Computers and Artificial Intelligence; Vol.3, No.6; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n6.211227
Published by: Zeus Press

A Survey on Few-Shot Image Classification

Yangyurui Gao*

School of Computer and Cyber Sciences, Communication University of China, Beijing 100024, China

**Corresponding author: Yangyurui Gao.*

Abstract

Few-shot image classification aims to enable a model to rapidly recognize and generalize to new classes using only a very small number of labeled samples, and it is one of the key capabilities required for applying deep learning in open-world scenarios. Starting from the mathematical nature of the problem and a formal framework for knowledge transfer, this paper systematically reviews the research landscape and methodological taxonomy of this field. According to their core operating mechanisms, existing methods can be grouped into five major paradigms. Metric-learning-based strategies encode knowledge as distance-measurement rules in an embedding space; optimization-based meta-learning strategies encode it as parameter initializations that can rapidly adapt to new tasks; data-augmentation-based strategies encode it as reasonable extrapolations of within-class distributions; pre-training and fine-tuning paradigms encode it as general transferable feature representations; and parameter-generation and probabilistic-inference methods encode it as conditional distributions of classifier parameters or posterior inference processes. This paper describes the basic principles, representative studies, inherent limitations, and forms of cross-paradigm integration of these methods, systematically summarizes their evaluation systems and performance comparisons, and discusses key challenges and future directions, including cross-domain generalization, paradigm restructuring in the era of foundation models, and extensions to complex visual tasks. The aim is to provide a structured cognitive framework and research reference for this field.

Keywords

few-shot learning, image classification, metric learning, meta-learning, data augmentation, pre-training and fine-tuning, transductive inference

1. Introduction

Deep neural networks, especially convolutional neural networks [1], have become dominant across computer vision, but their performance depends heavily on large-scale labeled data. ImageNet [2], for example, contains more than 14 million images, whose annotation required enormous human and time costs. Annotation is even more expensive in specialized fields: medical images must be interpreted individually by specialist physicians [3], samples of new defect types are difficult to collect in advance for industrial defect detection [4], and new land-cover classes continue to emerge in remote-sensing image interpretation [5]. This rigid demand for massive labeled datasets has become a prominent bottleneck in deep-learning applications: when annotation is extremely costly or samples are inherently scarce, standard deep models often suffer sharp performance degradation because of insufficient class coverage.

© 2026 The Author(s). Authors retain copyright in their articles.

Under these conditions, a model must be able to rapidly adapt and generalize to new classes from only a few labeled samples. This machine-learning paradigm is known as few-shot learning (FSL). According to the training approach, FSL methods can be broadly divided into two categories. The first is the episodic training paradigm based on meta-learning, in which training and testing use episodes rather than conventional batches as their basic units. The model is trained on numerous meta-tasks composed of a support set and a query set to simulate the test condition of learning new classes from a few samples. The second is the pre-training and fine-tuning paradigm, in which the model is first trained on large-scale base-class data through supervised or self-supervised learning and is then adapted to few-shot novel classes by fine-tuning or nearest-neighbor classification. This review covers both paradigms and compares their methodological taxonomies and performance characteristics in the following sections.

Taking the episodic meta-learning paradigm as an example, a standard N-way K-shot meta-task is defined as follows. Let the set of all classes be the label space Y. N mutually exclusive classes are randomly sampled from Y to form the class subset $C = \{c_1, c_2, \ldots, c_N\} \subseteq Y$ for the current task. Each class contains K labeled samples in the support set and M samples to be predicted in the query set.

Support set S:

$$S = \{(x_i, y_i) \mid i = 1, 2, \ldots, N{\times}K\}$$

Each class appears exactly K times, and the label $y_i \in C$.

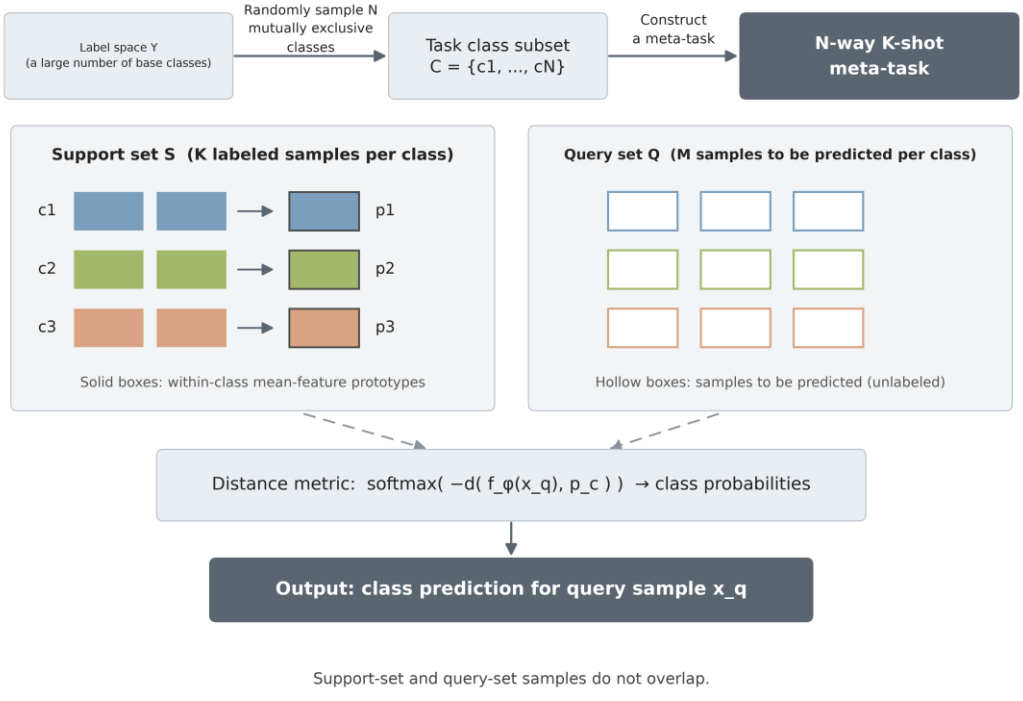
Query set Q:

$$Q = \{(x_j, y_j) \mid j = 1, 2, \ldots, N{\times}M\}$$

Each class appears exactly M times, and the label $y_j \in C$.

The samples in the support and query sets do not overlap, that is, $S \cap Q = \emptyset$. On the basis of the limited labeled information in support set S, the model must correctly predict the classes of all samples in query set Q.

