quality control. Each row shows an original image, the corresponding binary ground-truth mask, and a red foreground overlay. The examples were randomly selected from source class codes 05 (*Colpoda*), 16 (*Euglena*), and 21 (*Synchaeta*).

8.53 s/seed (interquartile range, 7.66–11.20 s). Peak CPU/GPU memory was not captured.

TABLE 1: Dataset structure and repeated partition design

Component	Per class	Across 21 classes	Purpose
Original images	40	840	Classification inputs
Ground-truth masks	40	840	Oracle spatial annotations
Training pool	24	504	Feature-classifier fitting and k-shot sampling
Validation set	8	168	Reserved; calibration not repeated in the corrected analysis
Test set	8	168	Final evaluation per repetition
Repeated seeds	—	20 (11–220, step 11)	Group-aware variability estimate
Label budgets	1, 3, 5, 10, 20, 24	Same per class	Label-efficiency analysis

TABLE 2: Representations evaluated in the corrected group-aware analysis

Method	Representation	Classifier	Role
DINO CLS	384-D global CLS	Linear SVM	Global baseline
Handcrafted	47-D morphology/texture	Linear SVM	Classical-only
Crop CLS	384-D crop CLS	Linear SVM	Mask-guided crop
Mask pool	384-D foreground pool	Linear SVM	Dense mask feature
Global+crop	768-D concatenation	Linear SVM	Partial fusion
Global+mask	1,152-D concatenation	Linear SVM	Partial fusion
Full hybrid	1,967-D concatenation	Linear SVM	Complete model

SVM: Support vector machine

4. RESULTS

4.1. Dataset Audit and Feature Construction

Representative original images, ground-truth masks, and foreground overlays used for quality control are shown in Fig. 1. The audit identified 840 original images and 840 masks, with 40 image-mask pairs in each of 21 classes. The feature bank contained six 384-dimensional DINOv2 blocks and 47 handcrafted descriptors; five learned blocks plus the handcrafted block formed the 1,967-dimensional hybrid. The single duplicated hash contained sample identifiers 764 and 766. Across 20 corrected repetitions, no hash appeared in more than one partition within a seed, and every class retained exactly 24 training-pool, 8 validation, and 8 test images. Table 1 summarizes the dataset structure and repeated partition design.

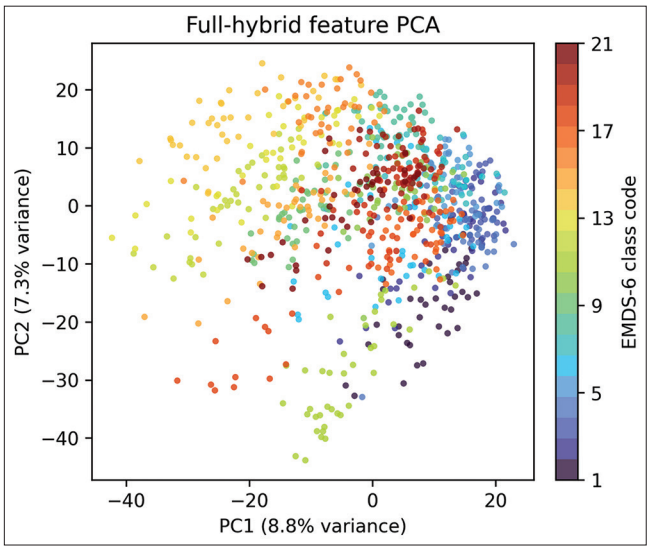


Fig. 2. Unsupervised two-dimensional PCA projection of all 840 standardized full-hybrid feature vectors. Colors denote environmental microorganism image dataset sixth version class codes; PC1 and PC2 explain 8.8% and 7.3% of total variance, respectively. The projection is descriptive and was not used for classifier training or selection.

TABLE 3: Duplicate-group-aware 24-shot performance across 20 seeds; fit time is the median wall-clock time for training-only hyperparameter selection and refitting in the nine-vCPU analysis environment

Method	Dim.	Accuracy, %	Macro-F1, %	MCC	Fit, s	Scope
Full hybrid	1,967	95.71±1.23	95.67±1.26	0.955	8.53	Full
Global+crop	768	95.36±1.39	95.33±1.40	0.952	6.08	Partial
Global+mask	1,152	94.97±0.96	94.92±1.02	0.948	6.57	Partial
Crop CLS	384	94.85±1.45	94.80±1.50	0.946	3.91	Single
Mask pool	384	94.11±1.88	94.08±1.89	0.939	3.15	Single
DINO CLS	384	93.01±1.80	92.93±1.82	0.927	4.12	Baseline
Handcrafted	47	64.67±3.51	64.00±3.44	0.630	0.21	Classical

