Chapter 3

Few-Shot Learning: Conceptual Framework, Methodological Developments, and Security Dimensions

Sara Naghib Zadeh[1]

Hatice Nur Gök[2]

Abstract

The emergence of deep learning has largely been founded on the assumption of access to large-scale labeled datasets. However, in many real-world domains—ranging from medical imaging and the preservation of low-resource language families to space exploration and financial fraud detection—data are not always abundant or readily available. Annotation costs, privacy constraints, and the inherent rarity of certain phenomena constitute structural barriers to the practical deployment of machine learning systems. In this context, efforts to endow machines with the ability to "generalize from a few examples," a fundamental characteristic of human cognition, have given rise to the field of Few-Shot Learning (FSL).

This review article reconstructs the FSL literature around three principal paradigms: model fine-tuning, data augmentation, and transfer learning. Within the transfer learning framework, meta-learning mechanisms are examined in depth, and metric-based, optimization-based, and model-based approaches are comparatively analyzed. Furthermore, in light of the security threats that have recently emerged in this domain, feature-level adversarial attacks (FAMF) against metric-based models are comprehensively evaluated for the first time within this framework. The attack mechanisms, empirical findings, and defensive strategies are critically discussed. By identifying the structural challenges facing this field and outlining future research directions, this study argues that FSL should be regarded not merely as a technical subfield, but as a strategic turning point for the reliability and robustness of artificial intelligence systems.

1 Dr. Lecture, Halic University, Vocational School, Department of Computer Programming, ORCID: 0009-0005-6959-1165.

2 Halic University, Vocational School, Department of Big Data Analytics, ORCID:0009-0002-4248-6793

https://doi.org/10.58830/ozgur.pub1351.c5536

1. Introduction: The Problem of Data Scarcity in the Age of Data

Over the past decade, artificial intelligence research has been profoundly shaped by the deep learning revolution, which enabled the learning of hierarchical and multilayered representations. Landmark achievements in ImageNet-based image classification, breakthroughs in machine translation, and advances in speech modeling have all relied on massive quantities of labeled data and, consequently, substantial computational resources. However, the implicit assumption underlying this paradigm—that data are abundant and inexpensive—is increasingly being challenged (Parnami & Lee, 2021).

In many domains, data collection is an expensive and expertise-intensive process. For example, annotating histopathological images for pathological diagnosis requires hours of work by trained medical specialists. In the case of rare diseases, only a limited number of cases may exist worldwide. Similarly, the number of native speakers available for documenting indigenous languages is inherently restricted, and in space exploration, each collected sample may cost millions of dollars. In such scenarios, the recommendation to simply "collect more data" is often impractical and, in some cases, ethically unacceptable (Zhao et al., 2020).

At this point, human cognition serves as a major source of inspiration. A child can learn the concept of a "dog" after observing only a few examples, while an adult can recognize an unfamiliar individual in a crowded environment after seeing only a handful of photographs. This capability stems not merely from raw computational power, but from the transfer and integration of prior experiences into new situations. A person who knows how to ride a bicycle, for instance, can learn to ride a motorcycle more quickly because foundational skills such as balance and coordination are transferable (Parnami & Lee, 2021).

Few-Shot Learning (FSL) is precisely the field of research that seeks to systematically equip machines with this human-like capability for generalization. Formally, FSL refers to the ability of a model to correctly classify previously unseen classes in tasks where only a very limited number of training examples are available (typically $K = 1, 5,$ or 10) (Zhao et al., 2025). Learning from a single example is referred to as *One-Shot Learning*, while learning without any training examples and solely through auxiliary semantic information is known as Zero-Shot Learning (Pourpanah et al., 2022).

The contribution of this article is multifaceted. First, it reconstructs the existing FSL literature by organizing current approaches into three principal categories: model fine-tuning, data augmentation, and transfer learning, thereby providing a comprehensive conceptual map of the field. Second,

within the transfer learning framework, it comparatively analyzes meta-learning mechanisms from metric-based, optimization-based, and model-based perspectives. Third, for the first time in a review study of this scope, the security dimension of FSL is placed at the center of discussion through a detailed examination of Feature-level Adversarial Attacks on Metric-based Few-Shot Learning (FAMF). Finally, by identifying the structural challenges facing this field and outlining future research directions, the study proposes practical insights and recommendations for real-world applications (Zhao et al., 2025).

2.Theoretical Foundations of Few-Shot Learning

2.1. Formulation of the Few-Shot Classification Problem

In standard supervised learning, the objective is to train a model on a large dataset D in order to solve a specific task T. However, in the Few-Shot Learning (FSL) setting, the model is confronted with tasks for which the available training data are extremely limited. In the scientific literature, this problem is commonly formalized using the M -way K -shot framework:

M (number of classes): represents the number of novel categories that the model must distinguish.

K (number of examples): denotes the number of training samples available for each class during the learning phase.

Within this framework, the dataset is divided into two essential components:

Support Set: consists of K examples for each of the M classes, which the model uses for temporary adaptation and learning.

Query Set: contains unseen samples from the same classes, which the model must classify based on the knowledge acquired from the support set.

