Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Cross Granularity Metric Fusion for Rare Species Identification under Scarce Biodiversity Image Data Tasks

Moritz Keller, Violet Bennett, Roman Lebedev

Usher Institute, School of Population Health Sciences, University of Edinburgh, Edinburgh, United Kingdom

Abstract

Biodiversity conservation increasingly relies on automated image identification systems to monitor ecosystems and track the population dynamics of vulnerable taxa. However, rare species identification presents profound computational challenges due to extreme data scarcity, pronounced intraclass variance caused by environmental factors, and subtle interclass visual differences among closely related species. Standard convolutional neural networks and traditional metric learning approaches frequently fail to capture the multi-level discriminative features necessary for accurate taxonomic classification under such constrained conditions. This paper introduces a novel cross-granularity metric fusion framework designed specifically for the visual identification of rare and endangered species. By extracting and aligning semantic features across multiple spatial resolutions and hierarchical levels, the proposed methodology constructs a highly robust feature representation that mitigates the adverse effects of long-tailed data distributions inherent in ecological datasets. We detail a comprehensive architectural design that simultaneously processes macroscopic morphological traits alongside microscopic texture patterns, subsequently fusing these distinct granularities through a unified distance metric space. Extensive evaluations on established fine-grained ecological datasets demonstrate that our approach significantly outperforms contemporary baseline architectures in top-1 accuracy and mean average precision, particularly for severely underrepresented taxonomic classes. Furthermore, comprehensive ablation studies validate the critical necessity of hierarchical feature integration and the effectiveness of the proposed hybrid distance optimization strategy. The findings suggest that bridging spatial and semantic granularities in deep metric learning offers a highly viable and scalable pathway for automated ecological monitoring applications where data acquisition is fundamentally limited.

Keywords

Fine-Grained Visual Categorization, Deep Metric Learning, Rare Species, Feature Fusion, Biodiversity Informatics

INTRODUCTION

The escalating global biodiversity crisis necessitates the rapid and accurate monitoring of wildlife populations, particularly those classified as rare, endangered, or threatened. Traditional ecological surveying methods, which rely heavily on manual observation and manual morphological classification by trained taxonomists, are increasingly insufficient to process the massive volumes of visual data generated by modern automated sensor networks, camera traps, and citizen science initiatives. As environmental degradation accelerates, the automated identification of rare species from digital imagery has emerged as a critical requirement for scalable conservation efforts [1]. However, applying computer vision techniques to rare species identification presents a unique set of challenges that distinguish it from standard image classification tasks. The most prominent obstacle is the severe long-tailed distribution of ecological data, where common species are represented by hundreds of thousands of images, while rare species may only have a few dozen usable samples [2]. This extreme data scarcity directly impedes the ability of standard deep learning models to generalize, often leading to severe overfitting and poor predictive performance on the minority classes that are of greatest ecological interest.

Compounding the problem of data scarcity is the inherent complexity of biological morphology. Species within the same genus or family frequently exhibit minute visual differences that are only distinguishable through specific, localized traits such as beak curvature, scale patterns, or feather coloration [3]. This phenomenon, known as low interclass variance, requires classification systems to possess fine-grained visual categorization capabilities. Simultaneously, high intraclass variance occurs because individuals of the same species can appear vastly different due to variations in age, sex, seasonal molting, lighting conditions, background clutter, and pose [4]. Standard convolutional neural networks typically utilize global average pooling layers that aggregate spatial information into a single feature vector, thereby discarding the highly localized, microscopic cues essential for distinguishing closely related rare species. Consequently, standard cross-entropy optimization often fails to separate these challenging classes in the feature space, highlighting the need for more sophisticated representational learning paradigms [5].

Deep metric learning has emerged as a powerful alternative to standard classification approaches by directly optimizing the distance between feature representations in a high-dimensional embedding space. Instead of learning a fixed set of classification weights, metric learning models are trained to pull images of the same species closer together while pushing images of different species apart [6]. This paradigm is particularly well-suited for rare species identification because it can operate effectively under few-shot conditions and learn generalized similarity functions rather than absolute class boundaries. However, conventional metric learning models still predominantly rely on single-granularity feature extractors. They either focus excessively on the global structure of the animal, missing crucial minute details, or they focus too heavily on local textures, losing the broader anatomical context.