MCC: Matthews correlation coefficient

4.2. Primary Performance and Core Ablation

The representations evaluated in the corrected group-aware analysis are summarized in Table 2. Table 3 reports the 24-shot classification performance and fitting times across 20 duplicate-group-aware repetitions, the full hybrid achieved 95.71% ± 1.23% accuracy and 95.67% ± 1.26% macro-F1. The global-plus-crop and global-plus-mask partial representations reached 95.33% ± 1.40% and 94.92% ± 1.02% macro-F1, respectively. Crop-only and mask-pool features reached 94.80% ± 1.50% and 94.08% ± 1.89%, while global DINOv2 reached 92.93% ± 1.82%. Handcrafted descriptors alone were substantially weaker at 64.00% ± 3.44%.

The full hybrid exceeded global DINOv2 and the global-plus-mask partial representation after correction, but it did not differ significantly from global-plus-crop. The unsupervised PCA projection in Fig. 2 shows broad class structure together with substantial overlap; it is descriptive rather than a supervised performance analysis.

4.3. Label Efficiency

Table 4 and Fig. 3 summarize the label-efficiency results for the global DINOv2 baseline and the full hybrid across the six label budgets. Both prespecified models improved as the number of labeled training images increased. Full-hybrid

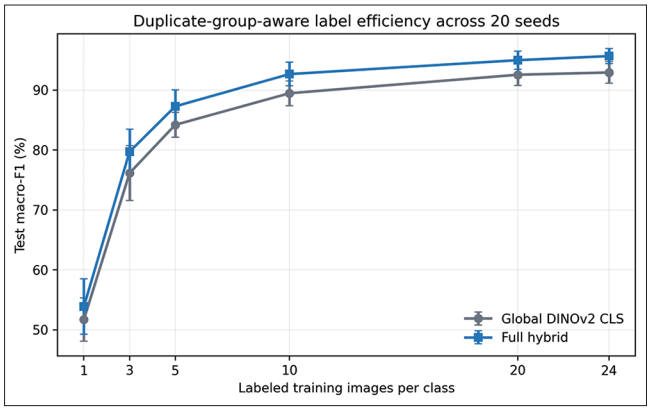


Fig. 3. Duplicate-group-aware test macro-F1 for the global DINOv2 baseline and full hybrid across 20 fixed seeds and six label budgets. Points are means and error bars are one standard deviation; the same partitions and labeled subsets were used for both models within each seed.

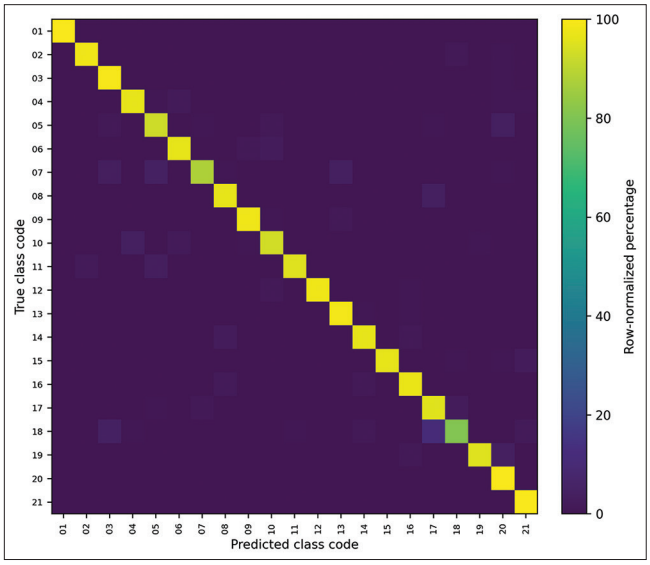


Fig. 4. Aggregate row-normalized confusion matrix for the full hybrid in the 24-shot condition across 20 duplicate-group-aware seeds. Each class contributes 160 test evaluations. Rows are true EMDS-6 class codes and columns are predicted codes; dataset labels are listed in Appendix A.