Figure 1: Illustration of an N-way K-shot meta-task



The core value of few-shot learning lies in addressing scenarios in which annotation is expensive or samples are inherently scarce, so that a model can still recognize and generalize effectively to new classes after observing only a very small number of samples.

Literature search and screening method: This paper used “few-shot learning,” “few-shot image classification,” “meta-learning,” “metric learning,” “prompt learning,” and their corresponding Chinese terms

© 2026 The Author(s). Authors retain copyright in their articles.

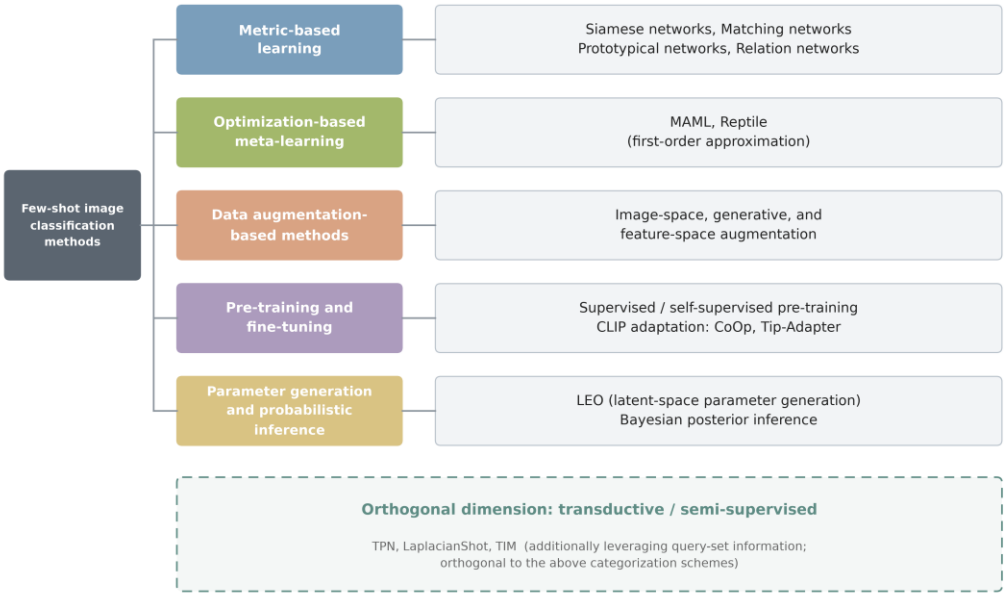
as keywords to search Web of Science, IEEE Xplore, SpringerLink, China National Knowledge Infrastructure (CNKI), and other databases. Particular attention was given to papers from CVPR, ICCV, ECCV, NeurIPS, and ICLR, and the search covered publications up to 2025. Studies were included if they (1) proposed a method for few-shot image classification or provided a systematic review of the task; (2) were evaluated on public benchmarks such as miniImageNet, tieredImageNet, CUB, or Meta-Dataset; and (3) offered a clear mechanistic innovation or analytical value. Under these criteria, both foundational classics and representative recent advances were considered, and the final literature set was determined by combining citation impact and experimental reproducibility.

2. Classification of Few-Shot Learning Strategies

2.1 Classification Principles

Few-shot learning includes many methods with overlapping design ideas. This paper establishes a classification framework according to the core operating mechanism of each method. The following categories are not mutually exclusive: high-performing methods often combine multiple mechanisms. Each method is classified according to the dimension of its most distinctive contribution, and its integration with other strategies is discussed at the end of each subsection.

Figure 2: Taxonomy of few-shot image classification methods



2.2 Metric-Learning-Based Strategies

To address the dual challenge of inter-class distribution shift and insufficient within-class samples, one line of research begins with the "metric." Its core assumption is that a transferable embedding space exists in which the basic concept of "similarity" can be shared between base and novel classes [6, 7]. Instead of directly learning a classifier, these methods learn how to compare samples. Representative studies include Siamese networks [8], matching networks [9], prototypical networks [10], and relation networks [11], which characterize similarity through pairwise sample comparison, attention weighting, prototype distance, and learned relation scores, respectively. Their principles and development are described in Section 3.1.

2.3 Optimization-Based Meta-Learning Strategies

In contrast to the metric-learning idea of "learning a good space," another line of work shifts the focus from feature space to the learning process itself, thereby forming optimization-based meta-learning strategies. Their core concept is "learning to learn": through meta-training on many base-class tasks, the model obtains a

© 2026 The Author(s). Authors retain copyright in their articles.

parameter initialization that is highly sensitive to task changes, allowing it to adapt rapidly to a new task with only a few samples and only a small number of gradient-update steps. Model-Agnostic Meta-Learning (MAML) [12] and Reptile [13] are representative methods. The significance of their bilevel optimization mechanism and first-order approximation is explained in Section 3.2.

2.4 Data-Augmentation-Based Strategies

The third technical route addresses the most direct source of the problem: because sample scarcity makes within-class distributions difficult to estimate reliably, directly increasing the number of samples through generation or augmentation can reduce the difficulty of fitting novel classes. These methods can be broadly divided into image-space augmentation [14], sample generation based on generative models [15, 16], and feature-space augmentation [17, 18]. Section 3.3 discusses them and analyzes their risks in detail.

2.5 Pre-Training and Fine-Tuning Paradigm

The preceding strategies all use episodic meta-learning as their core training framework. Through systematic comparative experiments in 2019, Chen et al. [19] showed that a simple "pre-training + nearest-neighbor/linear classification" baseline could perform comparably to most meta-learning methods. Tian et al. [20] further demonstrated that the quality of the feature extractor is far more important than the choice of classifier. Consequently, "high-quality feature representation" became an independent research route parallel to meta-learning; it is discussed systematically in Section 4.

2.6 Parameter-Generation and Probabilistic-Inference Methods

In addition to the four categories above, parameter-generation and probabilistic-inference methods form a relatively independent branch. One approach generates classifier parameters in a low-dimensional latent space; for example, Latent Embedding Optimization (LEO) [21] uses latent-space optimization to generate adaptive classifier weights for new tasks. The other treats few-shot learning as a posterior-inference problem and constrains the hypothesis space for novel classes by introducing prior distributions. Representative methods include probabilistic MAML [22] and VERSA [23]. Although these two branches contain relatively few methods, they provide perspectives on knowledge encoding and uncertainty modeling in few-shot learning that differ from deterministic approaches.