The fundamental distinction in FSL is that the model is trained on a set of "base" or "training" classes, yet during inference it is expected to recognize entirely new and previously unseen classes using only the limited information provided in the support set (Parnami & Lee, 2021).

2.2. Meta-Learning: The Mechanics of "Learning to Learn"

If classical deep learning can be described as the process of learning a specific skill, meta-learning may be understood as the process of learning how to learn. This concept, often regarded as the core driving mechanism of Few-Shot Learning, is inspired by the adaptive nature of human cognition.

2.2.1. Operational Principles of Meta-Learning

The primary objective of meta-learning is to expose the model to a wide variety of tasks rather than a single fixed problem. By solving thousands of small-scale tasks during the meta-training phase, the model learns which representations are generally transferable and how it can rapidly adapt to a new task with minimal parameter updates.

From a mathematical perspective, the goal is to identify optimal parameters θ that perform well not on a single task, but across a distribution of tasks $p(T)$:

$$\theta^* = arg\,min_\theta\, E_{\{T \sim p(T)\}}\left[L\left(D_{test}; f\left(D_{train}; \theta\right)\right)\right] \qquad (1)$$

In this formulation 1, L denotes the loss function evaluated on the test data D_{test}, while the model is allowed to adapt using only a very limited amount of training data D_{train} (Parnami & Lee, 2021).

Unlike conventional transfer learning, which primarily transfers knowledge from one domain to another, meta-learning aims to learn an adaptation strategy itself. This enables the model to achieve significantly higher robustness and adaptability when confronted with new tasks, even under conditions of limited labeled data or large-scale unlabeled environments, such as anomalous traffic patterns in distributed network systems.

2.2.2. Meta-Learning Taxonomy

Meta-learning approaches in Few-Shot Learning can be systematically categorized into three principal families based on the underlying mechanism of knowledge adaptation: optimization-based, model-based, and metric-based methods. Each family addresses the challenge of rapid adaptation from a distinct perspective, and their comparative characteristics are essential for understanding the methodological landscape of FSL.

Optimization-based meta-learning, exemplified by Model-Agnostic Meta-Learning (MAML) and its variants, seeks to learn an initialization of model parameters that can be fine-tuned with minimal gradient steps on new tasks. The core assumption is that a good initial parameter configuration exists in the loss landscape, from which rapid convergence to task-specific optima is possible (Finn et al., 2017).

Model-based meta-learning employs specialized architectures, such as memory-augmented neural networks or recurrent meta-learners, that explicitly store and retrieve task-relevant information. These models typically maintain an external memory module or use attention mechanisms to dynamically

adjust their internal state based on the support set, thereby enabling rapid task adaptation without extensive gradient-based optimization (Santoro et al., 2016; Munkhdalai & Yu, 2017).

Metric-based meta-learning, which will be examined in detail in Section 3.3.1, learns an embedding space where similarity between samples can be directly measured. Rather than adapting model parameters, these methods learn a fixed feature extractor and a distance metric, enabling classification by nearest-neighbor principles in the latent space (Snell et al., 2017; Sung et al., 2018).

The selection among these paradigms depends on the trade-off between computational cost, adaptation speed, and task complexity. While optimization-based methods offer greater flexibility, they incur higher computational overhead. Model-based approaches provide fast adaptation but require complex architectural designs. Metric-based methods excel in simplicity and inference speed but may lack discriminative power under extreme data scarcity (Parnami & Lee, 2021).

2.2.3. Meta-Learning in the Context of Few-Shot Learning

The integration of meta-learning into Few-Shot Learning represents a fundamental shift from conventional supervised learning paradigms. Whereas standard deep learning assumes access to abundant labeled data for each target class, meta-learning in FSL operates at the level of task distributions, enabling generalization across classes rather than within classes.

In the FSL setting, meta-learning is operationalized through episodic training. During each episode, a task is sampled consisting of a support set and a query set from a subset of classes. The model learns to minimize the classification error on the query set after being conditioned on the support set. Through thousands of such episodes, the model acquires meta-knowledge about how to extract discriminative features and construct decision boundaries from minimal examples.

This task-level learning distinguishes meta-learning from conventional transfer learning. In transfer learning, knowledge is transferred from source to target domains through shared representations or fine-tuned parameters. In meta-learning, the transfer occurs at the level of learning algorithms themselves—the model learns how to adapt, not merely what to transfer. This distinction is particularly critical in scenarios where target tasks are highly heterogeneous or where labeled data are too scarce to support effective fine-tuning (Parnami & Lee, 2021).

Furthermore, the meta-learning framework naturally accommodates the M-way K-shot formulation of FSL. The episodic training protocol ensures that the model is explicitly optimized for the few-shot scenario, rather than being implicitly expected to generalize from limited data after standard supervised pretraining. This alignment between training and evaluation conditions constitutes one of the primary reasons for the superior empirical performance of meta-learning-based FSL methods compared to simple fine-tuning approaches.