TABLE 4: Duplicate-group-aware label efficiency across 20 seeds

Shots	DINO CLS macro-F1, %	Full hybrid macro-F1, %	Gain, pp
1	51.70±3.62	53.86±4.63	+2.16
3	76.14±4.59	79.70±3.77	+3.56
5	84.19±2.06	87.27±2.77	+3.08
10	89.46±2.06	92.66±1.97	+3.20
20	92.55±1.79	94.99±1.51	+2.43
24	92.93±1.82	95.67±1.26	+2.74

Gains (pp, percentage points) are the mean of paired per-seed macro-F1 differences (full hybrid minus DINO CLS) across 20 seeds, calculated before rounding. They may differ by 0.01 pp from subtraction of the rounded aggregate means; for example, the gain at 20 shots is+2.43 pp

TABLE 5: Leave-one-block-out sensitivity analysis at 24 shots across 20 seeds

Removed block	Dim.	Macro-F1, %	Full−ablated, pp	Wilcoxon *P*-value	Holm *P*-value
Global CLS	1,583	95.62±1.18	+0.05	0.8895	1.0000
Crop CLS	1,583	94.81±1.55	+0.86	0.0192	0.1154
Patch mean	1,583	95.73±1.13	−0.06	0.7505	1.0000
Mask pool	1,583	95.81±1.34	−0.15	0.6197	1.0000
FG-BG contrast	1,583	95.48±1.30	+0.18	0.2467	1.0000
Handcrafted	1,920	95.75±1.04	−0.09	0.8603	1.0000

Positive differences indicate lower macro-F1 after removal. pp: Percentage points

macro-F1 rose from 53.86% ± 4.63% with one labeled image per class to 87.27% ± 2.77% at five shots, 92.66% ± 1.97% at 10 shots, and 95.67% ± 1.26% at 24 shots. The global DINOv2 baseline rose from 51.70% ± 3.62% to 92.93% ± 1.82% over the same range.

4.4. Incremental Value and Leave-One-Block-Out Analysis

The full hybrid had higher mean macro-F1 than global DINOv2 at every label budget. Paired gains were 2.16, 3.56, 3.08, 3.20, 2.43, and 2.74 percentage points (pp) at 1, 3, 5, 10, 20, and 24 shots, respectively. All six signed-rank comparisons remained significant after Holm correction; adjusted *P*-values ranged from 0.000011 to 0.00730.

In the primary 24-shot comparison, the paired macro-F1 gain was 2.74 pp, with a 50,000-resample bootstrap 95% confidence interval of 1.86–3.61 points. The exact sign-permutation Wilcoxon signed-rank *P*-value was 0.000019 and the Holm-adjusted value across label budgets was 0.000057. Thus, the corrected repeated analysis supports a stable hybrid-versus-global difference, although it does not show that every hybrid block is necessary. Table 5 further tests block removal and shows that no individual leave-one-out effect remained significant after Holm correction.

4.5. Class-level Performance and Error Patterns

Aggregating the 20 corrected 24-shot test sets gave 160 evaluations/class. The aggregate row-normalized confusion matrix for these repeated 24-shot evaluations is shown in Fig. 4. Class 01 had perfect precision and recall. The lowest class-level recall was 80.6% for class 18, which was most often predicted as class 17 (17 of 160 evaluations); class 17 had 95.6% recall but 86.4% precision. Classes 07 and 05 also showed lower F1 scores of 92.4% and 91.9%, respectively. These are repeated-partition error frequencies rather than counts of unique biological specimens.

The aggregate matrix reduces the instability of interpreting a single split, but repeated appearances of the same image across seeds are not independent biological replicates. Class-specific conclusions therefore remain descriptive and require confirmation on independently collected material.

4.6. Exploratory Calibration Scope

Calibration and conformal prediction were not recomputed under the corrected 20-seed group-aware protocol and are therefore not used to support the revised conclusions. The earlier five-seed diagnostic results came from the superseded sample-level partitions.

A reliable uncertainty analysis should be repeated after the full operational pipeline is fixed, including predicted masks, duplicate-safe partitions, and external data. Reporting nominal coverage without these conditions could give an unjustified impression of deployment readiness.

Accordingly, this revision treats calibrated confidence and conformal prediction as future validation tasks rather than evidence for clinical or environmental deployment.

5. DISCUSSION

5.1. Principal Findings

This proof-of-concept evaluated frozen DINOv2 features with ground-truth-mask spatial guidance and handcrafted descriptors on one small environmental microorganism dataset. Under 20 exact-content-group-aware partitions, the full hybrid achieved 95.71% mean accuracy and 95.67% mean macro-F1 across 21 EMDS-6 classes. The estimate is leakage-corrected for exact duplicates but remains an oracle, within-dataset result rather than expected performance on automatically segmented or externally acquired images.

