Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.20265055

Evolution of Few-Shot Image Classification: A Survey from Metric Learning to Large Model Fine-Tuning

Yuhang Ji*

School of Future Technology, South China University of Technology, Guangdong Province, 511400, China

**Corresponding author: Yuhang Ji.*

Abstract

While deep learning has achieved remarkable success in image recognition, it relies heavily on massive amounts of labeled data. In real-world scenarios like medical imaging, collecting such large datasets is often expensive or impossible. To overcome this critical bottleneck, Few-Shot Image Classification (FSIC) has emerged as a vital domain, enabling models to learn new categories from extremely limited samples. Based on a comprehensive review of recent literature, this paper systematically divides mainstream FSIC methods into three categories: Metric Learning, Optimization/Meta-Learning, and Transfer Learning with Large Model Fine-Tuning. This survey comprehensively evaluates the underlying mechanisms, inherent strengths, and specific weaknesses of each category. Additionally, the paper provides a clear performance comparison to illustrate the evolutionary impact of these different approaches. Finally, the paper deeply discusses remaining technical challenges, specifically detailing issues like cross-domain generalization and computational dependence, and outlines promising future research directions, such as multimodal knowledge fusion, to further advance the field.

Keywords

FSIC, metric learning, meta-learning, parameter-efficient fine-tuning, cross-domain generalization

1. Introduction

The human visual system is remarkably efficient, as humans have the ability to perform accurate classification after seeing just a single example of a new category. However, conventional computer models need thousands of images to learn the same concept. This heavy reliance on big data limits the application of AI in fields like medicine or military surveillance. FSIC was proposed to bridge this gap by enabling classifiers to generalize to new classes given only a small number of examples.

The main challenge of Few-Shot Image Classification (FSIC) is overfitting. With very few training samples, deep learning models tend to "memorize" specific details of those images instead of learning the general features of the category. To overcome this, recent literature has explored a variety of mechanisms. For instance, Prototypical Networks relieve overfitting by learning a metric space where classification is efficiently performed by computing distances to class prototype representations [1]. In the field of meta-learning, frameworks like Model-Agnostic Meta-Learning (MAML) optimize a model's initial parameters so that only

a few gradient steps are needed for rapid adaptation to complete new tasks [2]. More recently, the emergence of large vision-language models, such as CLIP, has shifted the field toward parameter-efficient fine-tuning [3]. Approaches such as Context Optimization (CoOp) automate prompt engineering by optimizing continuous context vectors [4]. Building on this, Multi-modal Prompt Learning (MaPLe) jointly combines vision and language prompts to ensure mutual collaboration between modalities, while Tip-Adapter utilizes a non-parametric key-value cache model to achieve highly efficient and training-free adaptation [5,6].

This paper provides a systematic survey of FSIC techniques. These methods are categorized into three cohesive groups based on how they emulate human problem-solving. The first group is metric learning, which revolves around finding similarities to determine whether a newly observed object is close to previously established category. The second group consists of meta-learning techniques, which concentrate on learning how to learn by mastering the problem-solving methodology for rapid adaptation. The final group explores transfer learning and large model fine-tuning, which applies foundational models that already possess vast world knowledge to act as a comprehensive reference for downstream tasks.

2. Overview of Related Technologies

2.1 Metric Learning-based Methods

The fundamental principle of metric learning-based methods revolves around mapping images into a high-dimensional embedding space where images of the same class are clustered closely together, while different classes are pushed far apart.

A seminal work in this area is Prototypical Networks, which calculates the average feature vector of the limited samples within a class to establish a prototype [1]. Classification for novel images is subsequently performed by assigning the image to the nearest class prototype based on a distance metric. Recent progress in this field includes PrototypeFormer, which leverages Transformer architectures to direct the model's attention toward the primary object rather than background elements [7]. Similarly, SS-FSL integrates text labels to provide semantic guidance to the model when visual information alone is insufficient [8].

The primary advantage of the metric learning approach lies in its simplicity and computational efficiency, making it intuitive and easy to implement. Nevertheless, its core limitation is that it often struggles with fine-grained classification tasks where different categories exhibit high visual similarity, such as distinguishing between specific animal breeds. In such scenarios, metric learning often fails to capture the subtle, discriminative features required for accurate class separation.

2.2 Optimization and Meta-learning Methods

Optimization and meta-learning methods focus on the paradigm of "learning to learn". Rather than optimizing for peak performance on the current dataset, these models aim to find an optimal set of initialization parameters that facilitate rapid adaptation to novel tasks.

Model-Agnostic Meta-Learning (MAML) is a foundational framework in this field; it trains a versatile starting point that requires only a few gradient updates to achieve high accuracy when encountering new categories [2]. More recent advancements include GAP, which accelerates this learning process dynamically based on task difficulty [9]. Additionally, UNEM introduces an effective approach that leverages both labeled support images and unlabeled query batches jointly to enhance prediction accuracy [10].

