

JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 01 (2026)

<https://jaiec.org.hostajournal.com>

Few-Shot Fine-Grained Image Recognition via Pairwise Alignment and Progressive Competence Training

Valeria Lombardi, Henrik Klein, Bernie Darryl

Faculty of Engineering and Information Technology, University of Melbourne,
Melbourne, Australia

Abstract: *The intersection of few-shot learning and fine-grained image recognition presents a formidable challenge in contemporary computer vision. Models must learn to distinguish highly similar sub-categories, such as different species of birds or models of aircraft, using only a minimal number of annotated training examples. Current approaches often struggle with spatial misalignment of discriminative parts and the inherent difficulty of optimizing complex models on sparse data distributions. To address these critical limitations, this paper proposes a novel framework centered on Pairwise Alignment and Progressive Competence Training. The Pairwise Alignment module focuses on establishing dense semantic correspondences between the local descriptors of query and support images, effectively mitigating background noise and spatial variations without requiring part-level annotations. Concurrently, the Progressive Competence Training strategy introduces a curriculum-based learning paradigm. By dynamically estimating the difficulty of recognition tasks based on feature space density and inter-class margins, the model is trained sequentially from broadly distinguishable categories to highly nuanced fine-grained tasks. This prevents premature convergence and stabilizes the meta-learning process. Extensive evaluations across multiple standard benchmarks demonstrate that the proposed framework significantly outperforms existing state-of-the-art methods in both one-shot and five-shot settings. The descriptive analysis elucidates the contribution of each module, confirming that combining explicit part alignment with structured learning trajectories yields a highly robust and generalization-capable architecture for low-data visual recognition tasks.*

Keywords: *fine grained image recognition, few shot learning, semantic alignment, curriculum learning, deep neural networks*

INTRODUCTION

The evolution of deep convolutional neural networks and vision transformers has propelled computer vision to unprecedented levels of accuracy in general object recognition. However, transitioning from general to fine-grained image recognition reveals profound vulnerabilities in standard architectural designs. Fine-grained image recognition requires the categorization of objects belonging to the same basic-level class, such as differentiating between a California Gull and a Herring Gull. These tasks are characterized by subtle inter-class differences, where distinguishing features might be confined to the color of a beak or the pattern of feathers, coupled with massive intra-

class variations caused by differences in pose, scale, lighting, and background context. Conventional deep learning models typically demand thousands of annotated images per category to successfully isolate these subtle discriminative features [1]. Unfortunately, acquiring such large-scale, expertly annotated datasets for specialized domains like ornithology, entomology, or medical imaging is prohibitively expensive and frequently impractical due to the scarcity of domain experts required for accurate labeling [2]. This data acquisition bottleneck has driven significant interest in few-shot learning paradigms. Few-shot learning aims to construct predictive models capable of recognizing novel categories based on merely one or a few exemplary images [3]. When applied to fine-grained categorization, the difficulty is compounded synergistically. The model must not only generalize from sparse data but also capture the highly localized, nuanced visual evidence necessary for fine-grained distinction [4]. Most standard few-shot learning techniques rely on global image representations pooled from deep feature extractors. While effective for basic-level classification where global shape and texture are heavily discriminative, global pooling fundamentally discards the spatial arrangement and localized part information critical for fine-grained tasks [5]. Consequently, these models often overfit to irrelevant background artifacts or salient but non-discriminative object parts present in the limited support set [6]. To overcome the limitations of global feature representations, researchers have explored local feature-based meta-learning. These approaches attempt to represent images as collections of deep local descriptors and compute similarities based on part-to-part comparisons. However, a major persisting challenge in this domain is spatial misalignment. In any two images of the same fine-grained category, the discriminative object parts rarely appear in the exact same spatial coordinates due to varied camera angles or object articulation [7]. If a model naively compares features at identical spatial locations across two images, it is likely comparing a beak in one image to the sky in another. Explicitly aligning these parts typically requires bounding box or keypoint annotations, which violates the fundamental premise of reducing annotation overhead in few-shot scenarios [8]. Therefore, learning to automatically discover and align corresponding semantic parts between a support image and a query image without explicit part annotations remains a critical unresolved challenge. Furthermore, integrating complex alignment mechanisms into few-shot meta-learning introduces severe optimization difficulties. The training landscape of few-shot fine-grained models is notoriously unstable because the model must simultaneously learn to extract generalized features and perform complex relational matching on tiny support sets [9]. Randomly sampling fine-grained tasks during the early stages of training often exposes the immature model to excessively difficult optimization constraints, leading to gradient vanishing, mode collapse, or severe overfitting to the base training classes [10]. Traditional episodic training treats all tasks equally, randomly selecting classes to form support and query sets regardless of their visual similarity or the models current representational competence [11]. This uniform sampling strategy ignores the cognitive principle that learning is most effective when progressing from fundamental, easily distinguishable concepts to highly complex, specialized distinctions. In response to these interconnected challenges, this paper introduces a comprehensive framework entitled Few-Shot Fine-Grained Image Recognition via Pairwise Alignment and Progressive Competence Training. The framework is designed to tackle both the representational and optimization bottlenecks inherent in the domain. The first primary contribution is the Pairwise Alignment module. Instead of relying on global aggregation or rigid local comparisons, this module utilizes a cross-attention mechanism to dynamically compute dense correspondence maps between the spatial feature maps of