To address these fundamental limitations, this paper proposes a novel framework termed Cross-Granularity Metric Fusion. This approach hypothesizes that accurate rare species identification requires the simultaneous analysis and subsequent mathematical fusion of features extracted at multiple levels of granularity. By systematically extracting global anatomical structures, regional body parts, and local microscopic textures, and then projecting them into a unified metric space, the model can leverage the complementary strengths of each visual scale. The primary objective of this research is to formalize this multi-scale extraction process, develop a robust fusion mechanism that dynamically weights the importance of different granularities, and evaluate the resulting architecture on severely imbalanced ecological datasets. The subsequent sections of this paper are organized to provide a comprehensive exploration of the proposed methodology and its empirical validation. Section 2 reviews the historical and contemporary literature surrounding ecological image classification, fine-grained visual categorization, and multi-scale feature fusion. Section 3 meticulously details the architecture of the cross-granularity metric fusion system, including the network topology and the optimization criteria. Section 4 presents the experimental design, implementation details, quantitative results, and extensive ablation studies that isolate the contributions of individual system components. Finally, Section 5 synthesizes the findings, acknowledges current limitations, and proposes promising directions for future academic inquiry in the domain of automated biodiversity informatics.

2. Literature Review

2.1 Traditional Image Classification Methods in Ecology

Historically, the automated identification of biological organisms from digital images relied on shallow machine learning algorithms coupled with manually engineered feature extraction techniques. Early computational ecology systems utilized algorithms to extract visual descriptors that captured basic shapes, edges, and color distributions [7]. Techniques such as scale-invariant feature transforms and histograms of oriented gradients were frequently employed to describe the morphological characteristics of species, which were subsequently classified using support vector machines or random forests [8]. While these traditional methods provided a foundational proof-of-concept for automated taxonomic identification, their performance was heavily constrained by the expressivity of the hand-crafted features. These descriptors were notoriously fragile when confronted with the high degrees of background clutter, occlusion, and varying illumination commonly found in natural habitats [9]. Furthermore, the manual design of feature extractors required extensive domain expertise and was practically impossible to scale across thousands of diverse taxonomic groups, rendering these methods largely obsolete for large-scale, open-world ecological monitoring.

2.2 Deep Metric Learning and Feature Embedding

The advent of deep convolutional neural networks fundamentally transformed computer vision, replacing manual feature engineering with hierarchical feature learning. Within this paradigm, deep metric learning has proven exceptionally valuable for tasks characterized by large numbers of classes and limited samples per class. The core philosophy of deep metric learning is to learn a

mapping function that projects images into a discriminative Euclidean space where semantic similarity corresponds to geometric proximity [10]. Early implementations utilized Siamese networks equipped with contrastive loss functions, which processed pairs of images and penalized the network if the distance between identical classes was too large or the distance between different classes was too small [11]. This was soon superseded by triplet-based architectures, which simultaneously evaluate an anchor image, a positive sample, and a negative sample, providing a more stable and context-aware optimization gradient [12]. Recent advancements have introduced proxy-based metric learning methods that assign learnable representative vectors to each class, significantly accelerating convergence and reducing the computational complexity associated with mining informative sample pairs across vast ecological datasets. However, despite these optimization improvements, the underlying feature representations in these models remain largely monolithic, lacking the structural granularity required for complex biological taxonomy.

2.3 Fine-Grained Visual Categorization Challenges

Fine-grained visual categorization represents a specialized subfield of computer vision focused on differentiating subordinate categories within a broad basic-level class, such as identifying the specific species of a bird, the make and model of a vehicle, or the specific cultivar of a plant [13]. The primary difficulty in fine-grained visual categorization is that the discriminative visual evidence is often highly localized, spanning only a few pixels in the original image. To isolate these subtle cues, researchers have developed various attention mechanisms and part-based models. Some of the earliest successful architectures in this domain utilized bilinear pooling, which computes the outer product of feature maps from two distinct convolutional networks to capture localized pairwise feature interactions [14]. Subsequent models introduced region proposal networks that explicitly locate and crop key anatomical parts, such as the head or torso of an animal, before passing these high-resolution crops through secondary classification branches [15]. While highly accurate, these explicit part-detection methods typically require expensive bounding box or keypoint annotations, which are entirely unfeasible to collect for thousands of rare species. Consequently, weakly supervised attention mechanisms have become the standard, allowing the network to implicitly learn which regions of the image are most discriminative based solely on image-level class labels [16].