3. Few-Shot Learning Approaches: Three Core Paradigms

Following the establishment of the theoretical framework of Few-Shot Learning (FSL), researchers have sought to develop methods capable of addressing the fundamental challenge of data scarcity in real-world environments. Unlike conventional deep learning approaches, which rely on large-scale annotated datasets, FSL requires the design of mechanisms that enable models to generalize from only a handful of training examples. In this context, the literature has gradually been organized around three main paradigms: (1) fine-tuning-based approaches, (2) data augmentation-based approaches, and (3) transfer learning and meta-learning-based approaches. Each of these paradigms addresses the data scarcity problem from a different perspective and aims to balance computational complexity, accuracy, generalization ability, and training cost (Parnami & Lee, 2021). A comparative overview of these paradigms is presented in Table 3-1.

Table 3-1. Comparison of the Main Paradigms in Few-Shot Learning (FSL)

Paradigm	Core Idea	Advantages	Limitations	Representative Studies
Fine-Tuning-Based Approaches	Transfer knowledge from pretrained models to low-data tasks by updating model parameters	Simple implementation, leverages prior knowledge, reduces need for training from scratch	Sensitive to domain shift, prone to overfitting in low-data regimes	ULMFiT (Howard & Ruder, 2018), Nakamura et al. (2019)
Data Augmentation-Based Approaches	Enrich datasets using unlabeled data, synthetic samples, or feature augmentation	Improves data diversity, enhances generalization, mitigates data scarcity	Difficulty in modeling true data distribution, limited generalization to unseen classes	Wang et al. (2016), Mehrotra et al. (2017), Chen et al. (2019)
Transfer / Meta-Learning-Based Approaches	Learn fast adaptation mechanisms across a distribution of tasks	High generalization, fast adaptation, suitable for dynamic environments	High computational complexity, difficult optimization	Finn et al. (2017), Santoro et al. (2016), Garcia & Bruna (2018)

3.1. Fine-Tuning-Based Approaches: Knowledge Transfer from Pretrained Models

Fine-tuning-based approaches represent the most traditional and direct strategy in Few-Shot Learning. The central idea is that a deep neural network pretrained on a large-scale and general-purpose dataset can transfer the representations learned in its layers to a new task with limited data. In this paradigm, the initial model is typically trained on datasets such as ImageNet, and then adapted using only a small number of samples from the target task. The underlying assumption is that early-layer representations capture general-purpose features that can be effectively reused across different domains (Parnami & Lee, 2021).

One of the most influential approaches in transfer learning for low-data settings is ULMFiT, introduced by Jeremy Howard and Sebastian Ruder (2018). Unlike earlier methods that only retrained the final classification layer, ULMFiT proposed a three-stage framework consisting of general language model pretraining, domain-specific fine-tuning, and task-specific classification training. The model also introduced layer-wise learning rates, where earlier layers are updated more conservatively and later layers more aggressively, enabling the preservation of general knowledge while adapting to new tasks. In addition, the use of slanted triangular learning rates, which initially increase and then gradually decrease during training, helps accelerate convergence and reduce overfitting, making the approach particularly effective in Few-Shot Learning scenarios with limited training data (Howard & Ruder, 2018).

Nakamura et al. (2019) further proposed a similar strategy, employing lower learning rates for low-shot classes alongside adaptive gradient-based optimization methods. The main objective was to prevent drastic parameter updates when training with scarce data. However, despite their simplicity and ease of implementation, fine-tuning-based approaches suffer from a key limitation: when there is a significant domain gap between source and target datasets, the risk of overfitting increases substantially, leading to poor generalization performance. A comparative summary of representative fine-tuning methods is presented in Table 3-2. These limitations have motivated the development of more advanced approaches based on transfer learning and meta-learning (Nakamura et al., 2019).

Table 3-2. Comparison of Representative Fine-Tuning Methods

Model / Method	Core Idea	Key Innovation	Limitation	Reference
ULMFiT	Fine-tuning a pretrained language model for low-resource tasks	Layer-wise learning rates and slanted triangular learning rates	Sensitive to domain shift	Howard & Ruder (2018)
Adaptive Fine-Tuning	Using lower learning rates for low-shot classes	Adaptive gradient-based optimization	Overfitting under domain mismatch	Nakamura et al. (2019)

3.2. Data Augmentation Approaches: Enriching the Learning Space

One of the most critical challenges in Few-Shot Learning (FSL) is the statistical scarcity arising from the limited number of training samples. When a model is exposed to only a small number of examples, it cannot adequately learn the true diversity of the underlying data distribution, which consequently reduces its generalization capability. Data augmentation approaches have been developed to address this limitation by enriching small datasets, either through generating new data or leveraging auxiliary information to expand the learning space (Parnami & Lee, 2021). A general categorization of these methods, along with their key characteristics, is presented in Table 3-3.

Table 3-3. Categorization of data augmentation methods in FSL

Subcategory	Core Idea	Representative Models / Studies	Advantages	Limitations
Unlabeled data utilization	Using unlabeled data to learn general representations	Wang et al. (2016), Boney et al. (2018)	Reduces dependence on labeled data	Sensitive to representation quality
Transductive learning	Incorporating test data structure into the learning process	TPN — Liu et al. (2019)	Improves generalization	Increased computational cost
Data synthesis	Generating synthetic samples for low-shot classes	Mehrotra et al. (2017), Hariharan & Girshick (2017), Wang et al. (2018)	Increases training diversity	Difficulty in modeling true data distribution
Feature augmentation	Enriching feature space instead of sample space	Dixit et al. (2017), Liu et al. (2018), Schwartz et al. (2018), Chen et al. (2019)	Produces richer representations	Limited interpretability of features

3.2.1. Use of Unlabeled Data

In many real-world applications, unlabeled data are significantly more abundant than labeled data. Therefore, one of the earliest strategies was to exploit unlabeled data to enhance model learning. Wang et al. (2016) proposed a transfer-based approach using convolutional neural networks, introducing a self-supervised pre-training phase. In this stage, the model learns a general and rich representation of the data space without being constrained to specific classes. This representation is subsequently used in few-shot tasks (Wang et al., 2016).