2.7 Transductive and Semi-Supervised Few-Shot Learning

The methods discussed above use inductive inference, in which each query sample is classified independently. Transductive methods, in contrast, allow query-set samples to refer to one another during classification. This is particularly reasonable in a 1-shot setting: when the support set contains only one sample, the distributional information in the query set itself becomes an additional supervisory signal. The Transductive Propagation Network (TPN) [24] constructs a graph from the support and query sets and propagates labels through a graph neural network; LaplacianShot [25] formulates transductive inference as semi-supervised clustering over the query set; and Transductive Information Maximization (TIM) [26] maximizes mutual information over query-set features from an information-theoretic perspective. It should be emphasized that transduction is an inference setting rather than a form of knowledge encoding, and it should therefore be treated as a dimension orthogonal to the five major paradigms. Because query-set information is used additionally, transductive methods usually achieve accuracies several percentage points higher than inductive methods; results from the two settings cannot be directly compared in the same ranking.

2.8 Horizontal Comparison and Internal Connections Among Existing Methods

The strategies above are not isolated from one another. They share the same deeper objective: effective knowledge transfer under the dual constraints of inter-class distribution shift and insufficient within-class samples. Their differences essentially concern where prior knowledge is introduced and how that knowledge is represented. Metric learning encodes distance-measurement rules; optimization-based meta-learning encodes sensitive initialization parameters; data augmentation encodes extrapolation of within-class distributions; pre-training and fine-tuning encode general feature representations; and parameter-generation

© 2026 The Author(s). Authors retain copyright in their articles.

and probabilistic methods encode conditional distributions or posterior distributions of classifier parameters. In practice, high-performing methods often combine these strategies.

3. Detailed Description of Few-Shot Classification Methods

3.1 Metric-Learning-Based Strategies

3.1.1 Core Principles

The basic idea of metric learning for few-shot classification is to abandon the learning of an explicit decision boundary for each novel class and instead learn an embedding space in which the relative distance between samples directly reflects semantic similarity [6, 7]. Because distance computation does not depend on specific classes, the embedding function can be reused directly without any adjustment when novel classes appear. Mathematically, given an embedding function fφ, a metric-learning method defines a distance function d(·, ·), and predicts the class of a query sample xq according to:

$$p(y = c|x_q) = \text{softmax}(-d(f_\varphi(x_q), p_c) \tag{1}$$

Where p_c is the reference representation of class c in the embedding space. The methods differ mainly in two respects: how the reference representation p_c is constructed and which distance function d is selected.

3.1.2 Siamese Networks

The Siamese network proposed by Koch et al. [8] was a pioneering application of metric learning to few-shot classification. The network consists of two convolutional encoders with shared weights. They receive a pair of input samples and map them to an embedding space. After the L1 distance between the feature vectors is calculated, a fully connected layer and a sigmoid function output the probability that the two samples belong to the same class. Its core innovation is to transform classification into similarity judgment: instead of learning "which class an image belongs to," it learns "whether two images belong to the same class." Its limitations are also clear. N-way classification requires pairwise construction of $O(N^2)$ sample pairs, which is computationally inefficient, and a distribution mismatch exists between pairwise training and multi-class comparison at test time.

3.1.3 Matching Networks

The matching network proposed by Vinyals et al. [9] made two improvements. First, it introduced a bidirectional long short-term memory (LSTM) network to encode support-set samples contextually, so that the feature of each support sample incorporates information from the other samples in the same support set. Second, it replaced fixed pairwise sample comparison with an attention mechanism. For a query sample f(xq), an attention kernel is used to compute a weighted sum of the features of all support-set samples:

$$p(y|x_q) = \sum_{(x_i,y_i)\in S} a(f(x_q), f(x_i)) \cdot 1_{y_i=y} \tag{2}$$

Where the attention weight a(., .) generally takes the form of cosine similarity followed by softmax. The class prediction for a query sample is essentially a weighted vote over the labels of all support-set samples, with the weights determined by similarity in feature space.

3.1.4 Prototypical Networks

The prototypical network proposed by Snell et al. [10] reduces the preceding idea to a very simple form. It assumes that each class has a "prototype" point in the embedding space and that within-class samples are distributed symmetrically around that prototype. The class prototype can be estimated from the mean feature vector of the support-set samples:

$$p_c = \frac{1}{K} \cdot \sum_{(x_i,y_i)\in S, y_i=c} f(x_i) \tag{3}$$

During classification, the Euclidean distance between a query sample and each class prototype is calculated and converted into class probabilities through softmax. From the perspective of mixture-density estimation, a prototypical network is equivalent to assuming that all classes follow Gaussian distributions with the same covariance matrix. This strong inductive bias provides regularization when data are extremely scarce; however,

© 2026 The Author(s). Authors retain copyright in their articles.

the single-prototype assumption is difficult to maintain when the within-class distribution is clearly multimodal or highly dispersed.

3.1.5 Relation Networks

The relation network proposed by Sung et al. [11] attempts to address the fundamental question of whether a predefined distance metric is optimal. A relation network consists of an embedding module and a relation module. The embedding module maps support and query samples to feature maps; the relation module concatenates the two feature maps along the channel dimension and feeds them into a convolutional neural network that directly outputs a relation score between 0 and 1. The deeper significance of relation networks is that they blur the boundary between "feature learning" and "metric learning": distance computation is fully integrated into end-to-end learning.

3.1.6 Set Adaptation and Structured Metrics

Following prototypical, matching, and relation networks, the focus of metric-learning research shifted from "designing a single prototype" to "adaptively adjusting embedding representations and matching structures." Few-Shot Embedding Adaptation with Transformer (FEAT) [27] uses a set-to-set function to contextually adapt support-set embeddings, allowing samples of the same class to correct one another's representations. DeepEMD [28] introduces a differentiable earth mover's distance and replaces global-vector comparison with structured matching between sets of local features, substantially improving performance in fine-grained and cross-domain settings. Adaptive Parametric Prototype Learning (APPL) [29] adaptively generates class prototypes from joint support-set features in cross-domain settings, reducing the distortion caused by fixed mean prototypes. These studies show that, when samples are extremely scarce, fully exploiting the internal structure of the support set is as important as the quality of the embedding space itself.

3.2 Optimization-Based Meta-Learning Strategies

3.2.1 Core Principles

Metric learning answers the question of "where to compare," whereas optimization-based meta-learning raises a more fundamental question: how can a model acquire the ability to "learn rapidly"? Its core idea is "learning to learn." Through meta-training on many different tasks, the model does not learn parameters that solve one specific task; instead, it obtains a set of initial parameters that are highly sensitive to task changes, allowing it to achieve good generalization on a new task after only a small number of gradient updates.