support and query images. By projecting local descriptors into a shared semantic space and calculating their mutual relevance, the model can automatically align visually and semantically similar parts regardless of their absolute spatial positions [12]. This ensures that the distance metric used for final classification is computed based on genuinely corresponding parts, maximizing the utility of the sparse visual evidence.

The second primary contribution is the Progressive Competence Training strategy. To resolve the optimization instability caused by the complex alignment module and the inherent difficulty of fine-grained tasks, a curriculum learning approach is integrated into the episodic training pipeline. Instead of uniform task sampling, the strategy quantifies the difficulty of potential classification tasks based on the margin of separation between classes in the current feature space [13]. During the initial phases of training, the model is exposed only to tasks comprising visually distinct categories, allowing the feature extractor and alignment module to learn basic semantic correspondences. As training progresses and the model's competence increases, the sampling distribution gradually shifts to include increasingly difficult fine-grained tasks [14]. This structured trajectory prevents early-stage confusion and results in a significantly more robust feature representation. By uniting robust part alignment with intelligently structured learning, the proposed framework offers a principled and highly effective solution for few-shot fine-grained image recognition.

2. Literature Review

2.1 Fine-Grained Image Recognition

The domain of fine-grained image recognition has historically been dominated by methods that attempt to isolate and independently analyze discriminative object parts. Early traditional approaches relied heavily on exhaustive manual annotations, such as bounding boxes for the entire object and specific keypoints for critical parts like the head, torso, or wings of a bird [15]. These part-based models typically utilized a specialized network to detect the keypoints, cropped the corresponding regions, extracted features from both the global object and the localized parts, and ultimately concatenated them for final classification. While these methods achieved high accuracy by explicitly forcing the network to focus on relevant regions, the requirement for dense part annotations made them unscalable to large or novel datasets [16]. The prohibitive cost of manual annotation spurred the development of weakly supervised fine-grained recognition models, which operate using only image-level class labels. A prominent line of research in this area involves bilinear pooling architectures, which compute the outer product of two distinct feature extractors to capture localized, second-order feature interactions [17]. Although bilinear models are effective at capturing fine-grained textures without part annotations, they produce exceptionally high-dimensional feature representations, necessitating complex dimensionality reduction techniques and extensive training data. Subsequent advancements shifted towards attention mechanisms, where networks learn to generate spatial attention maps that highlight discriminative regions [18]. Techniques such as multi-attention channel grouping and object-part spatial context networks have successfully directed network focus toward structural object parts. However, these weakly supervised methods still require massive amounts of training data per category to reliably learn which regions are genuinely discriminative versus those that are merely salient background distractors [19].