Similarly, Boney et al. (2018) introduced a semi-supervised variant of MAML, in which unlabeled data are used to refine the embedding function, while labeled data are used to train the classifier. This combination enables the model to learn more stable representations even under extreme data scarcity (Boney et al., 2018).

3.2.2. Transductive Learning

Transductive learning can be considered a subset of semi-supervised learning in which the structure of test data is incorporated into the learning process. Unlike inductive learning, which assumes test data are completely unseen, transductive learning jointly analyzes relationships between training and test samples.

In this context, Liu et al. (2019) proposed the Transductive Propagation Network (TPN). This model consists of four main stages: feature extraction, graph construction, label propagation, and loss computation. The key idea of TPN is that the geometric structure of unlabeled test data can help the model construct more accurate decision boundaries (Liu et al., 2019).

3.2.3. Data Synthesis and Generative Networks

Another important approach to data augmentation is the generation of synthetic samples for low-shot classes. Generative Adversarial Networks (GANs), due to their strong capability in modeling data distributions, have become a central tool in this domain.

Mehrotra et al. (2017) proposed a GAN-based architecture for One-Shot Learning. In this framework, the generator aims to produce synthetic samples that are close to the true data distribution (Mehrotra et al., 2017). Hariharan and Girshick (2017) introduced a two-stage approach consisting of representation learning and multi-shot classification, where synthetic data are used to improve classification performance (Hariharan & Girshick, 2017).

Furthermore, Wang et al. (2018) developed a model that integrates meta-learning with data generation to produce virtual samples for novel classes .

Despite their success, data synthesis methods still face significant challenges. Many generative models struggle to accurately capture complex data distributions, resulting in synthetic samples that may lack realistic structure. Moreover, their generalization to entirely novel classes remains difficult, and generated features often have limited interpretability (Parnami & Lee, 2021).

3.2.4. Feature Augmentation

Instead of directly generating samples, some researchers have focused on enriching the feature space. In this approach, the goal is to enable the model to learn intrinsic variations within data at the representation level.

Shu introduced the AGA model, which manipulates images based on object-level features (Shu et al., 2018). Schwartz et al. (2018) proposed the Delta Encoder, demonstrating that new features for unseen classes can be synthesized even from a few examples (Schwartz et al., 2022).

In addition, Chen et al. (2019) introduced TriNet, which establishes a bidirectional mapping between the semantic label space and the image feature space. This design allows each class to obtain a richer and more robust representation in the feature space (Chen et al., 2019).

3.3. Transfer Learning and Meta-Learning Approaches

3.3.1. Metric-Based Methods: Learning the Concept of Similarity

In many Few-Shot Learning (FSL) scenarios, the number of available samples is so limited that training a deep classifier directly becomes practically infeasible. Consequently, researchers have shifted from directly learning class labels toward learning the concept of *similarity*. Metric learning is based on the assumption that samples belonging to the same class should be close to each other in the feature space, while samples from different classes should be far apart (Parnami & Lee, 2021).

From a mathematical perspective, a metric is a function that measures the distance between two samples. In deep learning settings, commonly used metrics include Euclidean distance, Mahalanobis distance, and cosine similarity. A typical metric learning framework consists of two components: an embedding module that maps input data into a vector space, and a metric module that computes similarity or distance between embeddings (Kotovenko et al., 2023).The most important metric-based models and their characteristics are compared in Table 3-4.

Table 3-4. Comparison of metric-based learning models

Model	Core Idea	Distinct Feature	Advantages	Reference
Siamese Networks	Learning similarity between pairs of samples	Shared weights between twin networks	Suitable for one-shot learning	Koch et al. (2015)
Matching Networks	Using attention for sample comparison	LSTM + attention mechanism	High accuracy in low-shot settings	Vinyals et al. (2016)
Prototypical Networks	Defining a prototype for each class	Class-wise feature averaging	Simplicity and fast inference	Snell et al. (2017)
Relation Networks	Learning comparison function via CNN	Eliminates need for predefined metric	High flexibility	Sung et al. (2018)
CovaMNet	Using feature covariance information	Second-order statistical representation	Richer feature representation	Li et al. (2019)

Koch et al. (2015) introduced Siamese Neural Networks for one-shot recognition. These networks consist of two identical branches with shared weights and aim to minimize the distance between similar samples while maximizing the distance between dissimilar ones (Koch et al., 2015). Subsequently, Vinyals et al. (2016) proposed Matching Networks, which learn similarity using LSTM and attention mechanisms (Vinyals et al., 2016).