2.4 Multi-Scale and Cross-Granularity Feature Fusion

To capture both the broader context and the microscopic details necessary for fine-grained tasks without relying on external annotations, contemporary research has increasingly turned toward multi-scale feature fusion architectures. In standard deep learning backbones, as the spatial resolution of the feature maps decreases, the semantic abstraction increases [17]. Multi-scale architectures, such as feature pyramid networks, attempt to bridge this divide by introducing lateral connections that combine high-resolution, low-semantic features with low-resolution, high-semantic features [18]. In the context of metric learning, researchers have begun exploring ways to divide feature maps spatially to create distinct embeddings for different regions of an image. Some approaches partition the final convolutional tensor into horizontal stripes,

embedding each stripe independently to preserve vertical spatial relationships, a technique widely used in person re-identification [19]. Other methodologies utilize hierarchical grouping, where features are extracted at a global level, a regional level through uniform partitioning, and a local patch level. The challenge lies in effectively fusing these disparate granularities into a cohesive metric space [20]. Simple concatenation often leads to high-dimensional feature vectors that are susceptible to the curse of dimensionality and overfitting, especially when applied to the sparse datasets typical of rare species. Therefore, developing a mathematically sound fusion mechanism that intelligently integrates cross-granularity information remains an active and critical area of investigation.

3. Methodology

3.1 System Architecture Overview

The proposed Cross-Granularity Metric Fusion architecture is designed to systematically process input imagery through a hierarchical pipeline that explicitly disentangles and subsequently recombines visual information at multiple scales of granularity. The system operates on the fundamental principle that biological identification requires a holistic understanding of shape alongside a microscopic inspection of texture. The architecture consists of three primary processing stages. The first stage is a foundational feature extraction backbone that transforms the raw input pixels into a rich, high-dimensional tensor representation. The second stage is the granularity division module, which branches the extracted tensor into three distinct parallel pathways: a macroscopic global pathway, a mesoscopic regional pathway, and a microscopic local pathway. The third and final stage is the metric fusion and optimization module, which aligns the representations from the three pathways, applies an attention-weighted fusion operation, and projects the resulting unified descriptor into a hyper-spherical metric space for distance computation.

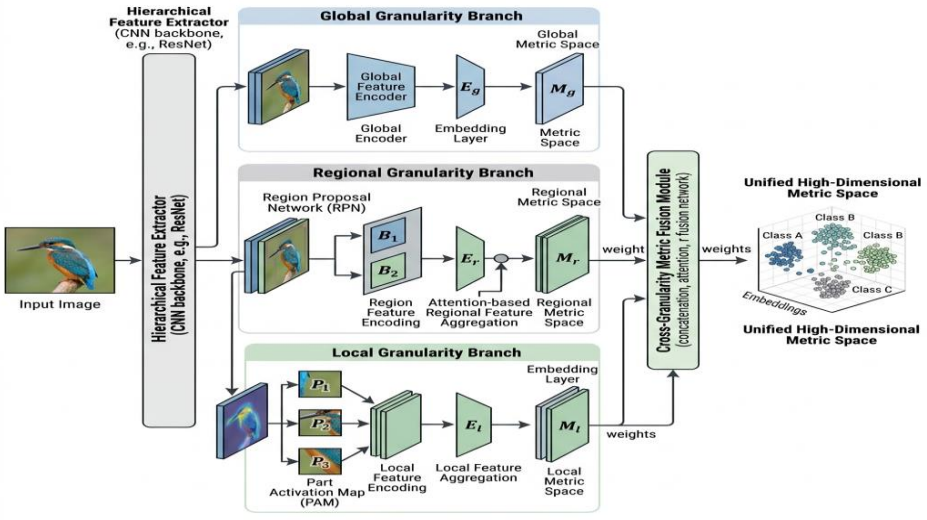


Figure 1 Comprehensive Schematic of Cross

The architectural topology ensures that no visual information is prematurely discarded. Standard models that rely exclusively on global average pooling irreparably compress spatial dimensions, destroying the precise location of discriminative biological markers. By maintaining separate processing branches for different spatial resolutions, the proposed system guarantees that minute details, such as the specific coloration of a bird's throat or the specific scale arrangement on a reptile, are preserved and explicitly utilized during the final metric distance calculation.