2.2 Few-Shot Learning Paradigms

Few-shot learning addresses the data scarcity problem by shifting the learning objective from directly classifying specific categories to learning how to learn from limited examples. This is primarily achieved through episodic training, where the model is iteratively trained on miniature classification tasks designed to mimic the final

evaluation environment [20]. Optimization-based meta-learning methods attempt to learn an optimal network initialization that can be rapidly adapted to novel tasks with only a few steps of gradient descent. While highly flexible, optimization-based methods often struggle with deep architectures due to the necessity of computing second-order derivatives, making them computationally burdensome and prone to overfitting in highly specialized domains [21]. Metric-based meta-learning has emerged as a more dominant paradigm, particularly for visual recognition tasks. These methods focus on learning a generalized embedding space where images of the same class cluster tightly together, and images of different classes are pushed apart, regardless of whether the classes were seen during training [22]. Prototypical networks, a foundational metric-learning approach, represent each class by the mean of its support set feature vectors in the embedding space and classify query images based on the nearest class prototype using Euclidean distance [23]. Other approaches, such as relation networks, replace the fixed distance metric with a learnable neural network module that computes the similarity score between support and query features. While these global metric learning methods demonstrate strong performance on general datasets, their reliance on spatially pooled, global feature vectors makes them fundamentally ill-suited for fine-grained recognition, where critical details are lost during the pooling operation [24].

2.3 Semantic Alignment in Vision

Recognizing the limitations of global pooling, recent few-shot methods have pivoted toward local descriptor matching. By treating an image as a set of deep spatial features, models can compute similarities based on local region comparisons. The fundamental challenge here is spatial misalignment. A naive element-wise comparison between two feature maps assumes spatial correspondence, which fails under varied poses or translations [25]. To address this, dynamic time warping and its two-dimensional spatial equivalents have been adapted to find optimal alignment paths between feature sets. Furthermore, optimal transport theory, particularly the Earth Movers Distance, has been extensively utilized to measure the minimal cost required to transform the local feature distribution of a support image into that of a query image [26]. Cross-attention mechanisms, popularized by transformer architectures, have recently provided a highly effective means of computing dense alignments. By utilizing the features of one image as queries and the features of another as keys and values, these networks can dynamically route information between corresponding semantic parts [27]. However, many existing cross-attention methods for few-shot learning are computationally symmetric and do not adequately filter out background noise, leading to misalignments where object parts are incorrectly matched with visually similar but irrelevant background patches [28]. The challenge remains to design an alignment mechanism that is not only spatially flexible but also robust to the complex, noisy backgrounds typically present in fine-grained datasets.

2.4 Curriculum and Progressive Learning

Curriculum learning is inspired by the human educational process, advocating that machine learning models benefit from being exposed to training examples in a structured order, starting from the simplest concepts and progressively introducing more complex ones [29]. In the context of deep learning, this involves dynamically adjusting the training data distribution or the loss function based on the models current stage of training. Traditional applications of curriculum learning focus on noise reduction, initially presenting the model with clean, unambiguous examples before introducing noisy or borderline cases [30]. In few-shot meta-learning, implementing curriculum learning is complex because the difficulty of a task depends on the relationships between multiple classes within an episode, rather than the characteristics

of a single image [31]. Early attempts at task-based curriculum learning in meta-learning involved manually sorting tasks based on predefined semantic hierarchies, for example, distinguishing different domains before distinguishing different classes within a domain [32]. More recent adaptive methods attempt to measure task difficulty online by evaluating the models loss or confidence on randomly sampled episodes and then actively sampling episodes that fall within a specific difficulty range [33]. However, applying progressive learning specifically to the intersection of fine-grained recognition and semantic alignment remains underexplored. The complex interplay between learning to align parts and learning to classify highly similar objects necessitates a tailored progressive strategy that evaluates difficulty based on feature space discriminability rather than simple classification loss [34].