Snell et al. (2017) introduced Prototypical Networks, where each class is represented by a prototype (i.e., the mean of its feature embeddings), and classification is performed based on the nearest prototype in feature space (Snell et al., 2017). Later, Sung et al. (2018) proposed Relation Networks, which learn the similarity function directly using a neural network instead of relying on predefined distance metrics (Sung et al., 2018).

Additionally, Li et al. (2019) introduced CovaMNet, which leverages covariance matrices to capture second-order statistical information, resulting in richer representations compared to mean-based approaches (Li et al., 2019).

The main advantages of metric-based methods include computational simplicity, fast inference, and rapid adaptation to new tasks. However, under extremely limited data conditions, simple distance metrics may lack sufficient discriminative power, which can reduce performance in dynamic environments (Parnami & Lee, 2021).

3.3.2. Meta-Learning Methods: Learning to Learn

Meta-learning, or "learning to learn," is one of the most fundamental frameworks in Few-Shot Learning. Unlike conventional machine learning, where a model is trained for a single task, meta-learning aims to enable models to rapidly adapt to new tasks with minimal data (Parnami & Lee, 2021).

In this framework, data are divided into two levels: meta-training and meta-testing. Each task consists of a support set and a query set. The model is trained over a distribution of tasks so that it can generalize to unseen tasks with only a few examples. A comparison of key meta-learning models is provided in Table 3-5.

Table 3-5. Comparison of meta-learning methods

Model	Architecture / Core Idea	Key Advantage	Limitation	Reference
MANN	External memory with neural Turing machine	Fast learning from limited data	High architectural complexity	Santoro et al. (2016)
Meta Networks	Combination of meta-learner and base learner	Rapid weight generation	High training cost	Munkhdalai & Yu (2017)
MAML	Learning adaptable initial parameters	Model-agnostic approach	Requires second-order gradients	Finn et al. (2017)
TAML	Regularization to reduce overfitting	Better generalization	Difficult hyperparameter tuning	Jamal et al. (2019)

One of the earliest influential models in this area is Memory-Augmented Neural Networks (MANN), introduced by Santoro et al. (2016). This model integrates external memory with a neural Turing machine architecture, enabling both short-term and long-term information storage (Santoro et al., 2016).

Munkhdalai and Yu (2017) proposed Meta Networks, consisting of a base learner and a meta-learner. The meta-learner generates fast weights that allow rapid adaptation to new tasks (Munkhdalai & Yu, 2017).

Finn et al. (2017) introduced Model-Agnostic Meta-Learning (MAML), one of the most influential meta-learning approaches. The core idea is to learn an initialization of model parameters that can be quickly adapted to new tasks using only a few gradient steps:

$$\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{T_i}\left(f_\theta\right) \qquad (2)$$

The main strength of MAML lies in its model-agnostic nature, meaning it can be applied to any gradient-based model. However, its reliance on second-order derivatives increases computational cost. To address this, variants such as FOMAML and Reptile were proposed (Finn et al., 2017).

Jamal et al. (2019) further proposed TAML, which introduces a regularization term to reduce overfitting on training tasks and improve generalization performance (Jamal et al., 2019).

3.3.3. Graph Neural Network-Based Methods: Modeling Structural Relationships

Graph Neural Networks (GNNs) represent another important direction in Few-Shot Learning, designed to model structural relationships among samples. In this approach, data are represented as graphs, where each sample is a node and relationships between samples are represented as edges (Parnami & Lee, 2021).

Garcia and Bruna (2018) proposed one of the first GNN-based models for few-shot image classification. In this architecture, the model jointly learns node and edge representations and explicitly models relationships among samples (Garcia & Bruna, 2018).

Later, Kim et al. (2019) developed EGNN, which focuses on edge classification. In this model, each edge is represented by a two-dimensional feature vector indicating whether two nodes belong to the same class or different classes (Kim et al., 2019). A comparison of these models is provided in Table 3-6.

Table 3-6. Comparison of GNN-based methods in FSL

Model	Core Idea	Key Feature	Advantages	Limitation	Reference
GNN	Modeling samples as graphs	Joint node-edge learning	High interpretability	Complexity grows with graph size	Garcia & Bruna (2018)
EGNN	Focus on edge classification	Probabilistic edge representation	Captures complex relations	High computational cost	Kim et al. (2019)

The main advantage of GNN-based methods is their strong interpretability and ability to model complex dependencies among samples. However, as the

number of samples increases, the number of edges grows exponentially, leading to significant computational overhead (Parnami & Lee, 2021).

Finally, a comparative summary of the main FSL families in terms of computational complexity, generalization capability, and robustness to overfitting is presented in Table 3-7.

Table 3-7. Comparative summary of major FSL families

Method Family	Computational Complexity	Inference Speed	Overfitting Resistance	Generalization	Interpretability
Model Fine-Tuning	Medium	High	Medium	Domain-dependent	Medium
Data Augmentation	Medium–High	Medium	High	Relatively good	Low
Metric Learning	Low	Very high	Medium	Good	High
Meta-Learning	High	High	High	Very strong	Medium
Graph Neural Networks	High	Medium	High	Good	High

3.4. Integrated Conceptual Framework: Synthesis of FSL Paradigms

The three paradigms examined in the preceding sections—fine-tuning, data augmentation, and transfer/meta-learning—are not mutually exclusive methodological alternatives, but rather complementary strategies that can be integrated to address the multifaceted challenge of data scarcity. This section presents an integrated conceptual framework that synthesizes these paradigms and clarifies their interrelationships within the broader FSL ecosystem.

