

FEW-SHOT IMAGE CLASSIFICATION: A SURVEY

Daneshpour A. ¹, Bucz S. , Jantassova D.,  Akhmedov R. 

Abstract Few-shot image classification (FSIC) addresses image recognition when only a few labeled examples are available for each novel class. This survey reviews five major FSIC method families: meta-learning, transfer learning, data augmentation, attribute-based methods, and vision-language model adaptation. The N -way K -shot setting is defined, and standard training, theoretical foundations, benchmark datasets, and evaluation protocols are examined. Results for 5-way 1-shot and 5-shot classification are compared across four widely used benchmarks. The survey also examines self-supervised and semi-supervised learning, computational efficiency, and few-shot class-incremental learning. The analysis shows that representation quality strongly affects FSIC performance. Large pre-trained and vision-language models achieve high benchmark accuracy, but their performance is associated with substantially larger pre-training datasets. Direct comparisons across methods remain difficult because backbone capacity, pre-training data, and evaluation settings vary. Cross-domain generalization and practical deployment also remain challenging. Connections with PAC-Bayes theory, information theory, and causal representation learning are discussed. Finally, open research challenges are identified in robustness, long-tail recognition, fine-grained classification, multimodal learning, efficient adaptation, and standardized evaluation.

Keywords: Few-Shot Learning, Meta-Learning, Image Classification, Vision-Language Models, Class-Incremental Learning, Transfer Learning, Prototypical Networks, CLIP, Data Augmentation, Episodic Training.

AMS Mathematics Subject Classification: 68T05, 68T07, 68T10, 68T45, 62H30.

DOI: 10.32523/2306-6172-2026-14-3-4-24

1 Introduction

Few-shot image classification (FSIC) addresses image recognition when only a few labeled examples are available for each novel class. The recognition accuracy is often achieved by deep models. Severe imbalance [1] and class imbalance should also come to consideration [2] where uncertain network design is addressed through mathematical optimization [3]. However, extensive labeled datasets are usually required [4, 5]. New objects are recognized by humans from few examples. Prior visual knowledge is used for this recognition [6, 7]. Few-shot image classification (FSIC) is motivated by this sample-efficiency gap. Novel categories are denoted by $\mathcal{C}_{\text{test}}$. Base categories are denoted by $\mathcal{C}_{\text{train}}$. The two category sets are disjoint. Only K labeled examples are provided for each novel category. K is usually limited to ten or fewer examples. Severe overfitting is caused by random initialization [8, 9]. This risk is reduced by pre-trained encoders. Dedicated meta-learning methods are often matched by strong transfer baselines [10, 11, 12]. Thus, empirical risk minimization is strongly affected by representation quality. Signal cannot be separated from noise by one example alone. External knowledge is therefore required. One knowledge source is supplied by episodic biases [13, 14].

¹Corresponding Author.

Another source is supplied by pre-trained visual features [12]. Semantic priors are also supplied by CLIP [15]. Large-scale annotation is costly and slow. Rare disease data are especially limited [16]. Species monitoring is affected by similar limits [17]. Industrial defect data are also scarce [18]. Similar imbalance is found in other data-constrained applications. FSIC is therefore relevant beyond academic benchmarks.

The N -way K -shot episodic protocol is used for standard evaluation. N novel classes are included in each episode. K support images and Q queries are provided per class. Inference-time scarcity is reproduced by this protocol [13]. Mini-batch training is also widely used [12]. Another evaluation setting is provided by transductive inference [19].

Modern episodic evaluation is established by Matching Networks [13]. Classification is simplified through Prototypical Network centroids [20]. An initialization for rapid adaptation is learned by MAML [14]. Episodic training is later challenged by strong transfer baselines [10, 12]. Large-scale vision-language priors are then introduced by CLIP [15]. Lightweight prompt adaptation is subsequently adopted [21, 22]. Accuracy is improved, but fair comparison is complicated. Five major FSIC method families are classified in this survey. PAC-Bayes and information-theoretic links are explained. Four standard benchmarks are compared. Episode-based and mini-batch training are assessed. FSCIL and its stability-plasticity conflict are reviewed. Self-supervised and semi-supervised extensions are examined. Seven open research problems are also identified. Related surveys are reviewed in Section 2. The problem is defined in Section 3. Theory is presented in Section 4. Methods are reviewed in Section 5. Extensions are discussed in Section 6. Benchmarks and results are presented in Sections 7–8. Efficiency and FSCIL are then examined. The survey is completed by discussion, challenges, and conclusions.

2 Related Surveys

Few-shot learning is broadly surveyed by Wang et al. [23]. Meta-learning is reviewed from an optimization perspective [24]. Vision-language models are not covered by earlier surveys [25]. Three additions are provided here. First, CLIP-based adaptation is included. Second, standard FSIC and FSCIL are reviewed together. Third, accuracy and computational cost are jointly assessed. Regarding the Scope and exclusions; Standard N -way K -shot image classification is covered. Its class-incremental extension is also included. Priority is given to methods on four common benchmarks. Distinct contributions from five families are selected. Historical variants are shown in Fig. 3. They are not all discussed separately. Different spatial predictions are required for detection and segmentation. They are treated briefly in Section 12. Most graph-based methods are excluded. Relevant label-propagation methods are retained. Cross-modal and video tasks are also excluded. Thus, image-level classification remains the main subject.

3 Problem Formulation and Training Protocol

3.1 Formal Problem Definition

Definition 1 (Few-Shot Image Classification). *A labeled base dataset is denoted by $\mathcal{D}_{train}$. Its classes are grouped within $\mathcal{C}_{base}$. Novel examples are contained in $\mathcal{D}_{test}$. Their classes are*

denoted by $\mathcal{C}_{\text{novel}}$. The base and novel classes are disjoint. An N -way K -shot task $\mathcal{T}$ contains:

$$\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N \times K} \quad (\text{support set}), \quad (1)$$

$$\mathcal{Q} = \{(x_j^q, y_j^q)\}_{j=1}^{N \times Q} \quad (\text{query set}), \quad (2)$$

where all labels belong to N novel classes. The class count is denoted by N . The support count per class is denoted by K . Common values are $K \in \{1, 2, 5, 10, 20\}$. A classifier f_θ is learned from $\mathcal{S}$. Its accuracy is measured on $\mathcal{Q}$.

3.2 Training Paradigms

Two main training paradigms are shown in Fig. 1.

Episode-based training. The training set is divided into few-shot episodes. The target inference task is imitated by each episode. The standard protocol is established by Matching Networks [13]. N base classes are sampled in each episode. K supports and Q queries are drawn from each class. Query classification is conditioned on the support set.

Mini-batch-based training. Standard gradient descent is applied to base-class batches. Few-shot transfer is supported by regularized features. Competitive results are often obtained by this approach [12]. Episodic optimization may therefore be outweighed by strong representations.

Hybrid paradigms. Episodic and standard phases are combined by hybrid methods. An auxiliary episodic loss may also be used [26]. Episodic bias is retained with stable standard optimization.

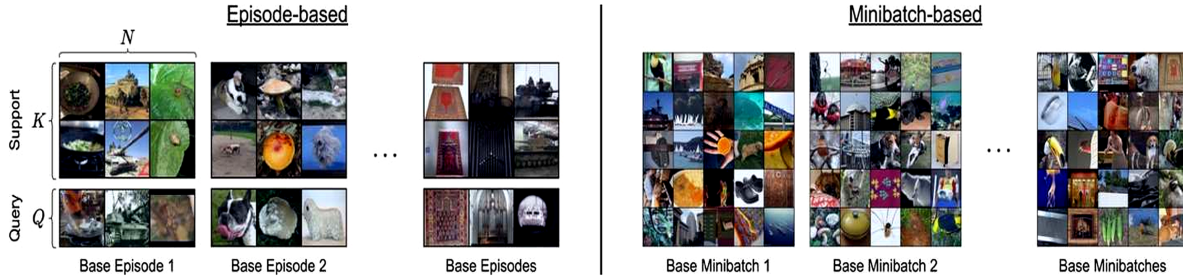


Figure 1: Two FSIC training paradigms. Episodes imitate few-shot tests. Mini-batches support representation transfer.