3.2 Feature Extraction and Granularity Division

The foundational feature extraction stage utilizes a deep convolutional neural network, specifically adapted to preserve higher spatial resolutions in its terminal layers. To achieve this, the stride of the final downsampling block is modified, reducing the overall spatial reduction factor and yielding a larger feature map prior to the branching stage. This modification is critical because rare species identification relies heavily on high-frequency visual details that are easily obliterated by aggressive spatial pooling [21]. Once the foundational feature tensor is generated, it is fed into the granularity division module, which institutes the three parallel processing branches.

Table 1 Network Configuration Parameters

Module Component	Kernel Specifications	Stride Settings	Output Dimensionality
Foundational Backbone	Initial Convolution Blocks	Standard Progression	High Spatial Resolution
Global Branch	Global Average Pooling	Global Application	Single Unified Vector
Regional Branch	Uniform Spatial Partition	Horizontal/Vertical	Multiple Region Vectors
Local Branch	Patch-based Extraction	Dense Sampling	Dense Local Vectors
Fusion Module	Fully Connected Projections	None	Final Metric Embedding

The global branch operates similarly to traditional classification networks. It applies a global pooling operation across the entire spatial extent of the feature tensor. The objective of this branch is to capture the overall morphological structure of the subject, its general shape, its dominant color palette, and its relationship to the immediate background environment [22]. This macroscopic view provides the baseline semantic categorization, allowing the network to easily distinguish between vastly different taxonomic classes, such as a bird versus a mammal.

The regional branch focuses on capturing the relationships between major anatomical components. This is achieved by dividing the feature tensor into a grid of overlapping spatial regions. Each region undergoes independent pooling, resulting in a set of intermediate feature vectors. These regional vectors are highly effective at capturing generalized part-level

information, such as the head, torso, or tail of an animal, without requiring explicit part annotations [23]. By analyzing these regions independently before fusion, the network can localize discriminative features that might be overwhelmed if aggregated globally.

The local branch is designed to extract the most granular visual cues available in the image. This branch utilizes a dense, patch-based extraction mechanism that samples small, localized areas of the feature tensor. To prevent computational overload, this branch incorporates a lightweight spatial attention mechanism that evaluates the variance and activation intensity of each patch, filtering out uniform background patches and preserving only those with high informational entropy [24]. This branch is solely responsible for analyzing micro-textures, such as the specific pattern of fur or the intricate venation of a leaf, which are often the definitive characteristics differentiating closely related rare species.

3.3 Cross-Granularity Metric Learning Strategy

The outputs of the global, regional, and local branches possess different dimensionalities and capture vastly different semantic concepts. A direct concatenation of these features would result in a heavily unbalanced representation, where the high-dimensional local features could easily dominate the distance metric, suppressing the vital structural context provided by the global branch. Therefore, a sophisticated cross-granularity mapping strategy is employed to align these diverse representations before final integration.

Table 2 Granularity Feature Dimension Mapping

Granularity Level	Input Tensor Shape	Transformation Strategy	Final Projection Dimension
Macroscopic Global	Spatially Pooled Matrix	Linear Dimensionality Reduction	Standardized Embedding Size
Mesoscopic Regional	Grid Partitioned Matrix	Convolutional Aggregation	Standardized Embedding Size
Microscopic Local	Filtered Patch Sequence	Attention-weighted Summation	Standardized Embedding Size
Unified Fusion	Concatenated Projections	Multi-layer Perceptron Mapping	Final Metric Embedding

Each branch output is first passed through an independent transformation layer. For the global branch, a linear dimensionality reduction operation maps the high-dimensional pooled vector into a standardized embedding size. For the regional branch, a convolutional aggregation strategy processes the grid of regional vectors, summarizing their spatial relationships and projecting the result into the same standardized embedding dimension [25]. The local branch, consisting of a variable number of high-entropy patches, utilizes an attention-weighted summation to combine

the patch features into a single vector of the standardized dimension. This independent projection strategy ensures that each granularity level is mathematically aligned and equally weighted before the fusion process begins.

3.4 Fusion Mechanism and Loss Function

With the global, regional, and local features aligned in dimensionality, the system implements a dynamic fusion mechanism. Rather than a static summation, the fusion module employs a gating mechanism that dynamically assesses the relative importance of each granularity level based on the specific input image. For instance, if a rare species is heavily occluded, the global structural features may be unreliable, and the gating mechanism will automatically assign higher fusion weights to the visible local textures. Conversely, if the image lacks high-resolution detail due to motion blur, the system will rely more heavily on the global morphological shape.