At the foundational level, fine-tuning-based approaches leverage pretrained representations as a starting point for adaptation. This paradigm is most effective when the source and target domains share substantial visual or semantic similarity, and when computational resources are limited. However, as the domain gap widens or data scarcity becomes more severe, fine-tuning alone proves insufficient, necessitating the incorporation of additional mechanisms.

Data augmentation operates at the sample level, enriching the training space through synthetic or transformed examples. When combined with fine-tuning, augmentation mitigates overfitting and enhances generalization. More critically, when integrated with meta-learning, augmentation can improve the diversity of episodic tasks, thereby strengthening the model's ability

to generalize across task distributions. For instance, feature augmentation strategies can be embedded within meta-learning episodes to provide richer support sets without requiring additional labeled data (Chen et al., 2019).

Transfer and meta-learning represent the highest level of abstraction in this framework. Meta-learning learns the adaptation mechanism itself, while transfer learning provides the representational substrate upon which this mechanism operates. The synergy between these approaches is evident in methods such as MAML, where pretrained initializations can accelerate meta-training, and in metric-based approaches, where transfer-learned embeddings serve as the feature space for similarity computation.

The selection of an appropriate strategy—or combination thereof—depends on several contextual factors: (i) the degree of domain shift between source and target data; (ii) the availability of unlabeled or auxiliary data; (iii) computational constraints; and (iv) the required adaptation speed. In practice, hybrid architectures that combine meta-learning with data augmentation and transfer learning have demonstrated superior performance over single-paradigm approaches, suggesting that the future of FSL lies in the principled integration of these complementary strategies rather than in their isolated application (Parnami & Lee, 2021; Song et al., 2023).

Figure 3-1 illustrates the proposed integrated framework, depicting the hierarchical relationships among the three paradigms and their potential intersections. The framework emphasizes that effective FSL systems should be designed with flexibility to incorporate multiple paradigms, adapting their composition to the specific requirements of the target application domain.

4. Experimental Evaluation and Performance Outlook in Few-Shot Learning

Experimental evaluation plays a central role in assessing the generalization capability of Few-Shot Learning (FSL) models. Unlike traditional supervised learning, where models are evaluated on large-scale data, FSL focuses on rapid adaptation to novel classes with extremely limited labeled samples. For this reason, standardized benchmark datasets such as miniImageNet and Omniglot have become widely used for fair comparison of models (Zhao et al., 2024).

Omniglot consists of handwritten characters from multiple alphabets and is considered a relatively simple benchmark due to its limited visual variability. In contrast, miniImageNet contains real-world images with significant variations in background, illumination, viewpoint, and object structure, making it a much more challenging benchmark for evaluating generalization performance (Parnami & Lee, 2021).

Most studies report results under standard evaluation settings such as 5-way 1-shot and 5-way 5-shot classification. In the 1-shot setting, the model observes only one example per class, whereas in the 5-shot setting it observes five examples per class, allowing a more robust estimation of class structure (Xian et al., 2020).

4.1. Empirical Results on miniImageNet

Table 4-1 summarizes the performance of representative FSL models on miniImageNet under 1-shot and 5-shot settings.

Table 4-1. Performance comparison of FSL models on miniImageNet (Parnami & Lee, 2021).

Model	Approach Type	1-shot Accuracy (%)	5-shot Accuracy (%)
Matching Networks	Metric learning	43.56	55.31
Prototypical Networks	Metric learning	49.42	68.20
MAML	Optimization-based meta-learning	48.70	63.11
Relation Networks	Metric learning	50.44	65.32
SimpleShot	Transfer learning	64.29	80.64

Several important observations can be drawn from these results. First, increasing the number of support samples from 1-shot to 5-shot consistently improves performance across all models. This indicates that even a small increase in labeled examples significantly enhances class representation quality. For instance, Prototypical Networks improve from 49.42% to 68.20% accuracy (Snell et al., 2017).

Studies in Few-Shot Learning consistently show a clear performance gap between simpler benchmarks such as Omniglot and more complex datasets like miniImageNet, where accuracy drops significantly due to higher visual diversity, emphasizing the importance of robust and transferable feature representations. In addition, results across methods indicate a steady improvement over time, with early approaches like Matching Networks achieving around 43% accuracy in the 1-shot setting, while later methods such as SimpleShot surpass 64%, reflecting advances in representation learning strategies. Notably, SimpleShot demonstrates that even without complex meta-learning mechanisms, strong

performance can be achieved through high-quality feature embeddings and proper normalization, sometimes outperforming more sophisticated models (Parnami & Lee, 2021).

4.2. Structural Perspective on Performance Trends

Empirical evidence shows that FSL performance is not solely determined by model complexity. Instead, it depends on multiple interacting factors, including feature representation quality, similarity metric design, task construction strategy, and domain shift between training and testing distributions.