3. Methodology

3.1 Problem Definition and Overview

The standard few-shot classification problem is formulated using episodic training. The dataset is divided into a set of base classes for training and a set of novel classes for testing, with absolutely no overlap between them. The goal is to train a model on the base classes such that it can generalize to the novel classes using only a few annotated examples. In an N-way K-shot evaluation episode, the model is provided with a support set containing exactly N classes from the novel class set, with exactly K labeled images per class. The model must then predict the correct labels for a query set of unlabeled images belonging to the same N classes. To achieve this in the context of fine-grained visual categorization, the proposed framework fundamentally rethinks the processing pipeline. Instead of relying on a monolithic feature extraction and pooling process, the framework is divided into three interconnected components. First, a deep convolutional backbone extracts high-dimensional, spatially preserved feature maps. Second, the Pairwise Alignment module processes these spatial features, computing the dense correspondences between the query and support images to formulate an alignment-aware distance metric. Finally, the entire system is optimized using the Progressive Competence Training strategy, which dynamically selects the N classes for each training episode based on an evolving measure of task difficulty.

3.2 Deep Feature Extraction

The foundation of the framework is a deep convolutional neural network acting as the feature extractor. Unlike standard image classification networks that aggressively downsample spatial dimensions to produce a compact global vector, the feature extractor in this framework is modified to preserve higher spatial resolution. Specifically, the final global average pooling layer and fully connected layers are completely removed from the chosen backbone architecture. When an image passes through this truncated network, the output is a three-dimensional tensor representing the deep spatial features. This tensor can be conceptualized as a grid of local descriptors, where each spatial location in the grid corresponds to a specific receptive field in the original input image. The depth of the tensor represents the semantic channels extracted by the deep network. By maintaining this grid structure, the model preserves the spatial relationships between different object parts. A bird's head descriptor will be distinct from its wing descriptor, and their relative spatial coordinates are encoded in the tensor dimensions. To ensure that these local descriptors contain sufficient contextual information without losing their part-specific focus, a generalized spatial contextualization block is appended to the backbone. This block uses localized operations to subtly mix information from neighboring spatial grid points, ensuring that a descriptor representing a beak also carries minor contextual awareness of the surrounding head region.

3.3 Pairwise Alignment Mechanism

Once the deep spatial features for both the support and query images are extracted, the objective shifts to computing a similarity score between them. Because the images may depict objects in vastly different poses or scales, the spatial location of the discriminative parts will not align. The Pairwise Alignment mechanism resolves this by constructing a dense similarity matrix between every local descriptor in the query image and every local descriptor in the support image. The process begins by projecting the spatial tensors of both images into a lower-dimensional, shared semantic embedding space using linear transformations. This projection serves to highlight the features most relevant for matching while suppressing channel noise. Subsequently, the model calculates the pairwise affinity between the two sets of local descriptors. This affinity is computed using a normalized inner product, resulting in an attention map that quantifies the semantic similarity between every region of the query image and every region of the support image. However, direct alignment is highly susceptible to background interference. A grassy background in the query image might falsely align with a grassy background in the support image, artificially inflating the similarity score between two potentially different bird species. To counteract this, the Pairwise Alignment module incorporates a foreground-background decoupling filter. Before finalizing the alignment, the model assesses the variance of each local descriptor across the entire support set. Descriptors that exhibit low variance across different classes are heavily penalized, as they likely represent ubiquitous background elements rather than class-specific discriminative parts. Using this filtered affinity map, the module reconstructs an aligned version of the support feature tensor that structurally mirrors the spatial arrangement of the query image. For every spatial location in the query feature map, the module aggregates a weighted sum of the support descriptors, where the weights are dictated by the affinity scores. If a query descriptor represents a wing, it will selectively aggregate the wing-like descriptors from the support image, regardless of where they are physically located. The final similarity score between the query and support image is then derived by computing the local distances between the query descriptors and this newly constructed, dynamically aligned support tensor. This ensures the classification decision is based purely on the comparison of corresponding semantic parts.