3.3 Episode-Based Objective

Parameters θ are optimized through expected episodic cross-entropy:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} \left[-\frac{1}{|\mathcal{Q}|} \sum_{(x,y) \in \mathcal{Q}} \log p_\theta(y \mid x, \mathcal{S}) \right], \quad (3)$$

Here, p_θ denotes the support-conditioned query probability.

3.4 Task Distribution and Generalization Gap

Performance is affected by the task distribution $p(\mathcal{T})$. Training tasks are sampled from base classes. Test tasks are sampled from held-out classes. The generalization gap is determined

by semantic class divergence. This gap is increased in tieredImageNet through WordNet [27]. A harder evaluation is therefore provided.

4 Theoretical Foundations

4.1 PAC-Bayes Perspective on Meta-Learning

Meta-learning generalization can be studied through PAC-Bayes theory [28, 29]. Assume that M independent tasks are sampled. Each task is drawn from $p(\mathcal{T})$. The task loss is bounded within $[0, 1]$. A sample-independent parameter prior is denoted by $P(\theta)$. A learned posterior is denoted by $Q(\theta)$. Empirical task risk is defined as

$$\hat{\mathcal{L}}_{\mathcal{T}_m}(\theta) = \frac{1}{|\mathcal{Q}_m|} \sum_{(x,y) \in \mathcal{Q}_m} \ell(f_{\theta}(x | \mathcal{S}_m), y), \quad (4)$$

The query risk is conditioned on support set $\mathcal{S}_m$. Population risk is denoted by $\mathcal{L}_{\mathcal{T}}(\theta)$. For $\delta \in (0, 1)$, the following bound is obtained:

$$\mathbb{E}_{\theta \sim Q} \left[\mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} [\mathcal{L}_{\mathcal{T}}(\theta)] \right] \leq \mathbb{E}_{\theta \sim Q} \left[\frac{1}{M} \sum_{m=1}^M \hat{\mathcal{L}}_{\mathcal{T}_m}(\theta) \right] + \sqrt{\frac{\text{KL}(Q \| P) + \ln(M/\delta)}{2M}}, \quad (5)$$

Here, posterior divergence is measured by $\text{KL}(Q \| P)$. Large deviations are penalized by the bound. Compact transferable representations are therefore favored.

Reusable initialization is learned by MAML [14]. A reusable embedding is learned by ProtoNet [20]. Both objects are conceptually related to a PAC-Bayes prior. No formal equivalence is implied. One parameter point is used by MAML. A deterministic map is used by ProtoNet. By contrast, P must be a probability distribution. An explicit probabilistic construction is therefore required. Gaussian constructions are used in PAC-Bayesian meta-learning [28, 29]. The connection is presented only as an analogy.

4.2 Information-Theoretic Lower Bounds

Discrete prediction error is bounded by Fano’s inequality [30, 31]. It can be applied to an N -way episode. Assume that query labels are sampled uniformly. Then, $H(y) = \log_2 N$. Any prediction rule is denoted by $\hat{y} = g(x, \mathcal{S})$. The resulting lower bound is

$$\mathbb{P}[\hat{y} \neq y] \geq 1 - \frac{I(y; \mathcal{S}) + 1}{\log_2 N}, \quad (6)$$

Here, support information is measured by $I(y; \mathcal{S})$. This FSIC form is derived from Chen et al. [32]. The general inequality is established by Fano [30]. A modern derivation is provided by Cover and Thomas [31]. Uniform class priors are required by this form. Otherwise, $\log_2 N$ is replaced by $H(y)$. Smaller gains are provided by additional shots with discriminative embeddings. This pattern is observed for strong pre-trained models.

4.3 Connections to Causal Representation Learning

Distributional robustness may be supported by causal representations [33]. Invariant mechanisms are preferred over dataset-specific correlations. However, this claim remains conceptual. Causal features are not proven requirements for FSIC. Strong results are achieved without explicit causal models. Some spurious correlations may be suppressed by large diverse datasets.

Cross-domain failures may still be caused by non-causal features. This explanation remains difficult to verify. This issue is addressed by domain-invariant features [34]. Another approach is provided by causal meta-learning [35]. Current evidence remains limited and benchmark-specific.

5 Methods

Five method families are shown in Fig. 2. Meta-learning, transfer, augmentation, attributes, and VLM adaptation are included. Meta-learning is divided into three subfamilies. These subfamilies are metric-based, model-based, and optimization-based.

Each family is defined by its main inductive bias. Episodic signals are used by meta-learning. Pre-trained backbones are used by transfer methods. Synthetic data are used by augmentation methods. Semantic descriptors are used by attribute methods. Image-text priors are used by VLM methods. These categories are not disjoint. Knowledge distillation is used by RFS [12]. CLIP-Adapter is also classified as transfer learning [22]. Augmentation and VLM adaptation are combined by text-guided diffusion [36]. The taxonomy is therefore used only for organization.

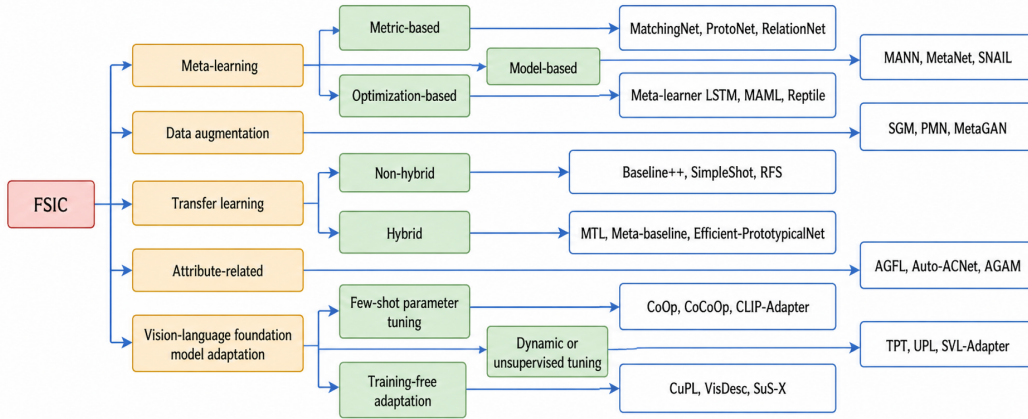


Figure 2: Classification of existing approaches for FSIC with three representatives for each category.

5.1 Meta-Learning Methods

Meta-learning is often described as learning to learn [37]. Models are trained across tasks from $p(\mathcal{T})$. Rapid adaptation is encouraged by this task distribution.

5.1.1 Metric-Based Methods

An embedding ϕ_θ is learned by metric methods. Classification is reduced to nearest-neighbor retrieval. The class prototype is defined as

$$\mathbf{p}_c = \frac{1}{K} \sum_{(x_i^s, y_i^s) \in \mathcal{S}: y_i^s = c} \phi_\theta(x_i^s), \quad (7)$$

The query probability is then defined as

$$p(y = c \mid x, \mathcal{S}) = \frac{\exp(-d(\phi_\theta(x), \mathbf{p}_c))}{\sum_{c'} \exp(-d(\phi_\theta(x), \mathbf{p}_{c'}))}, \quad (8)$$

Squared Euclidean distance is used by ProtoNet [20]. Cosine distance is used by MatchingNet [13].