Metric-based methods generally perform well due to their simplicity and fast inference. However, they may struggle in highly complex or noisy environments where simple distance measures are insufficient. In contrast, meta-learning approaches provide stronger adaptation capabilities but require higher computational cost and carefully designed task distributions for training (Finn et al., 2017).

Moreover, the success of models such as SimpleShot indicates that the quality of the embedding space can be more critical than architectural complexity. This observation has led to an increasing research focus on representation learning rather than purely on adaptation mechanisms (Payandeh et al., 2023).

Overall, experimental studies suggest that no single FSL approach is universally optimal. The best-performing method depends on dataset characteristics, task complexity, computational constraints, and the degree of domain shift between training and testing environments.

5. Security in Few-Shot Learning: Adversarial Attacks and Model Vulnerability

5.1. Feature-Level Adversarial Attacks and Model Vulnerability

Deep learning models are highly vulnerable to adversarial attacks, as it has been shown that even very small perturbations in the input can lead to significant decision-making errors (Goodfellow et al., 2015). More advanced methods such as PGD and Carlini–Wagner attacks further revealed and intensified this vulnerability in neural networks (Madry et al., 2018). This issue becomes even more critical in Few-Shot Learning (FSL), where models are trained with extremely limited samples, making their decision boundaries more fragile and highly sensitive to small input variations. For this reason, methods such as ADML, AQ, and DFSL have been proposed to improve robustness, although the security challenge has not yet been fully resolved (Kim et al., 2026).

In classical adversarial attacks, a small perturbation is added to the input:

$$x^{adv} = x + \delta s.t. \| \delta \| \leq \grave{o}$$

This approach is effective in image-based classification models, but it has limitations in metric-based Few-Shot models, since these models make decisions based on distances in the feature space rather than the pixel space. Therefore, effective attacks require manipulation at the feature level rather than the input level (Kim et al., 2026 ; Xu et al., 2025).

In this context, Kim et al. (2026) introduced FAMF (Feature-level Adversarial Attack on Metric-based Few-Shot Learning), the first attack specifically designed for metric-based FSL models. In this method, the perturbation is applied directly in the feature space:

$$f(x)^{adv} = f\left(x\right) + \arg\max_{\delta} L\left(\theta, f\left(x\right) + \delta, y\right) s.t. \| \delta \| \leq \grave{o}$$

The goal of this attack is to increase classification error by reducing intra-class similarity and increasing inter-class similarity in the feature space (Kim et al., 2026).

Experimental results on Omniglot and miniImageNet datasets show that FAMF is significantly more effective than traditional attacks such as PGD, achieving nearly 100% attack success rate in some models like FEAT. In the 1-shot setting, attacks are stronger than in the 5-shot setting because fewer support samples make decision boundaries more fragile. Moreover, FAMF remains effective even in the presence of defense mechanisms such as AQ, and it shows higher robustness against defensive strategies compared to image-level attacks. From a computational perspective, FAMF is also faster than PGD, since it operates directly in the feature space and does not require full backpropagation through the entire network (Kim et al., 2026 ; Xu et al., 2025).

These findings indicate that the main vulnerability of metric-based Few-Shot models lies in the feature representation space, rather than the input space. Therefore, the security of these models must be redefined beyond input-level defenses and extended toward feature-level defense mechanisms, which ensure the robustness of the latent representation space against adversarial perturbations. Additionally, FAMF can be used as a diagnostic tool to identify hidden weaknesses in models and to design more robust architectures (Zheng et al., 2025).

However, this type of attack is mainly developed under a white-box assumption, where the attacker has full access to the model architecture and embedding function. Therefore, developing black-box versions of such attacks

and designing robust defenses against feature-level adversarial strategies remain among the most important future research directions in the security of Few-Shot Learning (Cao et al., 2020 ; Kim et al., 2026).

5.2. Poisoning Attacks in Few-Shot Learning

Beyond adversarial perturbations at inference time, Few-Shot Learning models are vulnerable to poisoning attacks that compromise the integrity of the training or support data. In the FSL context, poisoning can occur at two distinct stages: during meta-training, where the attacker contaminates the base dataset used to learn the meta-parameters, or during inference, where malicious examples are injected into the support set.

Meta-training poisoning is particularly insidious because the attack is embedded in the model's learned prior. By strategically inserting mislabeled or crafted examples into the training tasks, an attacker can bias the meta-learner toward representations that fail under specific trigger conditions. For instance, a backdoored meta-model might perform normally on standard few-shot tasks but produce systematically incorrect predictions when a predefined visual pattern appears in the query image (Gu et al., 2019).

Support set poisoning targets the adaptation phase of FSL. Since metric-based and meta-learning models rely heavily on the support set for task-specific decision boundaries, even a small number of poisoned support examples can significantly degrade classification accuracy. Unlike standard supervised learning, where poisoning effects may be diluted across a large training set, FSL's extreme data scarcity amplifies the impact of each corrupted sample. Recent studies have demonstrated that injecting as few as one poisoned example per class into the support set can reduce accuracy by over 30% in prototypical networks (Shafahi et al., 2018).