The global DINOv2 class-token baseline achieved 92.93% mean macro-F1. The full hybrid was higher by 2.74 pp, with a 95% bootstrap interval of 1.86–3.61 points and a Holm-adjusted signed-rank *P*-value of 0.000057 across label budgets. This supports a repeatable difference from the global baseline. However, the global-plus-crop representation achieved 95.33% macro-F1 and was not significantly different from the full hybrid after correction (adjusted $P = 0.165$), so the data favor mask-guided spatial localization more clearly than they favor every element of the 1,967-dimensional fusion.

The core ablation qualified the original feature-fusion claim. Global-plus-mask was 0.74 pp below the full hybrid and remained different after correction (adjusted $P = 0.0385$), whereas global-plus-crop was only 0.33 points lower and was not significant. Handcrafted features alone performed poorly, and removing the handcrafted block from the full hybrid slightly increased mean macro-F1 by 0.09 points. Classical descriptors therefore offer interpretability but no demonstrated incremental accuracy in this dataset.

The cached feature bank required 4 min 19 s to extract on a Tesla T4. The float32 hybrid matrix occupied approximately 6.30 MiB, and training-only hyperparameter selection plus refitting of the 24-shot hybrid took a median of 8.53 s

(interquartile range, 7.66–11.20 s) in the nine-vCPU analysis environment. These measurements describe the frozen-feature classifier stage; they exclude operational segmentation and peak-memory profiling.

5.2. Relationship to Prior EMDS-6 Work

Previous studies on the EMDS-6 dataset have reported relatively modest classification performance using conventionally trained convolutional neural networks and transformer-based architectures. For example, Zhao *et al.* reported a test accuracy of 40.32% for the Xception model, which increased to 45.71% after applying data augmentation [3]. Yuan *et al.* reported 57.40% accuracy using few-shot learning with self-supervision [4]. The present one-shot result of 57.62% is numerically similar to the latter report, while the 24-shot result is substantially higher. These values should not be interpreted as a controlled head-to-head superiority claim because the partitioning schemes, supervision levels, augmentation procedures, model pre-training, and use of ground-truth masks differ. The current results instead show the potential of large-scale self-supervised transfer combined with explicit spatial priors.

The crop-only representation reached 94.80% macro-F1, and combining global and crop class tokens reached 95.33%. This parsimonious partial model was statistically indistinguishable from the full hybrid after correction. The result suggests that allocating more input resolution to the organism while retaining a global view accounts for much of the observed improvement. It also cautions against assuming that a larger concatenated vector is necessarily better.

5.3. Why Mask-Aware Dense Features May Help

DINOv2 produces a global class token and spatially resolved patch tokens [6]. Mask-guided crops, foreground pooling, and foreground-background contrast alter which spatial evidence is retained. In the leave-one-block-out analysis, removing the crop block reduced macro-F1 by 0.86 points (bootstrap interval, 0.30–1.42), but its signed-rank result did not remain significant after correcting six block-removal tests (adjusted $P = 0.115$). Removing any other block changed mean macro-F1 by no more than 0.18 points, and no leave-one-out comparison was significant after correction. The blocks are therefore substantially redundant, and the analysis does not establish a unique causal contribution for any single component.

5.4. Label Efficiency and Practical Annotation Implications

The learning curves indicated that the most substantial improvements in performance were achieved with only 1–5 labeled images per class. Classification accuracy surpassed 80% with three labeled samples and increased to nearly 89% with five. At 20 labeled images per class, the best-performing model achieved 95% accuracy, while incorporating the remaining four training images per class resulted in only a negligible improvement in mean performance. If these findings are confirmed through group-aware and external validation, they suggest that annotation efforts may be better directed toward selecting a small set of diverse and representative images rather than labeling a large number of highly similar samples. However, it is important to note that segmentation masks were already available in this study. In practice, generating accurate masks can require considerably more time and expertise than assigning image-level class labels. Therefore, the overall annotation effort should not be evaluated solely based on the number of labeled images.

5.5. Future Uncertainty Validation and Expert Review

The corrected analysis does not establish whether the model probabilities or prediction sets are suitable for expert review. The earlier five-seed calibration diagnostics came from superseded sample-level partitions and are excluded from the revised conclusions, as stated in Section 4.6. Future work should assess calibration and prediction-set performance with predicted masks, duplicate-safe partitions, and independent external data. Regularized adaptive prediction sets, class-conditional calibration, and selective prediction could be evaluated at several target coverage levels. Their practical value will depend on both the reliability of the uncertainty estimates and the number of cases referred to an expert. Until this validation is completed, probabilities and prediction sets should not be treated as validated confidence estimates or tools for deciding which cases require expert review.