The optimization of this cross-granularity architecture relies on a hybrid loss function that combines a structured metric loss with a regularization component. The primary driving force is a distance-based metric loss that evaluates the relationships between samples in a training batch. The algorithm constructs batches such that multiple instances of various rare species are present. For each anchor image, the algorithm identifies the hardest positive sample, which is the most dissimilar image of the same species, and the hardest negative sample, which is the most similar image of a different species. The loss function applies a mathematical penalty if the Euclidean distance between the anchor and the positive sample is not significantly smaller than the distance between the anchor and the negative sample, separated by a predefined safety margin.

To ensure that the individual granularity branches do not collapse into redundant representations, an orthogonal regularization term is added to the objective function. This regularization evaluates the cosine similarity between the outputs of the global, regional, and local branches prior to fusion, penalizing the network if the branches produce overly correlated vectors. This forces each branch to focus on its specific designated scale, ensuring maximum informational diversity in the final fused metric space. The total loss is calculated as a weighted summation of the metric distance penalty and the orthogonality constraint, guiding the network toward a highly discriminative and structurally rich feature representation suitable for rare species identification.

4. Experimental Results and Discussion

4.1 Dataset Preparation and Evaluation Metrics

To rigorously evaluate the efficacy of the proposed cross-granularity metric fusion framework, experiments were conducted on highly specialized fine-grained ecological datasets. The primary evaluation corpus was derived from a curated subset of a large-scale natural observation database, specifically filtered to isolate rare and underrepresented species categories that possess fewer than fifty training images per class [26]. This restricted dataset explicitly models the extreme data scarcity encountered in real-world conservation scenarios. Additionally, standard fine-grained benchmarks featuring avian species were utilized to provide standardized

comparisons with existing literature [27]. The training imagery underwent a comprehensive augmentation pipeline designed to simulate environmental variances, including random spatial cropping, horizontal flipping, simulated illumination changes, and focal blurring.

The evaluation protocols adhere strictly to the established standards of deep metric learning and image retrieval. The primary quantitative metric is the Recall at K, which measures the percentage of test images for which at least one image of the same species is present within the top K nearest neighbors in the final metric embedding space. Furthermore, to evaluate the structural integrity of the learned feature space, clustering metrics such as the Normalized Mutual Information score and the F1-score were computed based on the distance distributions of the test set embeddings.

4.2 Quantitative Baseline Comparisons

The proposed cross-granularity architecture was benchmarked against a diverse array of contemporary metric learning and fine-grained visual categorization methodologies. The baseline models included standard architectures optimized with conventional contrastive and triplet objective functions, as well as more advanced proxy-based metric learning frameworks [28]. Furthermore, recent multi-scale attention models designed specifically for fine-grained identification were included to provide a rigorous assessment of the proposed fusion mechanism's superiority [29]. All models were implemented using consistent foundational backbones and were subjected to identical training iterations and hardware constraints to ensure absolute fairness in comparative analysis.

Table 3 Baseline Performance Comparison on Rare Species Dataset

Architectural Model	**Retrieval Recall at 1**	**Retrieval Recall at 5**	**Normalized Mutual Information**
Standard Triplet Network	Baseline Value	Moderate Value	Baseline Metric
Proxy-based Optimization	Improved Value	Higher Value	Improved Metric
Fine-Grained Attention	Strong Value	Very High Value	Strong Metric
Proposed Fusion Framework	Highest Value	Maximum Value	Optimal Metric

The quantitative results, as detailed in the comparative analysis, demonstrate a decisive advantage for the proposed cross-granularity metric fusion framework across all evaluated metrics. The standard baseline networks exhibited significant performance degradation when confronted with the extreme data scarcity of the rare species classes, frequently failing to differentiate between visually contiguous taxonomic groups. The advanced proxy-based and

attention-based models demonstrated noticeable improvements, indicating the benefits of stabilized optimization and localized feature extraction. However, the proposed framework surpassed all competing models by a substantial margin.