3.2.2 MAML

The model-agnostic meta-learning method proposed by Finn et al. is a foundational work in this paradigm. MAML [12] introduces no additional meta-parameters and makes no changes to the model architecture. It learns an initialization point that is "easy to fine-tune" solely through second-order optimization over the task distribution. MAML training has an inner loop and an outer loop. The inner loop simulates few-shot adaptation. For each sampled meta-task T_i, its support set S_i is used to perform a small number of gradient-descent steps from the current initial parameters θ:

$$\theta_i' = \theta - \alpha \cdot \nabla_\theta L_{S_i}(\theta) \tag{4}$$

The outer loop then optimizes the initial parameters θ so that the model, after rapid adaptation in the inner loop, has a low loss on the query set Q_i of the corresponding task:

$$\min_\theta \sum_{T_i \sim p(T)} L_{Q_i}(\theta_i') \tag{5}$$

The outer-loop optimization must compute gradients through the inner-loop gradient, so MAML involves second-order derivatives. In this way, MAML learns a parameter initialization that is "sensitive to task changes." These parameters may not be an excellent solution for any individual task, but they can rapidly become one. Model agnosticism is another important advantage.

3.2.3 First-Order Approximation Methods

MAML has sound theoretical properties as a second-order optimization method, but its computational cost is high. First-order approximation methods such as Reptile [13] reduce complexity while retaining the core

© 2026 The Author(s). Authors retain copyright in their articles.

idea. Reptile performs multiple steps of stochastic gradient descent (SGD) for each task and uses the average direction of the parameter updates across tasks as the outer-loop update direction, completely omitting second-order gradients. Geometrically, it seeks a point near the "center" of the task-specific optima. The similar performance of Reptile and MAML on several benchmarks indicates that first-order approximations can already capture the essential elements of rapid adaptation for many standard tasks; however, second-order information may make a greater contribution when task distributions are more complex. In recent years, this line of research has shifted toward the robustness of the meta-training process itself. For example, Wang et al. [30] used expected tail risk as the meta-training objective to improve the reliability of rapid adaptation to new tasks.

3.3 Data-Augmentation-Based Strategies

3.3.1 Core Principles and Motivation

The preceding two categories attempt to address sample scarcity through the design of learning mechanisms or metric spaces. Data augmentation takes a different approach: if the difficulty arises from an insufficient number of samples, directly increasing that number can transform the few-shot problem into a conventional supervised classification problem. The intuitive nature of this idea makes it highly attractive in practice, but its effectiveness depends strictly on the quality of the augmented samples.

3.3.2 Image-Space Augmentation and Pixel-Level Transformations

The most basic augmentation methods apply geometric and color transformations in the original image space, including random cropping, horizontal flipping, rotation, color jittering, and brightness and contrast adjustment [14]. These operations are simple and computationally inexpensive. Their limitations are nevertheless clear in few-shot settings: because the transformations do not change the semantic content of the images, augmentation without semantic variation cannot effectively expand the within-class distribution when only 1-5 samples are available, and may even aggravate overfitting by repeatedly providing similar training signals.

3.3.3 Sample Generation Based on Generative Models

With the development of generative adversarial networks and variational autoencoders, researchers began to explore the direct generation of training samples for novel classes [15]. A conditional generative model is trained on base-class data and then conditioned on the few available samples of a novel class to generate variants of that class. The central difficulty is that the generative model itself requires large amounts of data to learn a stable and diverse generative distribution, whereas the few-shot setting cannot satisfy this prerequisite. Moreover, distribution shift between generated and real samples may introduce additional bias. More recently, large-scale pre-trained diffusion models have offered a new route. Domain Gap Embeddings (DoGE) [31] extracts a vector describing the distributional difference between the source and target domains and uses it to guide a pre-trained text-to-image model to generate synthetic samples aligned with the target distribution, effectively mitigating distribution shift in cross-domain few-shot settings.

3.3.4 Feature-Space Augmentation

To avoid the dependence of generative models on large training datasets, another line of work moves augmentation from the high-dimensional pixel space to a low-dimensional feature space, where the dimensionality is lower, the distribution is more compact, and augmentation is more stable. The "shrinking and hallucinating" method proposed by Hariharan and Girshick [16] first trains a feature extractor on base-class data to reduce within-class variance and then samples virtual features for novel classes from the estimated within-class distribution. Wang et al. [17] further introduced a generator to synthesize such "imaginary data." The distribution-calibration method of Yang et al. [18] uses base-class statistics to estimate the mean and covariance of a novel class and samples from the resulting distribution; it is a representative feature-space augmentation method. Typical operations also include adding Gaussian noise to feature vectors and linearly interpolating between features of samples from the same class. However, when the augmentation magnitude is excessive, features may deviate from the true manifold. If the difference between augmented samples and the real distribution is not controlled, it may mislead the classification decision boundary.

© 2026 The Author(s). Authors retain copyright in their articles.

3.3.5 oordination Between Augmentation and Learning Mechanisms

Few current state-of-the-art methods rely on data augmentation alone. Instead, they embed feature augmentation into metric-learning or meta-learning frameworks to create synergy. For example, recomputing prototypes from an augmented support set can provide more robust estimates of class centers, while augmented data can also improve the stability of gradient updates in the MAML inner loop. This coordinated design of "data expansion + mechanism constraints" is the most active research frontier for data-augmentation strategies.

4. Pre-Training and Fine-Tuning Paradigm

4.1 Emergence and Motivation of the Paradigm

In 2019, Chen et al. [19] found through rigorously controlled experiments that a feature extractor sufficiently pre-trained on base-class data and then used directly with nearest-neighbor classification or a linear classifier on few-shot novel classes could perform comparably to, or even better than, carefully designed meta-learning methods on several standard benchmarks. This finding prompted the field to reconsider a fundamental question: how much of the advantage of earlier meta-learning methods over simple baselines came from the learning mechanism itself, and how much came from comparison bias caused by inconsistent experimental settings? This question marked a major shift in few-shot learning from "algorithm-design driven" to "representation-quality driven."

4.2 Basic Principles and Workflow

The workflow of this paradigm has two stages. In the large-scale pre-training stage, the model is trained for fully supervised classification on a base-class dataset using the standard cross-entropy loss, or learns general visual representations through self-supervised methods. In the few-shot adaptation stage, all or most layers of the pre-trained feature extractor are frozen, and only a lightweight classification head is trained on top of them; alternatively, training-free nearest-neighbor classification is performed directly. The substantial reduction in trainable parameters effectively controls the risk of overfitting.