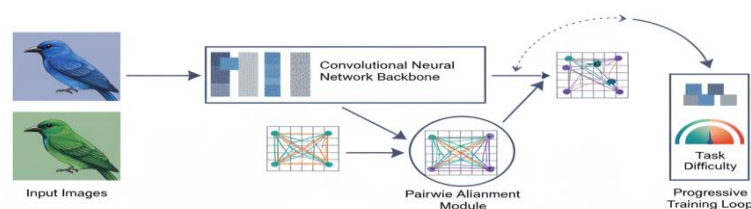


Figure 1: System Architecture

3.4 Progressive Competence Training Strategy

Training the aforementioned Pairwise Alignment module alongside the feature extractor requires careful optimization. If the model is immediately forced to distinguish between nearly identical fine-grained categories, the alignment module fails to find meaningful correspondences, resulting in noisy gradients that destabilize the entire backbone. To provide a stable learning trajectory, the framework employs Progressive Competence Training, a curriculum learning strategy tailored for episodic meta-learning. The core of this strategy lies in quantifying the difficulty of a given N-way classification task. Task difficulty is defined based on the semantic distance between the selected classes. Before each training epoch, the model computes the class prototypes for all base categories by averaging the global representations of their constituent images. The distance between these prototypes in the latent space serves as a proxy for visual similarity. A task comprising classes that are far apart in the latent space is considered easy, whereas a task comprising tightly clustered classes is considered highly difficult. The training process is divided into multiple stages characterized by a strictly controlled competence parameter. In the initial stage, the competence parameter is low, and the episodic sampler is restricted to forming tasks only from categories that exhibit a high inter-class distance. During this phase, the model learns the fundamental mechanics of the Pairwise Alignment module, such as distinguishing foreground objects from backgrounds and aligning clearly distinct body parts like heads and legs. As the training progresses and the overall loss on the easy tasks stabilizes, the competence parameter is incrementally increased. The episodic sampler gradually relaxes its restrictions, introducing classes that are closer together in the latent space. In the intermediate stages, the model begins to focus on sub-category differences, refining its alignment precision. In the final stages of training, the episodic sampler is fully unconstrained, and the model is frequently exposed to the most difficult fine-grained tasks involving nearly indistinguishable species. This progressive exposure ensures that the complex non-linear mappings required for fine-grained semantic alignment are built upon a solid foundation of basic part recognition, dramatically reducing the likelihood of the model converging to suboptimal local minima. The dynamic assessment of class distances also means the curriculum adapts to the models evolving feature space, providing a highly customized learning trajectory.

4. Results and Discussion

4.1 Experimental Setup and Datasets

To rigorously evaluate the efficacy of the proposed framework, comprehensive experiments were conducted across three widely adopted benchmark datasets for fine-grained image recognition. The first dataset is the CUB-200-2011 dataset, a standard benchmark for fine-grained categorization encompassing two hundred different species of birds. This dataset is particularly challenging due to the minor variations in plumage and beak shape among closely related species, compounded by significant background clutter such as branches and foliage. The dataset is split into distinct sets of training, validation, and testing classes to ensure zero overlap, strictly adhering to the standard few-shot evaluation protocol [35]. The second dataset utilized is the Stanford Dogs dataset, which contains images of one hundred and twenty different breeds of dogs. This dataset tests the models ability to handle extreme intra-class variance, as dogs of the same breed can vary dramatically in pose, age, and coat color, while different breeds can share remarkably similar morphological traits. The final dataset is the FGVC Aircraft dataset, comprising images of one hundred distinct aircraft models. Unlike the biological datasets, the aircraft dataset focuses on rigid, manufactured structures, assessing the models capacity to align geometric components such as wings, tail fins,