These methods boast strong adaptation capabilities and high flexibility across varying tasks. Conversely, their primary disadvantage is the initial meta-training phase, which is notoriously slow and computationally exhaustive. The requirement to compute complex bi-level optimizations and unroll computation graphs often leads to high memory consumption and challenging training dynamics.

2.3 Transfer Learning & Large Model Fine-tuning

The advent of large vision-language models, such as CLIP, has elevated Few-Shot Image Classification to a new level through transfer learning and parameter-efficient fine-tuning [3]. Because these foundational models are pre-trained on millions of internet images, they eliminate the need to train classifiers from scratch.

Context Optimization (CoOp) is a representative approach that utilizes learnable continuous prompt vectors to guide the frozen base model, effectively mitigating the risk of overfitting [4]. Extending this concept, MaPLe applies multi-modal prompt tuning to both the vision and language branches simultaneously to achieve a deeper alignment of features [5]. Taking a different approach to parameter efficiency, Task Residual Tuning (TaskRes) diverges from prompt manipulation by introducing a set of learnable, prior-independent parameters as an additive residual directly to the frozen text-based classifier, explicitly decoupling pre-trained prior knowledge from new, task-specific features [11]. Meanwhile, Tip-Adapter achieves competitive classification results without requiring gradient-based parameter updates, relying instead on a highly efficient key-value cache system constructed from few-shot supervisions [6].

The distinct advantage of these fine-tuning methods is their capacity to achieve state-of-the-art classification accuracy on diverse benchmarks by exploiting vast prior knowledge. However, their significant drawbacks include a heavy reliance on the existence of these expensive pre-trained models and the substantial hardware required to load them. Furthermore, if the downstream application involves highly specialized domains—such as microscopic medical imaging—that are absent from the general pre-training corpus, these models may still struggle to adapt effectively.

3. Analysis of Experimental Results

3.1 Evaluation Rationale and Dataset Selection

To ensure a logically rigorous comparison between traditional few-shot algorithms and foundation-model-based approaches, the paper adopts a decoupled evaluation strategy. Historically, datasets like *mini*ImageNet served as universal benchmarks. However, directly comparing large-scale pre-trained models (e.g., CLIP) against traditional models trained from scratch on these common datasets presents a fundamental flaw. Because foundation models undergo massive open-world pre-training, evaluating them on familiar distributions is akin to an "open-book exam"—their high performance may merely reflect pre-training memory rather than genuine few-shot adaptation. Therefore, the evaluation is divided into two tracks: *mini*ImageNet is utilized to illustrate the historical evolution of few-shot methods and highlight the capacity gap between different backbones (Table 1). Subsequently, to truly assess the few-shot generalization capabilities of large models, the paper shifts to robustness evaluations under severe distribution shifts, using a standardized backbone to fairly compare different adaptation mechanisms (Table 2).

3.2 Methodological Evolution on *mini*ImageNet

This paper analyzes the historical progression of few-shot learning on the *mini*ImageNet dataset. As shown in Table 1, early meta-learning and metric-learning approaches utilizing Conv-4 backbones (e.g., Prototypical Nets, MAML) established baseline 1-shot accuracies of 60.76% and 48.70%. Across all methods, providing 5 shots consistently yields a 17% to 10% improvement over 1-shot scenarios, proving that additional reference images effectively reduce model uncertainty.

Crucially, the table highlights the massive performance leap introduced by Vision Transformer (ViT) architectures and large-scale pre-training. Methods like PrototypeFormer (ViT-Small) and CLIP (ViT-Large/14) achieve staggering 1-shot accuracies of 90.88% and 83.86%. This disparity structurally validates the paper's core premise: the dominance of foundation models on standard benchmarks largely stems from their powerful backbones. Consequently, a more rigorous robustness evaluation is required to measure their actual adaptation capabilities.

Table 1: Performance Comparison on miniImageNet.

Method	Category	Backbone	5-way 1-shot (%)	5-way 5-shot (%)
Prototypical Nets	Metric Learning	Conv-4	60.76 ± 0.39	78.44 ± 0.21
PrototypeFormer	Metric Learning	ViT-Small	90.88 ± 0.31	97.07 ± 0.11
MAML	Meta-learning	Conv-4	48.70 ± 1.84	63.11 ± 0.92
GAP	Meta-learning	Conv-4	54.86 ± 0.85	71.55 ± 0.61
CLIP	Transfer & Fine-tuning	ViT-Large/14	83.86 ± 0.40	95.13 ± 0.14

3.3 Robustness and Domain Generalization of Foundation Models

To accurately assess few-shot generalization without the unfair advantage of pre-training familiarity, the paper evaluates model robustness against target domain distribution shifts (Target: -V2, -S, -A, -R). To ensure fairness, a uniform ViT-B/16 backbone is applied across all methods in Table 2, isolating the effectiveness of the adaptation strategies themselves.