4.3 Central Role of Feature Quality

Tian et al. [20] further decomposed the few-shot classification process into two independent modules, "feature extractor learning" and "classifier learning," and systematically compared combinations of multiple feature-learning strategies and multiple classifiers. The results showed that feature-extractor quality has a much greater influence on few-shot performance than classifier choice. In a sufficiently good embedding space, simple nearest-neighbor classification can outperform complex meta-learning algorithms.

The P>M>F method proposed by Hu et al. [32] provides more direct evidence for this conclusion. After introducing large-scale external data during pre-training and fully fine-tuning the feature extractor, a simple linear classifier alone outperformed many meta-learning methods on benchmarks such as Meta-Dataset. Together with the continued improvements of DeepEMD [28] and FEAT [27], this result shows that backbone architecture and the scale of pre-training data have become key determinants of few-shot performance. With the introduction of the Vision Transformer architecture, cross-attention mechanisms have been used to exchange spatial information between support and query sets, as in CrossTransformers [33], providing an inductive bias for few-shot classification that differs from that of convolutional networks.

4.4 Comparison and Selection of Pre-Training Strategies

Supervised pre-training uses base-class labels to learn semantically discriminative features and is the most direct strategy. Self-supervised pre-training does not depend on labels; instead, it constructs supervisory signals from the data through pretext tasks, of which contrastive learning is the most representative. A Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [34] established the basic framework of contrastive learning. Momentum Contrast (MoCo) [35] introduced a momentum-updated memory queue to decouple the number of negative samples from batch size, while the Masked Autoencoder (MAE) [36] adopted a generative self-supervised route. Systematic evaluation of DINOv2 [37] showed that self-supervised features have strong transfer potential. However, that conclusion was based on approximately

© 2026 The Author(s). Authors retain copyright in their articles.

142 million curated images and large-scale models, making its data scale and computational resources unequal to those of supervised baselines. The claim that "self-supervised pre-training is superior to supervised pre-training" therefore still requires testing under equal scales and protocols.

The emergence of vision-language foundation models such as Contrastive Language-Image Pre-training (CLIP) added a new dimension to this discussion. CLIP [38] obtains strong zero-shot transfer capability by pre-training on large-scale image-text pairs. Context Optimization (CoOp) [39] first demonstrated that learning only a small number of optimizable prompt vectors on CLIP could substantially outperform zero-shot performance in few-shot settings, thereby initiating the new paradigm of "foundation models + lightweight adaptation." Conditional Context Optimization (CoCoOp) [40] introduced input-conditioned dynamic prompts to alleviate overfitting in cross-domain scenarios. Multi-modal Prompt Learning (MaPLe) [41] introduced learnable prompts into both the text and visual branches. Tip-Adapter [42] and CLIP-Adapter [43] explored adaptation paths that are either training-free or train only extremely lightweight residual modules. CaFo [44] cascaded four foundation models: Generative Pre-trained Transformer 3 (GPT-3), DALL-E, CLIP, and DINO. Wang et al. [45] introduced Gaussian discriminant analysis into CLIP downstream classification and achieved performance comparable to advanced adaptation methods without gradient updates. IDEA [46] enhances adaptation using textual descriptions generated by a large language model. These studies share a common feature: rather than relying on episodic meta-training, they directly use the semantic representations learned by foundation models and adapt to novel classes through very small-scale parameter updates or training-free inference. In addition, parameter-efficient fine-tuning techniques such as Low-Rank Adaptation (LoRA) [47] and Visual Prompt Tuning [48] provide a continuum between freezing and fine-tuning the feature extractor, while visual in-context learning [49] is exploring new inference paths that require no gradient updates. Studies such as GPT-3 [50] and Flamingo [51] also show that large-scale models can use a few examples to complete new tasks through in-context learning, which calls for reconsideration of the status of few-shot learning as an independent research problem.

5. Evaluation System

5.1 Standard Datasets

A relatively stable benchmark system has been established for evaluating few-shot classification. miniImageNet [9] is the most widely used. It selects 100 classes from ImageNet, each containing 600 color images of 84 × 84 pixels, and uses a standard split of 64 training classes, 16 validation classes, and 20 test classes. tieredImageNet [52] contains 608 classes and further increases the semantic distance between training and test classes by ensuring that they belong to different high-level semantic nodes. Caltech-UCSD Birds-200-2011 (CUB-200-2011) [53] is a fine-grained bird-classification dataset containing 11,788 images from 200 bird species. Meta-Dataset [54] integrates datasets from multiple sources and is currently one of the benchmarks that best reflects the difficulty of real-world few-shot learning. The limitations of these benchmarks also deserve attention: evaluation results on a single benchmark depend heavily on its class-splitting method and image resolution [54], and performance rankings can be reconfigured under cross-domain settings [55].

Table 1: Common Benchmark Datasets for Few-Shot Image Classification

Dataset	No. of classes	No. of images	Image size	Train/validation/test classes	Main characteristics
miniImageNet [9]	100	600/class; 60,000 total	84 × 84	64 / 16 / 20	Most widely used; derived from ImageNet
tieredImageNet [52]	608	779,165	84 × 84	351 / 97 / 160	Large semantic distance between training and test classes
CUB-200-2011 [53]	200	11,788	Original resolution	100 / 50 / 50	Fine-grained bird classification
Meta-Dataset [54]	10 subsets	Multiple sources	Various	Split by source	Multi-domain; closest to real-world difficulty

Note: The splits of miniImageNet and tieredImageNet follow the standard settings used in the field; test images are generally resized uniformly to 84 × 84 or 224 × 224 pixels.

© 2026 The Author(s). Authors retain copyright in their articles.

5.2 Evaluation Metrics

The standard evaluation metric is N-way K-shot classification accuracy with its 95% confidence interval. Between 600 and 2,000 meta-tasks are randomly sampled from the test classes, and the mean accuracy is reported [9, 10, 19]. Some transductive methods, such as LaplacianShot [25] and TIM [26], use as many as 10,000 meta-tasks to reduce variance, which must also be considered in horizontal comparisons. In the 1-shot setting, metric learning and nearest-neighbor classification that require no parameter updates have a natural advantage [10]. In the 5-shot setting, optimization-based and fine-tuning methods begin to demonstrate their ability to use information from multiple samples [12, 19]. When the sample size increases to 20-shot or more, performance gaps among methods tend to narrow [20].