5.6. Strengths

The revised analysis used a balanced dataset, verified image-mask pairing, permanent identifiers, exact-content grouping, 20 repeated partitions, identical splits across methods, training-only hyperparameter selection, paired effect estimates, and multiplicity correction. Prediction-derived accuracy matched the saved result table to numerical precision. The analysis also retained negative findings: Handcrafted features did not improve the full hybrid, global-plus-crop was statistically comparable to the full vector, and no leave-one-block-out effect survived correction.

5.7. Limitations

Several limitations prevent the current results from being considered deployment-ready or definitive external evidence. First, ground-truth masks were used to construct every mask-

guided feature, including for test images. This oracle setting estimates the value of accurate spatial guidance but does not show that unseen images can be processed automatically. A practical system must predict masks using training-only supervision and measure the resulting classification loss.

Second, one exact duplicate image pair was present in class 20. This pair crossed training and test in the earlier analysis, but the final 20-seed protocol grouped both files by their shared hash. Automated checks confirmed that no exact-content group crossed training, validation, and test partitions. This addresses exact duplication, although near-duplicates were not assessed with perceptual hashing.

Third, all images came from EMDS-6, with no independent dataset or laboratory cohort. Images from one source may share acquisition characteristics that simplify discrimination. Consequently, the results apply only to the evaluated EMDS-6 setting and should not be generalized to other microscopes, magnifications, illumination conditions, operators, specimen-preparation procedures, laboratories, or microorganism collections.

Fourth, image-level perturbation tests for blur, noise, illumination, contrast, occlusion, and mask error were not available from the cached feature bank. Fifth, the predicted-mask extension and supervised ConvNeXt comparator were not executed, and peak memory was not captured. Sixth, orientation averaging had already been applied before caching and could not be ablated. Finally, the PCA projection and repeated class-level errors are descriptive; independent biological material is needed for generalizable conclusions.

5.8. Future Work

The next step is to evaluate the same fixed protocol with predicted masks and a modern supervised transfer-learning baseline. Raw images are also required for controlled blur, noise, illumination, contrast, occlusion, and mask-perturbation experiments. Near-duplicate screening should add perceptual hashes, and an independent dataset or laboratory cohort should test acquisition shift. The parsimonious global-plus-crop model deserves particular attention because it matched the full hybrid statistically while using 61% fewer input dimensions.

6. CONCLUSION

Across 20 duplicate-group-aware EMDS-6 partitions, the full hybrid achieved 95.71% mean accuracy and 95.67% mean macro-F1. Its 24-shot macro-F1 exceeded the global DINOv2 baseline by 2.74 pp, with a 95% bootstrap interval of 1.86–3.61 points and a Holm-adjusted signed-rank *P*-value of 0.000057. Core and leave-one-block-out ablations showed that the gain cannot be attributed to every block: global-plus-crop was statistically comparable to the full hybrid, and handcrafted descriptors provided no measurable increment. Ground-truth masks, one dataset, absent predicted-mask and perturbation tests, no supervised-transfer comparator, and no external validation still prevent claims of robustness, generalizability, or deployment readiness. The evidence supports a label-efficient, mask-guided proof-of-concept on EMDS-6.

REFERENCES

[1] P. Zhao, C. Li, M. M. Rahaman, H. Xu, P. Ma, H. Yang, H. Sun, T. Jiang, N. Xu and M. Grzegorzek. "EMDS-6: Environmental microorganism image dataset sixth version for image denoising, segmentation, feature extraction, classification, and detection method evaluation". *Frontiers in Microbiology*, vol. 13, p. 829027, 2022.

[2] P. Zhao, C. Li, M. M. Rahaman, H. Xu, P. Ma, H. Yang, H. Sun, T. Jiang, N. Xu and M. Grzegorzek. "*EMDS-6*." Figshare, United Kingdom, 2021.

[3] P. Zhao, C. Li, M. M. Rahaman, H. Xu, H. Yang, H. Sun, T. Jiang and M. Grzegorzek. "A comparative study of deep learning classification methods on a small environmental microorganism image dataset (EMDS-6): From convolutional neural networks to visual transformers". *Frontiers in Microbiology*, vol. 13, p. 792166, 2022.