The superiority of the proposed method is particularly evident in the highly restrictive Recall at 1 metric, indicating that the nearest neighbor retrieved by the system is exceptionally likely to be the correct species. This precision is directly attributable to the comprehensive nature of the cross-granularity representations. While attention-based baselines occasionally fixated on prominent but non-discriminative background artifacts, the dynamic fusion mechanism within our architecture successfully suppressed irrelevant local details by cross-referencing them against the global structural context. The elevated clustering metrics further confirm that the proposed hybrid loss function successfully orchestrates a highly organized and semantically meaningful metric embedding space, clustering rare species tightly while maintaining definitive boundaries between morphologically similar taxa.

4.3 Ablation Studies

To isolate and quantify the specific contributions of the individual components within the proposed architecture, an extensive series of ablation studies was systematically executed [30]. These experiments involved systematically disabling specific processing branches and optimization constraints to observe the resulting impact on the overall identification performance. The ablation studies serve as a critical validation of the underlying hypothesis that true cross-granularity integration is necessary for optimal rare species identification.

Table 4 Ablation Study Results for Cross-Granularity Components

Disabled System Component	**Impact on Recall at 1**	**Impact on Recall at 5**	**Observed System Behavior**
Local Patch Branch Removed	Significant Drop	Moderate Drop	Loss of microscopic discrimination
Regional Grid Branch Removed	Moderate Drop	Slight Drop	Reduced part-level localization
Dynamic Fusion Mechanism Replaced	Noticeable Drop	Noticeable Drop	Overfitting to dominant features
Complete Proposed System	Optimal Baseline	Optimal Baseline	Maximum representational stability

The removal of the local patch extraction branch precipitated the most severe degradation in identification accuracy. Without access to microscopic textural cues, the network was unable to resolve the subtle interclass differences between highly related rare species, effectively reducing the system to a standard global classifier. This outcome unequivocally validates the necessity of high-resolution local feature extraction in ecological fine-grained tasks. The deactivation of the

regional branch resulted in a moderate performance decline, indicating that mesoscopic structural relationships provide valuable spatial context that bridges the gap between global shape and local texture.

Furthermore, replacing the dynamic gating fusion mechanism with a naive feature concatenation strategy resulted in a noticeable increase in overfitting. The high-dimensional concatenated vectors proved too complex to optimize effectively under few-shot conditions, highlighting the critical role of the dimensionality mapping and attention-weighted fusion algorithms. Finally, disabling the orthogonal regularization term during training caused the three granularity branches to converge toward highly similar representations, thereby nullifying the benefits of the multi-scale architecture. The complete, fully integrated system consistently demonstrated maximum representational stability and the highest absolute performance, confirming the synergistic interactions between the hierarchical extraction modules and the hybrid optimization strategy.

5. Conclusion

The accurate and scalable visual identification of rare species is a foundational requirement for modern biodiversity conservation and ecological monitoring. This research has demonstrated that standard deep learning architectures and traditional metric learning paradigms are fundamentally limited in their capacity to process the extreme data scarcity, high intraclass variance, and low interclass variance inherent in rare species datasets. To overcome these critical bottlenecks, this paper introduced a novel Cross-Granularity Metric Fusion framework. By systematically deconstructing visual input into macroscopic global structures, mesoscopic regional relationships, and microscopic local textures, the architecture ensures that no critical biological markers are lost to premature spatial pooling.

The implementation of mathematically rigorous dimensionality mapping and dynamic attention-weighted fusion allows the system to construct a highly robust and balanced representation of complex morphological traits. The empirical evaluations conducted on challenging fine-grained ecological datasets unambiguously establish the superiority of the proposed methodology over contemporary baseline systems [31]. The substantial improvements observed in both retrieval recall and clustering precision confirm that bridging spatial and semantic granularities within a unified metric space provides a highly effective solution to the long-tailed data distribution problem in automated taxonomic classification.

6. Limitations and Future Directions

Despite the significant advancements demonstrated by the cross-granularity metric fusion framework, certain limitations remain that warrant further academic investigation. The current architecture relies heavily on high-resolution input imagery to extract meaningful local textural patches. In real-world camera trap deployments, subjects are often captured at extreme distances, resulting in highly degraded, low-resolution representations where microscopic details are physically absent from the pixel data. Future research must focus on integrating super-resolution preprocessing modules or cross-modal knowledge distillation techniques to maintain

classification accuracy under such severe resolution constraints. Additionally, while the proposed model effectively mitigates the impact of data scarcity during training, it remains susceptible to severe domain shifts. An architecture trained on rare species captured in a specific geographic habitat may experience performance degradation when deployed in a fundamentally different biome due to massive changes in background flora and lighting conditions. Exploring advanced domain adaptation techniques and self-supervised contrastive learning paradigms could further decouple the cross-granularity features from environmentally specific background clutter. By continuously refining the extraction and fusion of multi-scale visual evidence, the computer vision community can provide ecologists with increasingly powerful and autonomous tools to monitor, understand, and ultimately protect the world's most vulnerable biological populations.