Matching Networks [13] establish episodic training. Full-context attention is used for class scores:

$$\hat{y} = \sum_{i=1}^{N \times K} a(x, x_i^s) y_i^s, \quad a(x, x_i^s) = \frac{e^{c(\phi(x), \phi(x_i^s))}}{\sum_j e^{c(\phi(x), \phi(x_j^s))}}, \quad (9)$$

Here, $c(\cdot, \cdot)$ denotes cosine similarity.

Prototypical Networks use Euclidean class centroids [20]. Only N prototype distances are required. $N \times K$ support distances are required by Matching Networks.

Relation Networks replace fixed metrics with learned modules [38]:

$$r_c = g_\phi(\text{concat}[\phi_\theta(x), \mathbf{p}_c]), \quad (10)$$

Here, $r_c \in [0, 1]$ is the relation score. Greater metric flexibility is provided by this module.

FEAT contextualizes support embeddings with a Transformer [39]. The contextualized prototype is

$$\tilde{\mathbf{p}}_c = \frac{1}{K} \sum_{(x_i^s, y_i^s) \in \mathcal{S}: y_i^s = c} \text{Transformer}(\phi_\theta(x_i^s); \phi_\theta(\mathcal{S})), \quad (11)$$

No test-time gradient is required by this adaptation.

DeepEMD defines classification through optimal transport [40]. Dense feature maps are compared by Earth Mover’s Distance:

$$d_{\text{EMD}}(x, x') = \min_{\mathbf{F} \geq 0} \sum_{m,n} F_{mn} d_{mn}, \quad \text{s.t.} \sum_n F_{mn} = w_m, \sum_m F_{mn} = w'_n, \quad (12)$$

Here, d_{mn} denotes transport cost between local features. Complex spatial patterns are supported by this comparison.

Major metric choices are summarized in Table 1.

Table 1: Comparison of metric-based FSIC methods.

Method	Metric	Encoder	Episodic	Context
MatchingNet [13]	Cosine	Conv4/FCE	✓	Full
ProtoNet [20]	Euclidean	Conv4/ResNet	✓	Centroid
RelationNet [38]	Learned CNN	Conv4	✓	Centroid
FEAT [39]	Transformer	ResNet12	✓	Inductive
DeepEMD [40]	EMD	ResNet12	✓	Dense
DCAP [41]	Pooling	ResNet12	✓	Attentive

5.1.2 Model-Based Methods

Adaptation is encoded within model-based architectures. Adaptation can therefore be produced by one forward pass.

Episodic memory is added by MANN [42]. Fast episodic weights are used by MetaNet [43]. Temporal convolutions and attention are combined by SNAIL [44].

These methods are limited by memory capacity. Predictions may also be affected by support order. Order sensitivity is reduced by temporal convolutions [44]. Large values of N and K remain difficult.

5.1.3 Optimization-Based Methods

Parameters for rapid adaptation are learned by optimization-based methods. MAML provides the principal objective [14]:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{\mathcal{T}} [\mathcal{L}_{\mathcal{T}}(\theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}}(\theta))], \quad (13)$$

Here, α denotes the inner-loop rate. The initial parameters θ are learned by the outer loop.

Parameter-specific rates and directions are learned by Meta-SGD [45]. A first-order approximation is used by Reptile [46]. Lower computational cost is achieved by this approximation. Implicit differentiation is used by iMAML [47]. Explicit inner-loop unrolling is therefore avoided.

Feature reuse is identified as a major limitation [48]. Most MAML adaptation is performed within the final classifier. Little change is observed within the feature extractor. Stronger transfer baselines are motivated by this result.

Optimization-based methods are traced in Fig. 3.

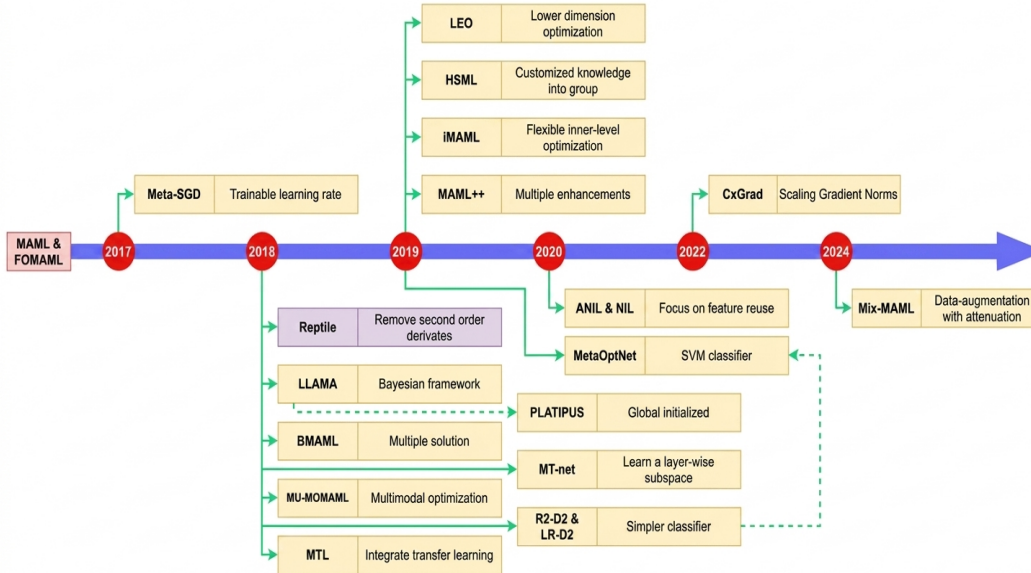


Figure 3: Optimization-based methods from 2017 to 2024. Dashed lines denote direct inheritance.

5.2 Transfer Learning Methods

Two stages are used by transfer learning. First, a backbone is pre-trained on a large dataset. Second, its features are adapted for few-shot inference. The process is shown in Fig. 4.

Non-hybrid methods use a fixed pre-trained backbone. Cosine distance is applied to ResNet-18 features by Baseline++ [10]. Many meta-learning results are matched by this baseline. L2 normalization and nearest neighbors are combined by SimpleShot [11].

Hybrid methods combine pre-training with task-specific adaptation. Self-supervised contrastive losses may be added [49]. Knowledge distillation is used by RFS [12]. The student’s 1-shot features are improved by teacher knowledge.

Fine-tuning strategies. Catastrophic forgetting may be caused by support adaptation [50, 19]. A self-supervised signal is provided by query entropy [50]. The unlabeled query manifold is used by label propagation [19]. Decision boundaries are improved in transductive settings.

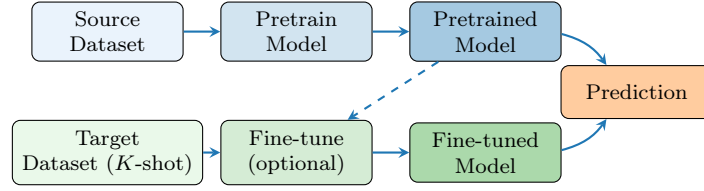


Figure 4: FSIC transfer pipeline. Pre-trained features receive lightweight target adaptation.

5.3 Data Augmentation Methods

Data scarcity can be reduced through synthetic novel-class examples. The process is shown in Fig. 5.

GAN-based synthesis. Image generation is conditioned on class embeddings by MetaGAN [51]. Base-class feature variance is transferred by SGM [52]. Image synthesis is avoided by this feature-space method.

Feature augmentation. Class statistics are transferred to novel classes by PMN [52]. Synthetic features are then produced for training. GAN instability is avoided by this method.