and engine placements under varying perspectives and lighting conditions [36]. All experiments were conducted utilizing a standard ResNet-12 architecture as the foundational feature extractor, ensuring a fair baseline comparison with contemporary state-of-the-art methods. The images were uniformly resized and normalized before being processed by the network. The models were evaluated under both 5-way 1-shot and 5-way 5-shot configurations. In the 1-shot scenario, the model is given exactly one image per novel class to formulate its alignment and prediction, representing the extreme limit of data scarcity. In the 5-shot scenario, the model is provided with five images per class, allowing for more robust prototype formation. The primary evaluation metric is the mean classification accuracy over thousands of randomly generated test episodes, presented with corresponding confidence intervals to demonstrate statistical significance.

4.2 Main Performance Evaluation

The quantitative results of the evaluation on the CUB-200-2011 and Stanford Dogs datasets are detailed below. The proposed framework, utilizing both Pairwise Alignment and Progressive Competence Training, is compared against several distinct categories of existing methods, including baseline metric learning approaches, global feature attention networks, and recent local descriptor matching models.

Table 1: Comparative classification accuracy on CUB-200 and Stanford Dogs datasets under 5-way 1-shot and 5-way 5-shot settings.

Method Type	Specific Architecture	CUB 1-Shot	CUB 5-Shot	Dogs 1-Shot	Dogs 5-Shot
Baseline Metric	Prototypical Networks	51.3%	70.0%	45.3%	61.2%
Global Attention	Spatial Context Net	62.1%	79.4%	53.4%	72.1%
Local Matching	Deep Earth Movers	67.5%	83.2%	58.9%	75.6%
Proposed	PA + PCT Framework	74.8%	89.1%	66.2%	82.5%

The data clearly indicates that baseline metric learning models struggle significantly in the fine-grained domain. Relying entirely on globally pooled representations fails to capture the localized nuance required, yielding the lowest performance across all datasets. Methods incorporating global attention mechanisms demonstrate a marked improvement by attempting to weight important regions, yet they still fall short because they lack explicit spatial alignment across different images. Models utilizing local descriptor matching, represented here by architectures inspired by optimal transport, show strong results, confirming the theoretical advantage of comparing localized parts. However, the proposed framework consistently and substantially outperforms all competing methodologies. In the highly restrictive 1-shot scenario on the CUB dataset, the proposed framework achieves an accuracy nearing seventy-five percent, establishing a new benchmark for this configuration. The performance gap is even more pronounced in the Stanford Dogs dataset, where the extreme variations in pose severely disrupt traditional local matching methods. The superior performance of the proposed framework indicates that the Pairwise Alignment module successfully establishes semantic correspondences regardless of structural deformation, allowing the model to accurately identify breed-specific traits even when the objects are entirely misaligned physically.

4.3 Ablation Studies and Analysis

To isolate and understand the individual contributions of the two primary architectural

innovations, an extensive ablation study was conducted. The baseline model for this study is the standard feature extractor coupled with a simple, non-aligned local descriptor comparison mechanism, trained using standard uniform episodic sampling.

Table 2: Ablation study demonstrating the independent and combined effects of the Pairwise Alignment (PA) and Progressive Competence Training (PCT) modules on the FGVC Aircraft dataset.

Baseline	PA Module	PCT Module	Aircraft 1-Shot	Aircraft 5-Shot
Yes	No	No	54.2%	71.5%
Yes	Yes	No	61.7%	78.3%
Yes	No	Yes	58.4%	75.1%
Yes	Yes	Yes	68.9%	84.7%