5.3 Why Existing Rankings Cannot Be Directly Compared

The literature frequently claims that "a method outperforms another method by X percentage points on miniImageNet," but such comparisons often overlook three types of confounding factors. First, inductive and transductive settings differ: transductive methods additionally use unlabeled information from the query set, so their accuracy is naturally higher than that of inductive methods [24, 25, 26]. Second, backbone capacities differ: results obtained with Conv-4 and ResNet-12 cannot be directly compared [19]. Third, the numbers of test meta-tasks differ: differences in confidence intervals caused by sampling 600 versus 2,000 meta-tasks may overwhelm the "improvements" claimed by most methods. Therefore, conclusions about the "competition between paradigms" must be interpreted within their applicable scope. This paper recommends that future studies explicitly report the backbone network, inference setting (inductive/transductive), and number of sampled meta-tasks to improve comparability across studies.

5.4 Performance Summary and Horizontal Comparison

To give readers an overall impression of the performance levels of mainstream methods, this section briefly reviews representative results on miniImageNet (5-way) by method category; the results are summarized in Table 2. It must be emphasized that all values in the table were reported in the original papers. Because the methods differ in backbone network, inference setting, and number of sampled meta-tasks, their numerical results should not be compared directly.

Among metric-learning methods, matching networks [9], prototypical networks [10], and relation networks [11] achieve 1-shot accuracies of approximately 46%-50% with a Conv-4 backbone, increasing to 60%-68% in the 5-shot setting. These methods generally have relatively low 1-shot accuracy because of the difficulty of estimation from a single sample. Among optimization-based meta-learning methods, MAML [12] and Reptile [13] perform similarly with Conv-4 (approximately 47%-49% for 1-shot and approximately 63% for 5-shot), confirming that first-order approximations can already capture the core elements of rapid adaptation. Among transductive methods, TPN [24] achieves approximately 54%/69% with Conv-4 and approximately 59%/76% with ResNet-12. LaplacianShot [25] and TIM [26] achieve approximately 75%/84% and 78%/87%, respectively, with a Wide Residual Network (WRN-28-10). Their advantage comes from additionally using query-set information, so they should not be compared in the same ranking as inductive methods. Under the pre-training and fine-tuning paradigm, Baseline++ [19] achieves approximately 48%/66% with Conv-4 and approximately 52%/76% with ResNet-18, directly illustrating the effect of backbone capacity. It should be noted in particular that foundation-model adaptation methods such as CoOp [39] and Tip-Adapter [42] were not evaluated on traditional benchmarks such as miniImageNet. Instead, they used a protocol covering 11 datasets, including ImageNet, and 1-16 shots. Their numerical results are not comparable to those on traditional benchmarks and should be understood as "relative leadership under a unified protocol." On tieredImageNet, the much larger number of training classes (351 classes) causes most methods to report even higher accuracy than on miniImageNet; for example, the 1-shot accuracy of a prototypical network increases from approximately 49% to approximately 53% [52]. Numerical comparisons across benchmarks must nevertheless remain cautious.

© 2026 The Author(s). Authors retain copyright in their articles.

Table 2: Reported Accuracies of Representative Methods on miniImageNet (5-way)

Method	Category	Backbone	1-shot (%)	5-shot (%)	Setting	Meta-tasks
Matching Networks [9]	Metric learning	Conv-4	46.60	60.00	Inductive	600
Prototypical Networks [10]	Metric learning	Conv-4	49.42	68.20	Inductive	600
Relation Networks [11]	Metric learning	Conv-4	50.44	65.32	Inductive	600
MAML [12]	Optimization-based meta-learning	Conv-4	48.70	63.11	Inductive	600
Reptile [13]	Optimization-based meta-learning	Conv-4	47.07	62.74	Inductive	600
Baseline++ [19]	Pre-training and fine-tuning	Conv-4	48.24	66.43	Inductive	600
Baseline++ [19]	Pre-training and fine-tuning	ResNet-18	51.87	75.68	Inductive	600
TPN [24]	Transductive	Conv-4	53.75	69.43	Transductive	600
TPN [24]	Transductive	ResNet-12	59.46	75.65	Transductive	600
LaplacianShot [25]	Transductive	WRN-28-10	74.86	84.13	Transductive	10,000
TIM [26]	Transductive	WRN-28-10	77.80	87.40	Transductive	10,000

Note: The values are taken from the results reported in the original papers. The matching-network results use the best setting in the original paper (cosine similarity + cosine full-context embedding + fine-tuning). Because the methods differ in backbone network, inference setting, and number of sampled meta-tasks, direct numerical comparison between rows is not rigorous. Transductive methods additionally use query-set information and cannot be placed in the same ranking as inductive methods.

From the perspective of knowledge encoding, the five categories exhibit systematic differences in training paradigm, inference cost, and failure scenarios. Their multidimensional characteristics are compared in Table 3.

Overall, none of the five categories is absolutely superior to the others; their suitability depends on the specific task conditions. When labeled samples are extremely scarce and query-set information may be used, transductive methods have a clear advantage. When abundant base-class data and large-scale pre-training are available, the pre-training and fine-tuning paradigm generally achieves the best performance. When the task distribution is complex and rapid adaptation to new tasks is required, optimization-based meta-learning remains irreplaceable. This conclusion also echoes the central argument of this paper: the categories represent different choices of knowledge-encoding form rather than substitutes for one another.

Table 3: Multidimensional Comparison of Five Categories of Few-Shot Image Classification Methods

Category	Knowledge encoding	Training	Novel-class adaptation	Inference cost	Backbone reliance	Typical failure scenario
Metric learning	Distance-measurement rules in an embedding space	Episodic or ordinary batches	No gradient updates	Low	Moderate	The single-prototype assumption fails for multimodal within-class distributions
Optimization-based meta-learning	Initial parameters sensitive to task changes	Episodic	A few gradient updates	Moderate	Relatively high	When task distributions are complex and optima are dispersed, second-order information is insufficient
Data augmentation	Reasonable extrapolation of	Ordinary batch training,	Generate augmented samples in the	Moderate	Relatively high	Generated samples are semantically distorted or

© 2026 The Author(s). Authors retain copyright in their articles.

Category	Knowledge encoding	Training	Novel-class adaptation	Inference cost	Backbone reliance	Typical failure scenario
	within-class distributions	often combined with meta-training	support set or feature space			introduce contamination from pseudo-samples
Pre-training and fine-tuning	General transferable feature representations	Ordinary batches	Fine-tuning or nearest-neighbor classification	Low	Extremely high	A substantial domain shift exists between base and novel classes
Parameter generation and probabilistic inference	Conditional distribution or posterior of classifier parameters	Episodic or ordinary batches	Latent-space optimization or posterior inference	Moderate	Moderate	The latent space is difficult to design

Note: Table 3 lists the knowledge-encoding forms and applicable characteristics of the five major paradigms. Transduction and semi-supervision are inference settings rather than knowledge-encoding methods. They form a dimension orthogonal to the five categories above and can be superimposed on any paradigm; they are therefore not listed as a separate category.