Diffusion-based augmentation. Realistic few-shot samples are generated by diffusion models [36]. Semantic control is provided through text prompts. Class-conditional synthesis is therefore supported.

The expected classification loss under augmented data is:

$$\mathcal{L}_{\text{aug}}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(f_{\theta}(x), y)] + \lambda \mathbb{E}_{(\tilde{x},y) \sim \mathcal{G}} [\ell(f_{\theta}(\tilde{x}), y)], \quad (14)$$

Here, $\mathcal{G}$ denotes the synthetic distribution. Real and synthetic losses are balanced by λ .

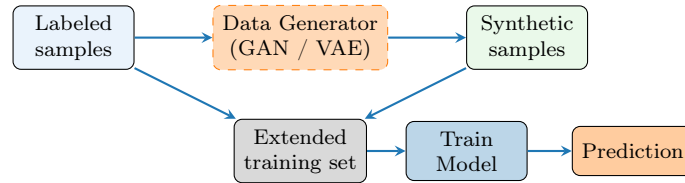


Figure 5: FSIC augmentation pipeline. A generator expands the novel-class training set.

5.4 Attribute-Related Methods

Auxiliary semantic descriptors are used by attribute methods. Visual attributes, word embeddings, and class names are included. Base and novel classes are linked by these descriptors.

Visual and semantic spaces are aligned by AGFL [53]. AADNet combines image-level and class-level similarities:

$$\hat{y} = \arg \max_c [\beta \cdot d_{\text{feat}}(\phi(x), \mathbf{p}_c) + (1 - \beta) \cdot d_{\text{attr}}(\psi(x), \mathbf{a}_c)], \quad (15)$$

Here, ψ is an attribute encoder. The class attribute vector is denoted by $\mathbf{a}_c$.

Cross-sample and intra-sample attention are added by AGAN [54]. More accurate prototypes are produced from feature dependencies. Common semantic attributes are supplied by Word2Vec [55]. Richer language representations are supplied by BERT [56].

5.5 Vision-Language Foundation Model Adaptation

FSIC performance is substantially improved by large VLMs. CLIP is trained on about 400 million image-text pairs [15]. Its contrastive loss is defined as

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^B \exp(\text{sim}(v_i, t_j)/\tau)}, \quad (16)$$

Here, v_i and t_i are paired embeddings. Temperature is denoted by τ . Batch size is denoted by B .

Images are compared with class-name prompts by zero-shot CLIP. Several few-shot adaptation strategies are used.

CoOp replaces fixed prompts with learned context vectors [21]. These vectors are fitted on the support set. The class score is

$$p(c | x) \propto \exp(\text{sim}(v_{\text{img}}(x), t_{\text{txt}}([v_1, \dots, v_M, c])) / \tau). \quad (17)$$

CoCoOp conditions prompts on each input image [57]. Unseen-class transfer is improved. Additional inference cost is required.

CLIP-Adapter adds residual modules to frozen CLIP [22]:

$$\phi'(x) = \alpha \cdot \text{Adapter}(\phi(x)) + (1 - \alpha) \cdot \phi(x), \quad (18)$$

Here, residual blending is controlled by α . A two-layer MLP is fitted on support examples.

Training-free adaptation uses generated class descriptions. This strategy is applied with GPT-3 by CuPL [58]. The resulting text embeddings are averaged. Few-shot gradients are not required.

Test-time prompt tuning (TPT) minimizes test-view entropy [59]. Context vectors are adapted for each test image. Labeled support examples are not used. Thus, TPT is not a conventional K -shot method. Its reported columns only preserve episodic comparability. TPT is better classified as zero-shot test-time adaptation. This distinction is recorded in Table 2.

6 Self-Supervised and Semi-Supervised Extensions

6.1 Self-Supervised Pre-Training for FSIC

Features are learned from unlabeled images by self-supervised methods. Supervised pre-training is complemented or replaced by these methods. Augmented image views are compared by SimCLR [60]. A momentum contrast mechanism is used by MoCo [61].

Table 2: Summary of VLM-based FSIC adaptation methods.

Method	Tunable Params	Requires Training	Strategy
Zero-shot CLIP	None	No	Hand-crafted prompt
CoOp [21]	Context vectors	Yes	Prompt tuning
CoCoOp [57]	Context + cond.	Yes	Input-cond. prompt
CLIP-Adapter [22]	Adapter MLP	Yes	Feature residual
TPT [59]	Context vectors	At test-time	Test-time prompt
CuPL [58]	None	No	LLM-generated prompts
SVL-Adapter	Adapter + prompts	Yes	Combined

Competitive miniImageNet features are learned without labels [62]. Nearest-neighbor classification is then applied. Domains with scarce base labels are supported by this result.

SSL accuracy is improved under matched pre-training domains. Self-supervised transfer is favored under some domain shifts [63]. CLIP features usually remain stronger.

6.2 Semi-Supervised Few-Shot Learning

Unlabeled novel-class images $\mathcal{U}$ are added in semi-supervised FSIC [27]. Class prototypes are refined through soft assignments:

$$\tilde{\mathbf{p}}_c = \frac{\sum_{(x_i^s, y_i^s) \in \mathcal{S}: y_i^s = c} \phi(x_i^s) + \sum_{u \in \mathcal{U}} w_{uc} \phi(u)}{|\{i : y_i^s = c\}| + \sum_{u \in \mathcal{U}} w_{uc}}, \quad (19)$$

Here, w_{uc} is the soft class assignment. Three to five EM iterations are typically required. Gains of 2–5% are reported [27].

7 Benchmark Datasets and Evaluation Protocol

7.1 Standard Datasets

Major FSIC datasets are listed in Table 3.

Table 3: Standard FSIC benchmark datasets.

Dataset	Classes	Train/Val/Test	Images/class	Domain
miniImageNet [13]	100	64/16/20	600	Natural
tieredImageNet [27]	608	351/97/160	≈ 1281	Natural
CIFAR-FS [64]	100	64/16/20	600	Natural
FC100 [65]	100	60/20/20	600	Natural
CUB-200 [66]	200	100/50/50	≈ 60	Fine-grained
Cars [67]	196	130/17/49	≈ 80	Fine-grained

miniImageNet contains 100 ImageNet classes [13]. Six hundred resized images are contained in each class. A 64/16/20 class split is standardized [68].

tieredImageNet uses WordNet for semantic class separation [27]. Thirty-four high-level categories are spanned by its 608 classes. More difficult generalization is therefore tested.

CIFAR-FS is derived from CIFAR-100 [64]. A 64/16/20 class split is used. **FC100** uses hierarchical groups [65]. Its split is more difficult. 32×32 images are contained in both datasets.

Table 4: MiniImageNet 5-way accuracy (%). Best within-group results are bold. CLIP uses external pre-training [15]. TPT uses no labeled support set [59].