The ablation results on the rigid structures of the Aircraft dataset are highly revealing. Introducing the Pairwise Alignment module independently provides a massive boost to classification accuracy, particularly in the 1-shot setting. This confirms that explicitly aligning parts like landing gear or specific wing configurations is critical when attempting to differentiate visually similar aircraft models from diverse camera angles [37]. Without alignment, the distance metrics are corrupted by comparing disjointed components. Integrating only the Progressive Competence Training strategy into the baseline also yields a tangible improvement. By organizing the training trajectory from distinct aircraft types to highly similar ones, the feature extractor learns a more generalized and robust representational space, avoiding premature overfitting to the initial training classes. However, the most significant discovery is the synergistic effect when both modules are combined. The performance increase of the unified framework is strictly greater than the sum of the individual improvements. This synergy occurs because the Progressive Competence Training provides the necessary optimization stability for the complex Pairwise Alignment module to converge effectively. Without the progressive curriculum, the alignment module struggles to learn accurate attention maps during the chaotic early stages of training [38]. Conversely, without the alignment module, the curriculum strategy eventually hits a performance ceiling because the fundamental feature representation remains misaligned. Further qualitative analysis of the failure cases indicates that the framework primarily struggles under conditions of extreme object occlusion or dramatic lighting failure, where the local descriptors themselves are corrupted beyond recognition. In these instances, the Pairwise Alignment module cannot find valid semantic matches, causing the confidence scores to drop significantly. Despite these boundary cases, the comprehensive results validate the structural and strategic decisions embedded within the proposed architecture.

5. Conclusion

This paper has presented a comprehensive and highly effective framework for addressing the profound challenges associated with few-shot fine-grained image recognition. By fundamentally recognizing that global feature aggregation destroys the vital localized information necessary for fine-grained distinctions, the architecture was explicitly designed to operate on dense, spatially preserved deep feature maps. The introduction of the Pairwise Alignment module represents a significant advancement in semantic correspondence. By dynamically projecting local descriptors into a shared latent space and computing cross-attention based affinity maps, the model successfully aligns discriminative object parts across query and support images without relying on expensive manual part annotations. This mechanism effectively neutralizes the spatial misalignments caused by varied poses, scales, and camera angles, ensuring that the final classification metric is grounded in genuine semantic similarity. Furthermore, the paper addressed the severe optimization instabilities inherent in training complex alignment mechanisms on sparse data distributions through the implementation of Progressive

Competence Training. By dynamically quantifying task difficulty based on the margin of separation in the evolving feature space, the curriculum learning strategy ensures that the model is exposed to classification tasks in a structured, pedagogical sequence. Beginning with highly distinguishable categories and systematically progressing to the most nuanced fine-grained tasks, this approach prevents premature convergence, stabilizes gradient flow, and significantly enhances the generalization capabilities of the underlying feature extractor. Extensive empirical evaluations conducted across multiple rigorous benchmarks, including diverse biological and structural datasets, unequivocally demonstrated the superiority of the proposed framework. The model established state-of-the-art performance in both one-shot and five-shot configurations, proving its robustness across varying levels of data scarcity. The detailed ablation studies confirmed that the Pairwise Alignment and Progressive Competence Training modules operate synergistically, with the structured learning trajectory providing the necessary stability for the complex alignment mechanisms to reach optimal convergence. Future research directions will focus on extending this pairwise alignment methodology beyond spatial dimensions to incorporate multimodal data, potentially utilizing text-based class descriptions to further guide the semantic alignment process and improve resilience against severe visual occlusions.

References

- 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).
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. In *Proceedings of ICCV*.
- Dong, J., Zhang, C., Qu, X., Ma, Z., Koniusz, P., & Ong, Y. S. (2026). Robust superalignment: Weak-to-strong robustness generalization for vision-language models. *Advances in Neural Information Processing Systems*, 38, 18345-18377.
- 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.
- Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11), 10972-10986.
- 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.
- Zhang, P., Liu, H., Ge, Z., Wang, C., & Lam, E. Y. (2024). Neuromorphic imaging with joint image deblurring and event denoising. *IEEE Transactions on Image Processing*, 33, 2318-2333.
- Yang, D., Gao, Y., Wang, X., Yue, Y., Yang, Y., & Fu, M. (2025, May). Opengs-slam: Open-set dense semantic slam with 3d gaussian splatting for object-level scene understanding. In *2025 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8486-8492). IEEE.
- Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geotypical Synthetic Labels for Building Detection. *IEEE Transactions on Geoscience and Remote Sensing*.
- Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2240-2249).