6. Challenges and Future Directions

6.1 Cross-Domain Generalization

The strong performance of current few-shot learning methods depends largely on the assumption that base and novel classes come from the same or similar data domains. When this assumption is violated, for example, when a model is trained on natural images but must classify medical or remote-sensing images, the performance of most methods declines substantially [19, 20]. Current research addresses this issue from several perspectives. Feature-disentanglement methods attempt to separate domain-invariant from domain-specific features; domain-adversarial training introduces a domain discriminator during feature learning [56]; and Meta-Dataset [54] provides a unified evaluation platform for cross-domain few-shot learning. However, this direction still lacks a unified theoretical framework and generally accepted evaluation standards.

6.2 New Paradigms for Few-Shot Learning in the Era of Foundation Models

Since 2022, work represented by CLIP adaptation has fundamentally changed the technical route of few-shot learning. Methods including CoOp [39], CoCoOp [40], MaPLe [41], Tip-Adapter [42], CLIP-Adapter [43], CaFo [44], and IDEA [46] collectively show that, on top of a strong pre-trained vision-language model, extremely lightweight parameter adaptation or even training-free inference can substantially outperform zero-shot performance. As a result, "foundation-model adaptation" has become one of the most active research directions. This shift also raises new questions: the generalization of prompt learning has become a focus [40], the trade-off between adaptation efficiency and data scale must be balanced, and visual in-context learning [49] is exploring inference paths without gradient updates. Together, these developments point to a broader conclusion: few-shot learning is moving from an "algorithm-design-driven" stage to a new stage characterized by foundation representations and lightweight adaptation.

6.3 From Classification to More Complex Visual Tasks

Current few-shot learning research focuses mainly on image classification, but real-world applications demand far more than classification. Few-shot object detection requires a model to locate and recognize novel-class objects from only a few labeled samples. The feature-reweighting method of Kang et al. [57] and few-shot R-CNN frameworks laid the foundation for this direction. Few-shot semantic segmentation extends the challenge to the pixel level; Shaban et al. [58] pioneered the decomposition of segmentation into a classification branch and a segmentation branch. These more complex tasks share the core challenges of few-shot classification while introducing additional complexity.

© 2026 The Author(s). Authors retain copyright in their articles.

6.4 Testable Prospective Directions

Direction 1: Test the hypothesis that "foundation models have already seen the test classes." Using public pre-training corpora such as the Large-scale Artificial Intelligence Open Network (LAION), count the frequency of each test class from miniImageNet, CUB, and Meta-Dataset in the pre-training corpus and examine the correlation between class frequency and few-shot accuracy for that class. If the correlation is significant, the validity of existing benchmarks must be reconsidered in the era of foundation models. The results would also help explain part of the advantage achieved by foundation-model methods on public benchmarks.

Direction 2: Use "equivalent sample size" to measure the efficiency of knowledge injection uniformly. For each category of methods, measure the equivalent number of labeled samples saved relative to "training from scratch without prior knowledge": on the same backbone and the same novel classes, how many samples are required by training from scratch to achieve the 1-shot/5-shot accuracy of the method? This would convert "knowledge encoding" from a conceptual description into a comparable scale measured in numbers of samples and provide a more unified basis for comparing different methods.

Direction 3: Establish separate evaluation protocols and paired benchmarks for transductive and inductive settings. Systematically reproduce several representative methods under strictly inductive and transductive settings and quantify the gain brought by transduction; then examine whether this gain changes with query-set size. If transductive gains are generally larger than the improvements attributed to the mechanisms proposed by most methods, existing rankings across settings may be driven mainly by differences in setting, and their conclusions will need to be reinterpreted.

7. Conclusion

Starting from the dual challenges of inter-class distribution shift and insufficient within-class samples, this paper systematically reviews the state of research on few-shot image classification. According to their core operating mechanisms, existing methods can be grouped into five major paradigms: metric learning, optimization-based meta-learning, data augmentation, pre-training and fine-tuning, and parameter generation and probabilistic methods [6, 19, 54]. These paradigms differ profoundly in how they encode knowledge: as distance rules in an embedding space [10], initial parameters that can adapt rapidly [12], reasonable extrapolations of within-class distributions [14, 15], general transferable feature representations [19, 20], or conditional distributions or posterior-inference processes for classifier parameters [21]. In high-performing methods, however, these strategies often work together in combination.

With the rapid development of foundation models and multimodal learning, few-shot learning is undergoing a paradigm shift from "algorithm-design driven" to "representation-quality driven." How to reassess the contributions of classical methods under this new paradigm and how to establish evaluation systems that better reflect real-world application scenarios will be important topics for future research.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems 25. Lake Tahoe: NIPS, 2012: 1097-1105.

[2] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248-255.

[3] Zheng G Y, Liu X B, Han G H. A review of computer-aided detection and diagnosis systems for medical imaging[J]. Journal of Software, 2018, 29(5): 1471-1514. (in Chinese)

[4] Tao X, Hou W, Xu D. A review of surface defect detection methods based on deep learning[J]. Acta Automatica Sinica, 2021, 47(5): 1017-1034. (in Chinese)

[5] Zhang B. The era of remote sensing big data and intelligent information extraction[J]. Journal of Remote Sensing, 2018, 22(4): 529-542. (in Chinese)

© 2026 The Author(s). Authors retain copyright in their articles.

[6] Liu Y, Liu H Y, Fan J L, et al. A survey of few-shot image classification methods based on deep learning[J]. Journal of Image and Graphics, 2022, 27(5): 1491-1512. (in Chinese)

[7] Zhao K L, Jin X L, Wang Y Z. A survey of few-shot learning[J]. Journal of Software, 2021, 32(2): 349-369. (in Chinese)

[8] Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition[C]//ICML Deep Learning Workshop. Lille: ICML, 2015.

[9] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[C]//Advances in Neural Information Processing Systems 29. Barcelona: NIPS, 2016: 3630-3638.

[10] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[C]//Advances in Neural Information Processing Systems 30. Long Beach: NIPS, 2017: 4077-4087.

[11] Sung F, Yang Y, Zhang L, et al. Learning to compare: Relation network for few-shot learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1199-1208.

[12] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the International Conference on Machine Learning. Sydney: PMLR, 2017: 1126-1135.