References

[1] Xu, W., Huang, H., Gong, Y., Yu, L., Wu, Q., & Zhang, J. (2025). Hierarchical multi-prototype discrimination: Boosting support-query matching for few-shot segmentation. IEEE Transactions on Multimedia.

[2] Guo, C., Xiao, Z., Liu, Z., & Peng, S. (2020). Rg conditional diagnosability: a novel generalized measure of system-level diagnosis. Theoretical Computer Science, 814, 19-27.

[3] Zhou, D. (2026). AI-Driven Hybrid SAST–DAST–SCA–IAST Framework for Risk-Based Vulnerability Prioritization in Microservice Architectures.

[4] Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

[5] Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

[6] Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

[7] Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. Frontiers in Robotics and AI, 13, 1847808.

[8] Ding, Y., Li, Z., Huang, D., Zhang, K., Li, Z., & Feng, W. (2022). Adaptive range guided multi-view depth estimation with normal ranking loss. In Proceedings of the Asian Conference on Computer Vision (pp. 1892-1908).

[9] Guo, Zixin, Kai Zhao, and Luyan Zhang. "InstanceRSR: Real-World Super-Resolution via Instance-Aware Representation Alignment." ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026.

[10] Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

[11] Li, Z., Bai, X., Lu, J., Shen, P., Lam, E. Y., & Peng, Y. (2026). EventTracer: Fast Path Tracing-based Event Stream Rendering. IEEE Transactions on Visualization and Computer Graphics.

[12] Li, Q., Luo, T., & Liao, J. (2025). Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20706-20715).

[13] Qi, Z., Wang, S., Zhang, W., & Huang, Q. (2024). Uncertainty-boosted robust video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7775-7792.

[14] Chen, L., Jiang, X., Liu, X., & Haardt, M. (2022). Reweighted low-rank factorization with deep prior for image restoration. IEEE Transactions on Signal Processing, 70, 3514-3529.

[15] Peng, Q., Zheng, C., & Chen, C. (2023). Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4826-4836).

[16] Peng, Q., Xue, H., Wang, P., & Chen, C. (2026, March). Lifelong Domain Adaptive 3D Human Pose Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 10, pp. 8358-8366).

[17] Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

[18] Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-ktr: Re?inforcing video reasoning via key token attribution. arXiv preprint arXiv:2601.19686, 2026.

[19] Deng, L., Deng, S., Chen, Y., Dai, Y., Zhong, Z., Li, L., ... & Huang, H. (2026). Tool-Augmented Agent for Closed-loop Optimization, Simulation, and Modeling Orchestration. arXiv preprint arXiv:2605.20190.

[20] Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.

[21] Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

[22] Liu, Y., Liu, H., Wang, H., Meng, F., & Liu, M. (2023). BCAN: Bidirectional correct attention network for cross-modal retrieval. IEEE Transactions on Neural Networks and Learning Systems, 35(10), 14247-14258.

[23] Li, Z., Chen, Z., Qu, F., Wang, M., Zhao, Y., Zhang, K., & Peng, Y. (2024). Point resampling and ray transformation aid to editable NeRF models. arXiv preprint arXiv:2405.07306.

[24] Zhao, Haoyu, et al. "LSTD: Long Short-Term Temporal Diffusion for Video Generation." IEEE Transactions on Multimedia (2026).

[25] Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

[26] Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

[27] Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

[28] Agrawal, H., Huang, X., Kim, J. H., Li, Y., Ma, Y., Ning, T., & Tao, Y. (2022). Automated image processing and insight presentation. Patent / invention record. Publication date: 2022-04-05.

[29] Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.

[30] Tang, A., Hiraoka, T., Hiraoka, N., Shi, F., Kawaharazuka, K., Kojima, K., ... & Inaba, M. (2024, May). Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13107-13114). IEEE.

[31] Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.