Method	Backbone	Type	miniImageNet	
			1-shot	5-shot
MatchingNet [13]	Conv4	Metric	43.56 $\pm$ 0.84	55.31 $\pm$ 0.73
ProtoNet [20]	Conv4	Metric	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66
RelationNet [38]	Conv4	Metric	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70
MAML [14]	Conv4	Optim.	48.70 $\pm$ 1.84	63.11 $\pm$ 0.92
Reptile [46]	Conv4	Optim.	49.97 $\pm$ 0.32	65.99 $\pm$ 0.58
Baseline++ [10]	Conv4	Transfer	51.87 $\pm$ 0.77	75.68 $\pm$ 0.63
FEAT [39]	ResNet12	Metric	66.78 $\pm$ 0.20	82.05 $\pm$ 0.14
DeepEMD [40]	ResNet12	Metric	65.91 $\pm$ 0.82	82.41 $\pm$ 0.56
RFS-simple [12]	ResNet12	Transfer	62.02 $\pm$ 0.63	79.64 $\pm$ 0.44
RFS-distill [12]	ResNet12	Transfer	64.82 $\pm$ 0.60	82.14 $\pm$ 0.43
MELR [69]	ResNet12	Metric	67.40 $\pm$ 0.40	83.40 $\pm$ 0.28
protoLP [70]	ResNet12	Metric	68.30 $\pm$ 0.34	84.12 $\pm$ 0.24
CoOp [21]	ViT-B/16	VLM	71.51	84.43
CLIP-Adapter [22]	ViT-B/16	VLM	72.80	85.60
TPT [59]	ViT-B/16	VLM	75.44 $\pm$ 0.11	88.13 $\pm$ 0.12

Cross-domain benchmarks. Ten diverse datasets are combined within Meta-Dataset [26]. Cross-domain transfer is tested across distinct visual domains. Different results are often reported for Meta-Dataset and miniImageNet. Single-benchmark conclusions are therefore limited.

7.2 Evaluation Protocol

Six hundred random test episodes are used for standard evaluation. Mean accuracy and 95% confidence intervals are reported. Conv4 and ResNet12 are common classical backbones. ViT-B/16 is commonly used for VLM methods.

Transductive and inductive evaluation. Each query is classified independently by inductive methods. Statistics from all queries are used by transductive methods. Higher accuracy is often reported under transductive inference. The two settings must be distinguished.

Cross-method comparison. Reported results are taken from original studies. A shared experimental protocol is not used. Substantial differences are found across experimental components. Different inference settings are also used across studies. Algorithmic superiority cannot be inferred from unmatched results. Results are grouped by backbone when possible. Controlled reimplementations remain necessary.

8 Quantitative Results

8.1 Performance on miniImageNet

Representative miniImageNet results are shown in Table 4. Methods are grouped by backbone and pre-training regime. Conv4, ResNet12, and ViT-B/16 are not directly comparable. Large differences are found in capacities and pre-training datasets. Within-group comparisons are therefore preferred. The K -shot support set is not used by TPT [59]. Zero-shot test-time adaptation is reflected by its accuracy.

Table 5: 5-way classification accuracy (%) on tieredImageNet.

Method	Backbone	1-shot	5-shot
ProtoNet [20]	Conv4	53.31 $\pm$ 0.89	72.69 $\pm$ 0.74
MAML [14]	Conv4	51.67 $\pm$ 1.81	70.30 $\pm$ 1.75
FEAT [39]	ResNet12	70.80 $\pm$ 0.23	84.79 $\pm$ 0.16
RFS-distill [12]	ResNet12	71.52 $\pm$ 0.69	86.03 $\pm$ 0.49
DeepEMD [40]	ResNet12	71.16 $\pm$ 0.87	86.03 $\pm$ 0.58

Table 6: 5-way classification accuracy (%) on CIFAR-FS and FC100. Best results per backbone group in **bold**.

Method	Backbone	CIFAR-FS		FC100	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [20]	Conv4	55.50 $\pm$ 0.70	72.00 $\pm$ 0.60	35.30 $\pm$ 0.60	48.60 $\pm$ 0.60
MAML [14]	Conv4	58.90 $\pm$ 1.90	71.50 $\pm$ 1.00	38.00 $\pm$ 1.20	52.00 $\pm$ 0.90
Baseline++ [10]	Conv4	62.00 $\pm$ 0.60	79.00 $\pm$ 0.60	40.10 $\pm$ 0.60	55.40 $\pm$ 0.70
FEAT [39]	ResNet12	68.87 $\pm$ 0.22	82.90 $\pm$ 0.15	41.10 $\pm$ 0.15	55.00 $\pm$ 0.19
RFS-distill [12]	ResNet12	73.90 $\pm$ 0.80	86.90 $\pm$ 0.50	44.60 $\pm$ 0.70	60.90 $\pm$ 0.60
DeepEMD [40]	ResNet12	77.34 $\pm$ 0.82	88.40 $\pm$ 0.43	46.47 $\pm$ 0.78	63.22 $\pm$ 0.71

8.2 Performance on tieredImageNet

Selected tieredImageNet results are shown in Table 5. Only Conv4 and ResNet12 methods are included. TieredImageNet base classes are used by these methods. Verified VLM results are not widely reported. Unverified values are therefore excluded. Standard VLM evaluation remains an open reproducibility need.

8.3 Performance on CIFAR-FS and FC100

CIFAR-FS and FC100 results are shown in Table 6. Higher accuracy is usually reported for CIFAR-FS than miniImageNet. A difficult hierarchical split is used by FC100. Only Conv4 and ResNet12 results are included. Verified VLM results remain unavailable under this protocol. Unverified or extrapolated values are excluded.

8.4 Performance Trend Analysis

MiniImageNet trends from 2016–2024 are shown in Fig. 6. Sources are listed in Table 7. Different backbone capacities are used across the plotted years. Pre-training scale is also increased sharply. These factors are confounded with method changes. The largest rise is observed after CLIP-scale pre-training. Thus, isolated algorithmic progress is not measured.

8.5 Evaluation Metrics

Mean episodic classification accuracy is the standard metric:

$$\text{Acc} = \frac{1}{T} \sum_{t=1}^T \frac{1}{N \times Q} \sum_{(x,y) \in \mathcal{Q}^{(t)}} \mathbf{1}[f_{\theta}(x \mid \mathcal{S}^{(t)}) = y], \quad (20)$$

The 95% confidence interval is provided by a normal approximation. Other metrics include:

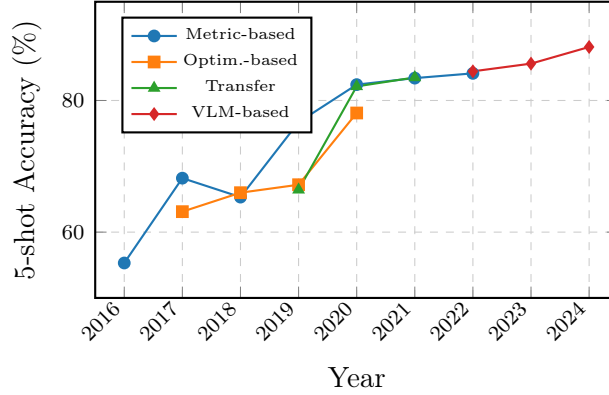


Figure 6: Reported miniImageNet 5-shot accuracy from 2016–2024. Backbone and pre-training changes are included.

Table 7: Sources for miniImageNet trend points in Fig. 6. Original studies are cited.

Family	Year	Method	5-shot Acc. (%)
Metric-based	2016	MatchingNet [13]	55.31 (Table 4)
Metric-based	2017	ProtoNet [20]	68.20 (Table 4)
Metric-based	2018	RelationNet [38]	65.32 (Table 4)
Metric-based	2019	TADAM [65]	76.70
Metric-based	2020	DeepEMD [40]	82.41 (Table 4)
Metric-based	2021	MELR [69]	83.40 (Table 4)
Metric-based	2022	protoLP [70]	84.12 (Table 4)
Optim.-based	2017	MAML [14]	63.11 (Table 4)
Optim.-based	2018	Reptile [46]	65.99 (Table 4)
Optim.-based	2019	iMAML [47]	67.20
Optim.-based	2020	MAML++ (see [48])	78.10
Transfer	2019	Baseline++ [10]	75.68 (Table 4)
Transfer	2020	RFS-distill [12]	82.14 (Table 4)
Transfer	2021	SKD (see [49])	83.50
VLM-based	2022	CoOp [21]	84.43 (Table 4)
VLM-based	2023	CLIP-Adapter [22]	85.60 (Table 4)
VLM-based	2024	TPT [59]	88.13 (Table 4)

- **Top-5 accuracy:** correct classes among five highest predictions.
- **Cross-domain accuracy:** accuracy on unseen target domains.
- **Relative improvement:** gain over zero-shot CLIP.