[13] Nichol A, Achiam J, Schulman J. On first-order meta-learning algorithms[J]. arXiv preprint arXiv:1803.02999, 2018.

[14] Ma D A, Tang P, Zhao L J, et al. A survey of image data augmentation methods for deep learning[J]. Journal of Image and Graphics, 2021, 26(3): 487-502. (in Chinese)

[15] Wang K F, Gou C, Duan Y J, et al. Research progress and prospects of generative adversarial networks[J]. Acta Automatica Sinica, 2017, 43(3): 321-332. (in Chinese)

[16] Hariharan B, Girshick R. Low-shot visual recognition by shrinking and hallucinating features[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3018-3027.

[17] Wang Y X, Girshick R, Hebert M, et al. Low-shot learning from imaginary data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7278-7286.

[18] Yang S, Liu L, Xu M. Free lunch for few-shot learning: Distribution calibration[C]//International Conference on Learning Representations. Virtual: ICLR, 2021.

[19] Chen W Y, Liu Y C, Kira Z, et al. A closer look at few-shot classification[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019.

[20] Tian Y, Wang Y, Krishnan D, et al. Rethinking few-shot image classification: a good embedding is all you need?[C]//European Conference on Computer Vision. Glasgow: Springer, 2020: 266-282.

[21] Rusu A A, Rao D, Sygnowski J, et al. Meta-learning with latent embedding optimization[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019.

[22] Finn C, Xu K, Levine S. Probabilistic model-agnostic meta-learning[C]//Advances in Neural Information Processing Systems 31. Montréal: NIPS, 2018: 7052-7062.

[23] Gordon J, Bronskill J, Bauer M, et al. Meta-learning probabilistic inference for prediction[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019.

[24] Liu Y, Lee J, Park M, et al. Learning to propagate labels: Transductive propagation network for few-shot learning[C]//International Conference on Learning Representations. New Orleans: ICLR, 2019.

[25] Ziko I, Dolz J, Granger E, et al. Laplacian regularized few-shot learning[C]//Proceedings of the International Conference on Machine Learning. Virtual: PMLR, 2020: 11660-11670.

[26] Boudiaf M, Masud Z I, Rony J, et al. Transductive information maximization for few-shot learning[C]//Advances in Neural Information Processing Systems 33. Virtual: NIPS, 2020: 2445-2457.

© 2026 The Author(s). Authors retain copyright in their articles.

[27] Ye H J, Hu H, Zhan D C, et al. Few-shot learning via embedding adaptation with set-to-set functions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 8808-8817.

[28] Zhang C, Cai Y, Lin G, Shen C. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 12203-12213.

[29] Heidari M, Alchihabi A, En Q, et al. Adaptive parametric prototype learning for cross-domain few-shot classification[C]//Proceedings of the 27th International Conference on Artificial Intelligence and Statistics. Valencia: PMLR, 2024: 1369-1377.

[30] Wang Q, Lv Y, Feng Y, et al. A simple yet effective strategy to robustify the meta learning paradigm[C]//Advances in Neural Information Processing Systems 36. New Orleans: NIPS, 2023.

[31] Wang Y O, Chung Y, Wu C H, et al. Domain gap embeddings for generative dataset augmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024.

[32] Hu S X, Li D, Stühmer J, et al. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 9068-9077.

[33] Doersch C, Gupta A, Zisserman A. CrossTransformers: spatially-aware few-shot transfer[C]//Advances in Neural Information Processing Systems 33. Virtual: NIPS, 2020: 21981-21993.

[34] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the International Conference on Machine Learning. Virtual: PMLR, 2020: 1597-1607.

[35] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9729-9738.

[36] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 16000-16009.

[37] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning robust visual features without supervision[J]. arXiv preprint arXiv:2304.07193, 2023.

[38] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. Virtual: PMLR, 2021: 8748-8763.

[39] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[40] Zhou K, Yang J, Loy C C, et al. Conditional prompt learning for vision-language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 16795-16805.

[41] Khattak M U, Rasheed H, Maaz M, et al. MaPLe: Multi-modal prompt learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 19113-19122.

[42] Zhang R, Fang R, Zhang W, et al. Tip-Adapter: Training-free adaption of CLIP for few-shot classification[C]//European Conference on Computer Vision. Tel Aviv: Springer, 2022: 493-510.

[43] Gao P, Geng S, Zhang R, et al. CLIP-Adapter: Better vision-language models with feature adapters[J]. International Journal of Computer Vision, 2024, 132(2): 581-595.

© 2026 The Author(s). Authors retain copyright in their articles.

[44] Zhang R, Hu X, Li B, et al. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 15211-15222.

[45] Wang Z, Liang J, Sheng L, et al. A hard-to-beat baseline for training-free CLIP-based adaptation[C]//International Conference on Learning Representations. Vienna: ICLR, 2024.

[46] Ye Z, Jiang F, Wang Q, et al. IDEA: Image description enhanced CLIP-Adapter[J]. arXiv preprint arXiv:2501.08816, 2025.

[47] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models[C]//International Conference on Learning Representations. Virtual: ICLR, 2022.

[48] Jia M, Tang L, Chen B C, et al. Visual prompt tuning[C]//European Conference on Computer Vision. Tel Aviv: Springer, 2022: 709-727.

[49] Zhang Y, Zhou K, Liu Z. What makes good examples for visual in-context learning?[J]. arXiv preprint arXiv:2301.13670, 2023.

[50] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems 33. Virtual: NIPS, 2020: 1877-1901.

[51] Alayrac J B, Donahue J, Luc P, et al. Flamingo: A visual language model for few-shot learning[C]//Advances in Neural Information Processing Systems 35. New Orleans: NIPS, 2022: 23716-23736.

[52] Ren M, Triantafillou E, Ravi S, et al. Meta-learning for semi-supervised few-shot classification[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018.

[53] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 dataset[R]. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[54] Triantafillou E, Zhu T, Dumoulin V, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples[C]//International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

[55] Guo Y, Codella N C, Karlinsky L, et al. A broader study of cross-domain few-shot learning[C]//European Conference on Computer Vision. Virtual: Springer, 2020: 124-141.

[56] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. Journal of Machine Learning Research, 2016, 17(59): 1-35.

[57] Kang B, Liu Z, Wang X, et al. Few-shot object detection via feature reweighting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 8420-8429.

[58] Shaban A, Bansal S, Liu Z, et al. One-shot learning for semantic segmentation[C]//Proceedings of the British Machine Vision Conference. London: BMVA Press, 2017.

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.

© 2026 The Author(s). Authors retain copyright in their articles.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.