The results reveal a stark contrast in domain generalization. While a basic Linear Probe on CLIP achieves 65.85% under the 16-shot setting, it suffers a catastrophic degradation on challenging shifted domains. Conversely, advanced prompt-learning and adaptation methods successfully mitigate this collapse. TaskRes achieves the highest source accuracy of 73.07%, and alongside MaPLe, maintains significantly higher robustness across all target domains, both retaining over 49% on Target: -A and over 76% on Target: -R. By testing under strict distribution shifts, the results prove that advanced residual adaptation and prompt optimization are essential to unlock the true out-of-distribution generalization potential of foundation models.

Table 2: Robustness Comparison with ViT-B/16 backbone (M: CoOp's context length)

Method	16-shot acc. (%)	Target: -V2	Target: -S	Target: -A	Target: -R
Zero-shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
MaPLe	70.72	64.07	49.15	50.90	76.98
CLIP+CoOp(M=16)	70.72	64.18	46.71	48.41	74.32
Co-CoOp	71.02	64.07	48.75	50.63	76.18
TaskRes	73.07	65.30	49.13	50.37	77.70

4. Challenges and Future Research Directions

Despite the remarkable progress evidenced by recent benchmarks, several critical challenges persist that define the frontier of future research in few-shot learning.

First, the Cross-Domain Generalization Problem remains a significant hurdle. Most current state-of-the-art models are evaluated under the assumption that training and testing data share the same underlying distribution. However, performance often collapses when there is a significant "domain shift"—for instance, when a model trained on high-resolution natural photographs is tasked with identifying pencil sketches, satellite imagery, or artistic renderings. Future research must move beyond simple feature alignment and focus on learning domain-invariant representations that can generalize to unseen environments without requiring extensive fine-tuning.

Second, there is a growing concern regarding the Heavy Dependence on Large-Scale Pre-trained Models. While "giants" like CLIP have propelled accuracy scores to new heights, their success is largely tethered to the massive datasets they encountered during pre-training. In specialized or highly private sectors—such as microscopic medical imaging, rare industrial defect detection, or specialized satellite telemetry—these general-purpose models often lack the necessary prior knowledge. This necessitates the development of more "data-efficient" architectures that can achieve high performance from scratch or through the adaptation of smaller, more specialized foundation models.

Finally, the integration of Multimodal Knowledge Fusion represents a vital evolutionary step. Human recognition is rarely based on visual pixels alone; instead, it draws upon a vast reservoir of "common sense" and semantic context to identify objects. Future AI systems should transcend pure visual matching by incorporating external knowledge graphs, hierarchical taxonomies, or linguistic descriptions into the learning loop. By bridging the gap between low-level visual features and high-level symbolic reasoning, models can achieve a more robust, human-like understanding of novel categories, even when visual information is ambiguous or scarce.

5. Conclusion

This paper provides a comprehensive review of the evolution of Few-Shot Image Classification (FSIC) methods from 2017 to 2025, addressing the critical bottleneck of data scarcity in deep learning. The paper systematically categorizes the mainstream approaches into Metric Learning, Optimization/Meta-learning, and Transfer Learning with Large Model Fine-tuning. While classification accuracy on benchmarks like

*mini*ImageNet has reached unprecedented peaks—largely driven by the integration of robust, large-scale multimodal foundation models such as CLIP—significant hurdles remain. Persistent challenges, including cross-domain generalization, severe domain shifts, and high computational dependencies, limit broader real-world application. Consequently, future research must pivot toward developing lighter, data-efficient model architectures, advancing multimodal knowledge fusion, and integrating smarter, human-like common-sense reasoning to achieve truly robust artificial intelligence.

References

[1] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, *30*.

[2] Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126-1135). PMLR.

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

[4] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International journal of computer vision*, *130*(9), 2337-2348.

[5] Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023). Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19113-19122).

[6] Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., ... & Li, H. (2021). Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

[7] Su, M., He, F., Li, G., & Li, F. (2025). PrototypeFormer: Learning to explore prototype relationships for few-shot image classification. *Neurocomputing*, *640*, 130326.

[8] Liu, T., Basu, A., Caverlee, J., & Kong, S. (2025). Solving Semi-Supervised Few-Shot Learning from an Auto-Annotation Perspective. arXiv preprint arXiv:2512.10244.

[9] Kang, S., Hwang, D., Eo, M., Kim, T., & Rhee, W. (2023). Meta-learning with a geometry-adaptive preconditioner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16080-16090).

[10] Zhou, L., Shakeri, F., Sadraoui, A., Kaaniche, M., Pesquet, J. C., & Ben Ayed, I. (2025). UNEM: UNrolled generalized EM for transductive few-shot learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 9665-9675).

[11] Yu, T., Lu, Z., Jin, X., Chen, Z., & Wang, X. (2023). Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10899-10909).

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.