9 Computational Efficiency Analysis

9.1 Training Complexity

Computational cost is compared in Table 8. Costly second-order meta-gradients are required by MAML. Let S denote the inner-loop step count. Its cost is scaled roughly linearly with S . A large differentiation cost is included in each step. Each step’s batch size is changed by shot count K . The second-order step count is not determined by K . Per-episode cost

is approximated by $O(SC_{2\text{nd-order}})$. First-order approximation is used by Reptile. Implicit differentiation is used by iMAML.

Table 8: Qualitative computational complexity. “Train” denotes training. “Adapt” denotes episodic inference.

Method Family	Train Cost	Adapt Cost	Params Tuned
Metric-based	Medium	Low	Full model
Model-based	High	Low	Full model
Optimization-based	Very High	Medium	Full model
Transfer (non-hybrid)	Low	Very Low	None (frozen)
Transfer (hybrid)	Medium	Low	Classifier
VLM (prompt)	Low	Low	Prompts only
VLM (adapter)	Low	Low	Adapter MLP
VLM (training-free)	None	Very Low	None

9.2 Inference-Time Efficiency

Inference cost is important on resource-constrained devices. About 149 million parameters are contained in ViT-B/16 CLIP. No episode-specific gradient is required by many VLM methods. Embedding and classification stages are included in ProtoNet inference. Embedding cost is expressed as $O((NK + NQ)C_\phi)$. Prototype construction cost is expressed as $O(NKd)$. Query comparison cost is expressed as $O(NQd)$. These smaller costs are usually dominated by backbone computation. Fixed-backbone prototype and VLM methods have similar cost patterns.

Compact students can be learned through VLM distillation [71]. Ten times fewer parameters are used by some students. Accuracy within three percent of CLIP is reported.

10 Few-Shot Class-Incremental Learning

FSIC is extended to sequential class arrival by FSCIL. Few labeled examples are supplied in each new session. Previously learned classes must be retained. The stability–plasticity dilemma is created by this requirement. Forgetting is caused by excessive plasticity. Novel-class learning is restricted by excessive stability.

10.1 Problem Setup

T sequential sessions are processed by FSCIL. Many examples from C_0 classes are contained in the base session. N classes are introduced in each later session. Only K support images are provided per class. Evaluation includes every class observed so far:

$$A_t = \frac{1}{|\bigcup_{s=0}^t \mathcal{C}^s|} \sum_{c \in \bigcup_{s=0}^t \mathcal{C}^s} \text{Acc}(c), \quad (21)$$

Here, $\text{Acc}(c)$ denotes test accuracy for class c .

The forgetting metric. Backward transfer measures retention [72]:

$$\text{BWT} = \frac{1}{T} \sum_{t=1}^T (A_t^{(0)} - A_T^{(0)}), \quad (22)$$

Here, $A_t^{(0)}$ denotes base-class accuracy after session t . Forgetting is indicated by negative BWT. Perfect retention is indicated by zero BWT.

10.2 Key Strategies

Backbone freezing. The feature extractor is fixed after base training. Only the classifier is updated later. Forgetting is reduced, but adaptation is restricted. Context is propagated between class prototypes by CEC [73].

Regularization-based methods. Updates are constrained by TOPIC’s neural gas regularizer [74]. Large deviation from base parameters is prevented. Important parameters are protected through EWC Fisher information [75].

Prototype-based approaches. Class prototypes are preserved and rectified by SPPR. New-session semantics are used for this correction. Labels are propagated between old and new prototypes by protoLP. Strong CUB-200 and CIFAR-100 results are reported.

Architecture-based approaches. New capacity is added through the GKEAL knowledge ensemble [76]. Existing parameters are preserved. Parameter regularization is replaced by architectural expansion. The FSCIL setup is shown in Fig. 7.

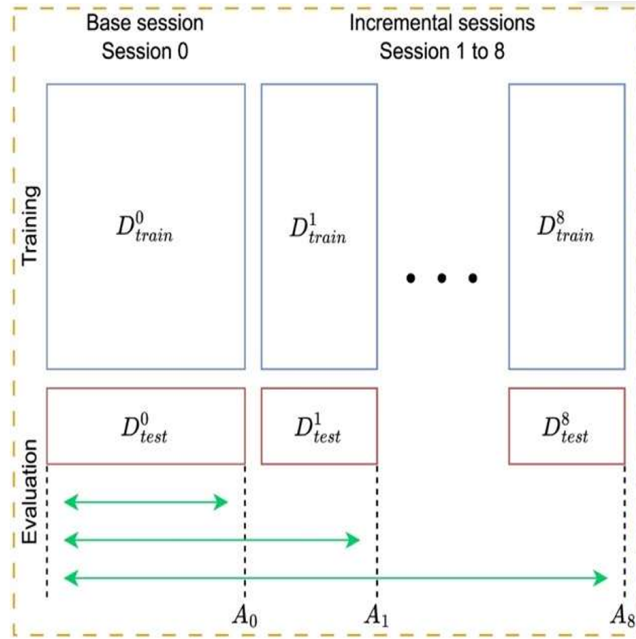


Figure 7: FSCIL sessions. Accuracy A_i covers all classes observed through session i .

10.3 FSCIL Benchmarks and Results

MiniImageNet is included among common FSCIL benchmarks [13]. CUB-200 is also widely used [66]. Another standard benchmark is provided by CIFAR-100. Eight or nine incremental sessions are usually evaluated. Final-session gaps of 15–25% remain on miniImageNet. Catastrophic forgetting therefore remains severe.

11 Discussion

FSIC accuracy is strongly affected by pre training scale. Conv4 baselines report approximately 50% 1 shot accuracy, whereas CLIP based ViT methods exceed 75%. Representation quality explains much of this difference. RFS features support strong cosine classifiers [12]. Transfer accuracy increases approximately logarithmically with data scale [77]. CLIP is trained on 400 million image text pairs [15], whereas ImageNet 1K contains approximately 1.2 million images. CLIP therefore benefits from substantially greater pre training scale. Episodic training matches training tasks with test tasks. However, regularized mini batch training remains competitive [10, 12, 11]. Limited backbones benefit most from episodic training, while its importance decreases with strong pre trained features. An episodic bias remains useful [12], but backbone quality has a larger effect. Future episodic training may benefit from hard task sampling. Metric design has shifted from fixed distances toward learned similarity. CLIP learns discriminative similarity during large scale pre training, which reduces task specific metric gains. Future metrics should adapt CLIP features to target domains. Generalization remains limited because most benchmarks share domains with training data. Cross domain tests report absolute losses of 15–25% [10, 26]. No single method dominates across all Meta Dataset domains. Domain adaptive methods are therefore still required. Adaptation gains decrease when strong calibrated representations are used. This pattern motivates the representation bottleneck hypothesis. FSIC may therefore be limited mainly by feature quality. However, current evidence comes from unmatched experimental studies. Different backbones and datasets are used across reported results. Controlled comparisons are therefore required, with fixed backbones and pre training data. Only adaptation methods should be varied in such experiments. The representation bottleneck hypothesis therefore remains provisional.