[4] J. Yuan, P. Zhao, N. Xu, C. Li and T. Jiang. "A Few-Shot Learning Based Environmental Microorganism Image Classification Method". In: *Proceedings 2024 2nd International Conference on Machine Vision, Image Processing and Imaging Technology (MVIPIT)*. pp. 199-203, 2024.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *Proceedings International Conference Learning Representations (ICLR)*. arXiv, New York, 2021.

[6] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes.,… & P. Bojanowski. "DINOv2: Learning robust visual features without supervision". *Transactions on Machine Learning Research,* 2024. Available: https://openreview.net/forum?id=a68SUt6zFt

[7] M. K. Hu. "Visual pattern recognition by moment invariants". *Transactions on Information Theory*, vol. 8, no. 2, pp. 179-187, 1962.

[8] T. Ojala, M. Pietikäinen and T. Mäenpää. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002.

[9] R. M. Haralick, K. Shanmugam and I. Dinstein. "Textural features for image classification". *IEEE Transactions on Systems, Man, and*

Cybernetics, vol. SMC-3, no. 6, pp. 610-621, 1973.

[10] C. M. Liang, C. C. Lai, S. H. Wang and Y. H. Lin. "Environmental microorganism classification using optimized deep learning model". *Environmental Science and Pollution Research*, vol. 28, no. 24, pp. 31920-31932, 2021.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski and A. Joulin. "Emerging Properties in Self-Supervised Vision Transformers". In: *Proceedings IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 9650-9660, 2021.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick. "Masked autoencoders are scalable vision learners. In: *Proceedings IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR)*". pp. 16000-16009, 2022.

[13] M. Nielsen, L. Wenderoth, T. Sentker and R. Werner. "Self-supervision for medical image classification: State-of-the-art performance with ~100 labeled training samples per class". *Bioengineering (Basel)*, vol. 10, no. 8, p. 895, 2023.

[14] D. Kang, P. Koniusz, M. Cho and N. Murray. "Distilling Self-Supervised Vision Transformers for Weakly-Supervised Few-Shot Classification and Segmentation". In: *Proceedings IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR)*. pp. 19627-19638, 2023.

[15] F. Kulwa, C. Li, J. Zhang, K. Shirahama, S. Kosov, X. Zhao, T. Jiang and M. Grzegorzek. "A new pairwise deep learning feature for environmental microorganism image analysis". *Environmental Science and Pollution Research*, vol. 29, no. 34, pp. 51909-51926, 2022.

[16] L. Nanni, E. De Luca, M. L. Facin and G. Maguolo. "Deep learning and handcrafted features for virus image classification". *Journal of Imaging*, vol. 6, no. 12, p. 143, 2020.

[17] H. Zheng, R. Wang, Z. Yu, N. Wang, Z. Gu and B. Zheng. "Automatic plankton image classification combining multiple view features via multiple kernel learning". *BMC Bioinformatics*, vol. 18, no. Suppl 16, p. 570, 2017.

[18] T. Panaïotis, E. Amblard, G. Boniface-Chang, G. Dulac-Arnold, B. Woodward and J. O. Irisson. "Benchmark of plankton images classification: Emphasizing features extraction over classifier complexity". *Earth System Science Data*, vol. 18, pp. 945-967, 2026.

[19] C. Cortes and V. Vapnik. "Support-vector networks". *Machine Learning*, vol. 20, pp. 273-297, 1995.

[20] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger. "On Calibration of Modern Neural Networks". In: *Proceedings 34th International Conference Machine Learning (ICML)*. *Proceedings of Machine Learning Research*, vol. 70, pp. 1321-1330, 2017.

[21] Y. Romano, M. Sesia and E. J. Candès. "Classification with Valid and Adaptive Coverage". In: *Proceedings Advances in Neural Information Processing Systems 33*, 2020.

[22] S. Holm. "A simple sequentially rejective multiple test procedure". *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65-70, 1979.

APPENDIX

APPENDIX A: EMDS-6 class-code mapping

Class codes and dataset labels used in the analysis

Class code	Dataset label
01	Actinophrys
02	Arcella
03	Aspidisca
04	Codosiga
05	Colpoda
06	Epistylis
07	Euglypha
08	Paramecium
09	Rotifera
10	Vorticella
11	Noctiluca
12	Ceratium
13	Stentor
14	Siprostomum
15	K. Quadrala
16	Euglena
17	Gymnodinium
18	Gymlyano
19	Phacus
20	Stylonychia
21	Synchaeta

EMDS-6: Environmental microorganism image dataset sixth version