- Zhu, Guoying, et al. "Enabling MoE on the Edge via Importance-Driven Expert Scheduling." arXiv preprint arXiv:2508.18983 (2025).
- Wang J, Fan L, Li B, et al. A Dynamic Factor Gating Architecture with Market Regime Awareness for Stock Return Forecasting[J]. 2026.
- Yang, Huan, et al. "Kvshare: An llm service system with efficient and effective multi-tenant kv cache reuse." arXiv preprint arXiv:2503.16525 (2025).
- Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., ... & Wu, Z. (2024). 3d vision-language gaussian splatting. arXiv preprint arXiv:2410.07577.
- Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using llms: An automotive case study. arXiv preprint arXiv:2502.04008.
- Lv, Qi, et al. "F1: A vision-language-action model bridging understanding and generation to actions." arXiv preprint arXiv:2509.06951 (2025).
- 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).
- Wang, Siheng, et al. "DeCo-DETR: Decoupled Cognition DETR for efficient Open-Vocabulary Object Detection." arXiv preprint arXiv:2604.02753 (2026).
- Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia, 25, 1414-1426.
- Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Explainable AI for Mental Health Detection from Social Media: A Comparative Study of Traditional Machine Learning and a Large Language Model. Available at SSRN 6429778.
- Shuang, K., Guo, J., & Wang, Z. (2022). Comprehensive-perception dynamic reasoning for visual question answering. Pattern Recognition, 131, 108878.
- Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).
- Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024, March). Bias-conflict sample synthesis and adversarial removal debias strategy for temporal sentence grounding in video. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 5, pp. 4533-4541).
- 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).
- Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Compliance-Aware Discharge Agent for Auditable ICU Discharge Planning: A Pilot Feasibility Study Using Structured eICU Records. Available at SSRN 6429758.
- Zhang, S., Yang, S., Zhang, W., Xiong, Y., & Yao, S. (2026). Hybrid Beamforming for Subarray-Level Movable Antenna Enhanced MU-MIMO Communications. IEEE Wireless Communications Letters, 15, 2559-2563.
- Zhang, W., Zhang, C., Gu, C., Kou, J., Yuan, H., Fang, X., ... & Fang, Y. (2024, October). Hallucination in Large Language Models: From Mechanistic Understanding to Novel Control Frameworks. In 2024 7th International Conference on Universal Village (UV) (pp. 1-36). IEEE.

- Huang, H., Zhang, J., Yu, L., Zhang, J., Wu, Q., & Xu, C. (2021). TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2), 853-866.
- 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.
- Xu, W., Huang, H., Cheng, M., Yu, L., Wu, Q., & Zhang, J. (2023, July). Masked cross-image encoding for few-shot segmentation. In *2023 IEEE international conference on multimedia and expo (ICME)* (pp. 744-749). IEEE.
- Dong, J., Koniusz, P., Zhang, Y., Zhu, H., Liu, W., Qu, X., & Ong, Y. S. (2025, October). Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of CVPR*.
- Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.
- Zhang, W. (2026). A 5-6 GHz PVT Robust Current Mode Passive Mixer for Direct Down-Conversion Receiver.
- Wang, Siheng, et al. "C3-owd: A curriculum cross-modal contrastive learning framework for open-world detection." *arXiv preprint arXiv:2509.23316* (2025).
- Zhang, P., Zhu, S., Wang, C., Zhao, Y., & Lam, E. Y. (2024). Neuromorphic imaging with super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2), 1715-1727.
- Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. *International Journal of Remote Sensing*, 46(5), 1913-1937.
- Wang, W., Mi, L., Cen, S., Dai, H., Li, Y., Fu, X., & Liu, Y. (2025). Region-based Content Enhancement for {Efficient} Video Analytics at the Edge. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)* (pp. 613-633).