12 Challenges

Real applications are affected by substantial domain shifts. Large accuracy losses are observed under these shifts. Possible solutions are provided by domain-adaptive prompts and generalized features [78, 34]. Invariant representations can also be promoted through cross-domain episodes. Long-tailed class distributions remain insufficiently studied in FSIC. Balanced sampling and robust losses are therefore required. Minority classes can be supported through logit adjustment [79]. Additional support can be provided by class-conditional synthesis. Compositional and fine-grained recognition also remain difficult. Class attributes can be separated from identity by structured embeddings [80]. Discriminative regions can be identified through part-aware attention. Multimodal FSIC remains relatively underexplored. Additional information can be provided by depth, audio, language, and point clouds. Visual support examples can be extended through textual descriptions. Text-based recognition can be supported by large vision-language models. Detailed descriptions can be generated with GPT-4 [81]. FSIC principles can also be extended to detection and segmentation. Spatial localization must then be considered alongside recognition. Dense feature matching and spatial attention provide possible solutions [40]. Large performance gaps from full supervision remain. Efficient adaptation also requires further investigation. Compact student models can be produced through knowledge distillation [71]. Quantization and hardware-aware architecture search require further study. Evaluation protocols remain inconsistent across existing FSIC studies. Comparisons are limited by different backbones, episodes, procedures, and splits. A standardized evaluation suite is therefore required. Cross-domain, fine-grained, and long-tail tests should be included.

13 Conclusion

Five major FSIC method families are reviewed and compared. Meta-learning, pre-trained encoders, and vision-language models provide different priors. Results show that representation quality strongly affects FSIC accuracy. Higher results are often associated with larger pre-trained models. However, direct superiority among adaptation methods is not established. Simple transfer methods can match several meta-learning approaches. These findings suggest a strong representation bottleneck in FSIC. Standard benchmarks also reveal limited cross-domain robustness. Long-tail, fine-grained, and multimodal settings remain insufficiently evaluated. FSCIL results show persistent catastrophic forgetting across sessions. Stable learning and effective adaptation must therefore be achieved together. Future studies require controlled backbones, datasets, splits, and evaluation protocols. Training and inference costs should also be reported. Smaller models should be assessed for practical deployment. Standardized evaluation is required for reliable method comparison.

Appendix A. ProtoNet Derivation

A.1. Equivalence to Gaussian Mixture Models

Isotropic class-conditional Gaussians are assumed with equal variance σ^2 . Then, ProtoNet softmax equals a Gaussian-mixture posterior:

$$p(y = c \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}\|\phi(x) - \mathbf{p}_c\|^2\right), \quad (\text{A.1})$$

A probabilistic interpretation is therefore obtained.

A.2. Bregman Divergences and Metric Choice

ProtoNet is optimal for matching exponential-family distributions [20]. The corresponding distance is a Bregman divergence:

$$d_F(\mathbf{u}, \mathbf{v}) = F(\mathbf{u}) - F(\mathbf{v}) - \langle \nabla F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle, \quad (\text{A.2})$$

Here, the exponential family is defined by F . Multivariate Gaussians are associated with squared Euclidean distance.

Acknowledgments

Minor AI assistance is declared for language editing. ChatGPT and openai are used for grammar correction and text refinement. It is also used for summarization and structural clarification. All concepts, methods, and interpretations are produced by the authors. All assisted text is reviewed and verified by the authors. Full responsibility is retained by the authors.

References

- [1] B. Yousefimehr and M. Ghatee, “A systematic survey and empirical comparison of hybrid methods for imbalanced fraud detection: Combining resampling and machine learning,” *AUT J. Math. Comput.*, vol. 7, no. 1, pp. 85–116, 2026, doi: 10.48550/arXiv.2004.05439

- [2] B. Yousefimehr, M. Ghathe, and R. Razavi-Far, “Multi-stage self-distillation for class-imbalanced anomaly detection,” *IEEE Trans. Emerg. Topics Comput. Intell.*, early access, 2026. doi:10.1109/TETCI.2026.3671131
- [3] M. Ghathe and S. M. Hashemi, “Some concepts of the fuzzy multicommodity flow problem and their application in fuzzy network design,” *Math. Comput. Model.*, vol. 49, no. 5–6, pp. 1030–1043, 2009, doi: 10.1016/j.mcm.2008.08.009.
- [4] Y. Zhuang, P. Liu, H. Yang, K. Zhang, Y. Wang, and Z. Pu, “Few-shot learning for novel object detection in autonomous driving,” *Commun. Transp. Res.*, vol. 5, Art. no. 100194, 2025, doi: 10.1016/j.commtr.2025.100194.
- [5] A. Dosovitskiy *et al.*, “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [6] H. Zhang, J. Xu, S. Jiang, and Z. He, “Simple semantic-aided few-shot learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 28588–28597, doi: 10.1109/CVPR52733.2024.02701.
- [7] M. Hiller, R. Ma, M. Harandi, and T. Drummond, “Rethinking generalization in few-shot classification,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 3582–3595. doi:10.48550/arXiv.2206.07267
- [8] R.-Q. Wang, F. Zhu, X.-Y. Zhang, and C.-L. Liu, “Training with scaled logits to alleviate class-level over-fitting in few-shot learning,” *Neurocomputing*, vol. 522, pp. 142–151, 2023, doi: 10.1016/j.neucom.2022.12.011.
- [9] Y. Zou, S. Zhang, Y. Li, and R. Li, “Margin-based few-shot class-incremental learning with class-level overfitting mitigation,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 27267–27279. doi: 10.48550/arXiv.2210.04524
- [10] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, “A closer look at few-shot classification,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019. arXiv:1904.04232
- [11] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, “SimpleShot: Revisiting nearest-neighbor classification for few-shot learning,” 2019, arXiv:1911.04623.
- [12] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, “Rethinking few-shot image classification: A good embedding is all you need?” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 266–282, doi: 10.1007/978-3-030-58568-6_16.
- [13] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, 2016, pp. 3630–3638. arXiv:1606.04080
- [14] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1126–1135. arXiv:1703.03400
- [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 8748–8763. arXiv:2103.00020
- [16] C. Medina, A. Devos, and M. Grossglauser, “Self-supervised prototypical transfer learning for few-shot classification,” 2020, arXiv:2006.11325.
- [17] Z. Huang, Z. Yang, H. Hu, and Z. Gan, “Few-shot image classification via contrastive self-supervised learning,” 2020, arXiv:2008.09942.
- [18] S. Sheynin, S. Benaim, and L. Wolf, “A hierarchical transformation-discriminating generative model for few shot anomaly detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8495–8504, doi: 10.1109/ICCV48922.2021.00838.
- [19] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang, “Learning to propagate labels: Transductive propagation network for few-shot learning,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [20] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4077–4087. arXiv:1703.05175
- [21] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022, doi: 10.1007/s11263-022-01653-1.
- [22] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H. & Qiao, Y., “CLIP-Adapter: Better vision-language models with feature adapters,” *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, 2024, doi: 10.1007/s11263-023-01891-x.

- [23] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” *ACM Comput. Surv.*, vol. 53, no. 3, Art. no. 63, 2020, doi: 10.1145/3386252.
- [24] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, “Meta-learning in neural networks: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, 2022, doi: 10.1109/TPAMI.2021.3079209.
- [25] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, “Few-shot image recognition with knowledge transfer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 441–449, doi: 10.1109/ICCV.2019.00053.
- [26] E. Triantafillou *et al.*, “Meta-Dataset: A dataset of datasets for learning to learn from few examples,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. arXiv:1903.03096
- [27] M. Ren *et al.*, “Meta-learning for semi-supervised few-shot classification,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018. arXiv:1803.00676
- [28] A. Pentina and C. H. Lampert, “A PAC-Bayesian bound for lifelong learning,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 32, 2014, pp. 991–999. arXiv:1311.2838
- [29] R. Amit and R. Meir, “Meta-learning by adjusting priors based on extended PAC-Bayes theory,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 80, 2018, pp. 205–214. arXiv:1711.01244
- [30] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA, USA: MIT Press, 1961.
- [31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006, doi: 10.1002/047174882X.
- [32] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, “Meta-Baseline: Exploring simple meta-learning for few-shot learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9062–9071, doi: 10.1109/ICCV48922.2021.00893.
- [33] B. Schölkopf *et al.*, “Toward causal representation learning,” *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, 2021, doi: 10.1109/JPROC.2021.3058954.
- [34] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, “Cross-domain few-shot classification via learned feature-wise transformation,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. arXiv:2001.08735
- [35] H. Yao *et al.*, “Improving generalization in meta-learning via task augmentation,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 11887–11897. arXiv:2007.13040
- [36] S. He, R. Gheorghiu, J. E. Solem, and H. Krim, “Diffusion-based data augmentation for few-shot image classification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.
- [37] S. Thrun and L. Pratt, “Learning to learn: Introduction and overview,” in *Learning to Learn*, S. Thrun and L. Pratt, Eds. Boston, MA, USA: Springer, 1998, pp. 3–17, doi: 10.1007/978-1-4615-5529-2_1.
- [38] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1199–1208, doi: 10.1109/CVPR.2018.00131.
- [39] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, “Few-shot learning via embedding adaptation with set-to-set functions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8808–8817, doi: 10.1109/CVPR42600.2020.00883.
- [40] C. Zhang, Y. Cai, G. Lin, and C. Shen, “DeepEMD: Few-shot image classification with differentiable earth mover’s distance and structured classifiers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 12203–12213, doi: 10.1109/CVPR42600.2020.01222.
- [41] J. He, R. Hong, X. Liu, M. Xu, and Q. Sun, “Revisiting local descriptor for improved few-shot classification,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 2s, Art. no. 127, 2022, doi: 10.1145/3511917.
- [42] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 48, 2016, pp. 1842–1850.
- [43] T. Munkhdalai and H. Yu, “Meta networks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 2554–2563.
- [44] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, “A simple neural attentive meta-learner,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018. arXiv:1707.03141
- [45] Z. Li, F. Zhou, F. Chen, and H. Li, “Meta-SGD: Learning to learn quickly for few-shot learning,” 2017, arXiv:1707.09835.
- [46] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” 2018, arXiv:1803.02999.

- [47] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine, “Meta-learning with implicit gradients,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 113–124. arXiv:1909.04630
- [48] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, “Rapid learning or feature reuse? Towards understanding the effectiveness of MAML,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. arXiv:1909.09157
- [49] J. Y. Lim, K. M. Lim, C. P. Lee, and Y. X. Tan, “SCL: Self-supervised contrastive learning for few-shot image classification,” *Neural Netw.*, vol. 165, pp. 19–30, 2023, doi: 10.1016/j.neunet.2023.05.037.
- [50] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, “A baseline for few-shot image classification,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020. arXiv:1909.02729
- [51] R. Zhang, T. Che, Z. Ghahramani, Y. Bengio, and Y. Song, “MetaGAN: An adversarial approach to few-shot learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 2365–2374.
- [52] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, “Low-shot learning from imaginary data,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7278–7286, doi: 10.1109/CVPR.2018.00760.
- [53] Y. Zhu, W. Min, and S. Jiang, “Attribute-guided feature learning for few-shot image recognition,” *IEEE Trans. Multimedia*, vol. 23, pp. 1200–1209, 2021, doi: 10.1109/TMM.2020.2993952.
- [54] Z. Wang, H. Li, Z. Zhang, F. Chen, and J. Zhai, “Attribute- and attention-guided few-shot classification,” *Multimedia Syst.*, vol. 30, no. 1, Art. no. 60, 2024. DOI:10.1007/s00530-023-01236-1
- [55] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 26, 2013, pp. 3111–3119. arXiv:1310.4546
- [56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [57] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Conditional prompt learning for vision-language models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16816–16825, doi: 10.1109/CVPR52688.2022.01631.
- [58] S. Pratt, I. Covert, R. Liu, and A. Farhadi, “What does a platypus look like? Generating customized prompts for zero-shot image classification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 15691–15701, doi: 10.1109/ICCV51070.2023.01438.
- [59] M. Shu *et al.*, “Test-time prompt tuning for zero-shot generalization in vision-language models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 14274–14289. arXiv:2209.07511
- [60] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 119, 2020, pp. 1597–1607. arXiv:2002.05709
- [61] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9729–9738, doi: 10.1109/CVPR42600.2020.00975.
- [62] J.-C. Su, S. Maji, and B. Hariharan, “When does self-supervision improve few-shot learning?” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 645–666, doi: 10.1007/978-3-030-58571-6_38.
- [63] M. M. R. Siddiquee *et al.*, “Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 191–200, doi: 10.1109/ICCV.2019.00028.
- [64] L. Bertinetto, J. F. Henriques, P. H. S. Torr, and A. Vedaldi, “Meta-learning with differentiable closed-form solvers,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019. arXiv:1805.08136
- [65] B. N. Oreshkin, P. Rodríguez López, and A. Lacoste, “TADAM: Task dependent adaptive metric for improved few-shot learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018. arXiv:1805.10123
- [66] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 dataset,” California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
- [67] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2013, pp. 554–561, doi: 10.1109/ICCVW.2013.77.
- [68] S. Ravi and H. Larochelle, “Optimization as a model for few-shot learning,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.
- [69] N. Fei, Z. Lu, T. Xiang, and S. Huang, “MELR: Meta-learning via modeling episode-level relationships for few-shot learning,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

- [70] H. Zhu and P. Koniusz, “Transductive few-shot learning with prototype-based label propagation by iterative graph refinement,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23996–24006, doi: 10.1109/CVPR52729.2023.02298.
- [71] Z. Li, S. Zhang, and W. Ouyang, “Knowledge distillation from large language models for few-shot image classification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.
- [72] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, “Efficient lifelong learning with A-GEM,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019. arXiv:1812.00420
- [73] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, “Few-shot incremental learning with continually evolved classifiers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 12455–12464, doi: 10.1109/CVPR46437.2021.01227.
- [74] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, “Few-shot class-incremental learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 12183–12192, doi: 10.1109/CVPR42600.2020.01220.
- [75] J. Kirkpatrick *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017, doi: 10.1073/pnas.1611835114.
- [76] P. Zhu, Y. Liu, and J. Gu, “Ease: Unsupervised Discriminant Subspace Learning for Transductive Few-Shot Learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 9078–9088.
- [77] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, “Do ImageNet classifiers generalize to ImageNet?” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 97, 2019, pp. 5389–5400. arXiv:1902.10811
- [78] C. Ge *et al.*, “Domain adaptation via prompt learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1160–1170, 2025, doi: 10.1109/TNNLS.2023.3327962.
- [79] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. arXiv:2007.07314
- [80] P. Tokmakov, Y.-X. Wang, and M. Hebert, “Learning compositional representations for few-shot recognition,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6372–6381, doi: 10.1109/ICCV.2019.00647.
- [81] J. Achiam *et al.* (OpenAI), “GPT-4 technical report,” 2023, arXiv:2303.08774.

A. Daneshpour,
 Department of Mathematics and Computer
 Science,
 Amirkabir University of Technology,
 Tehran, Iran,
 Email: alidaneshpour@aut.ac.ir

Stefan Bucz,
 Faculty of Economics and Informatics,
 J. Selye University, Slovakia.

Damira Jantassova,
 Karaganda Technical University,
 Karaganda, Kazakhstan.

Ramis Akhmedov,
 Obuda University, Budapest, Hungary,
 SDU University, Almaty, Kazakhstan.

Received 06.11.2025, Accepted 09.03.2026, Available online 30.09.2026