The defense against poisoning attacks in FSL requires robust data validation mechanisms at both the meta-training and inference stages. Techniques such as spectral signature detection, activation clustering, and outlier removal have shown promise in identifying poisoned samples, though their adaptation to the few-shot setting remains an active area of research.

5.3. Backdoor Attacks

Backdoor attacks represent a specialized form of poisoning in which the attacker embeds a hidden trigger pattern into the model during training, causing the model to behave normally on clean inputs but produce attacker-chosen outputs when the trigger is present. In Few-Shot Learning, backdoor

attacks are especially dangerous due to the limited number of training examples, which makes it easier to introduce trigger patterns without detection.

In metric-based FSL, backdoor attacks can be implemented by manipulating the embedding space such that trigger-embedded samples from any class map to a specific region associated with a target class. During inference, any query image containing the trigger will be misclassified into the attacker's chosen category, regardless of its true class. This attack is particularly effective against prototypical networks, where the class prototype can be shifted toward the trigger-embedded region through a small number of poisoned support examples (Liu et al., 2020).

The stealthiness of backdoor attacks in FSL stems from the fact that clean task performance remains largely unaffected, making detection through standard accuracy evaluation difficult. Furthermore, the episodic training protocol of meta-learning provides additional cover for the attacker, as the trigger can be distributed across multiple tasks without appearing suspicious in any single episode.

Defensive strategies against backdoor attacks in FSL include neural cleanse techniques, which reverse-engineer potential triggers by analyzing model behavior across classes, and fine-pruning methods that remove dormant neurons associated with backdoor functionality. However, these defenses were originally designed for standard supervised learning and require significant adaptation for the meta-learning setting, where model parameters are optimized for rapid task adaptation rather than fixed classification boundaries.

5.4. Membership Inference Attacks

Membership inference attacks aim to determine whether a specific data sample was included in the training set of a machine learning model. While traditionally studied in the context of large-scale supervised learning, membership inference poses distinct challenges and risks in Few-Shot Learning due to the intimate relationship between support set composition and model predictions.

In FSL, membership inference can be targeted at two levels: the meta-training set and the task-level support set. At the meta-training level, an attacker with query access to the trained meta-model can infer whether a particular class or sample was included in the meta-training distribution. This is particularly concerning in applications involving sensitive data, such as medical diagnosis or biometric identification, where the mere presence of an individual's data in the training set may constitute a privacy violation (Shokri et al., 2017).

At the support set level, membership inference becomes even more direct. Since the support set is explicitly used to condition the model's predictions during inference, an attacker can exploit the model's confidence patterns to infer which samples were present in the support set. For example, in prototypical networks, the distance between a query sample and its corresponding class prototype tends to be smaller when the query was part of the support set, providing a discriminative signal for membership inference.

The vulnerability of FSL models to membership inference is exacerbated by their reliance on distance metrics in the embedding space. These metrics often leak information about the support set composition, particularly when the number of support examples is extremely small. Defense mechanisms such as differential privacy, which adds calibrated noise to model outputs or gradients, have been proposed to mitigate membership inference risks. However, applying differential privacy in FSL is challenging because the noise may degrade the already limited information available for task adaptation, creating a tension between privacy preservation and few-shot performance.

5.5. Model Extraction Attacks

Model extraction attacks involve an adversary with only black-box query access to a target model, attempting to construct a functionally equivalent copy of that model. In Few-Shot Learning, this threat is particularly acute because FSL models are often deployed as lightweight, specialized services where query access is provided to users for rapid task adaptation.

The extraction process in FSL differs from standard model extraction due to the episodic nature of inference. An attacker can query the target model with carefully constructed support sets and query samples, observing the predicted labels or confidence scores. By systematically exploring the input space across multiple episodes, the attacker can train a surrogate model that approximates the target model's embedding function and decision boundaries.

Metric-based FSL models are especially susceptible to extraction attacks because their decision logic is relatively simple: compute embeddings, measure distances, and select the nearest class. This simplicity enables accurate extraction with fewer queries compared to complex deep classifiers. Furthermore, because FSL models are designed for rapid adaptation, they often lack the defensive depth—such as input preprocessing pipelines or ensemble structures—that might otherwise impede extraction (Tramer et al., 2016).

The consequences of successful model extraction in FSL extend beyond intellectual property theft. An extracted surrogate model can be used to mount more effective white-box attacks, including the feature-level adversarial

attacks and backdoor injections described in previous sections. Moreover, if the original model was trained on proprietary or sensitive data, the extracted model may retain traces of that data, creating secondary privacy risks.

Defenses against model extraction in FSL include rate limiting on query access, output perturbation to obscure confidence scores, and watermarking techniques that embed identifiable signatures in model predictions. However, these defenses must be carefully calibrated to avoid degrading the core few-shot adaptation capability that makes FSL valuable in the first place.

5.6. Comprehensive Defense Strategies

The diverse attack surface of Few-Shot Learning—spanning feature-level adversarial perturbations, data poisoning, backdoor triggers, membership inference, and model extraction—necessitates a multi-layered defense strategy that addresses vulnerabilities at each stage of the FSL pipeline. This section synthesizes the defensive approaches discussed throughout Section 5 and proposes an integrated security framework for robust FSL deployment.

At the input level, adversarial training remains the most effective defense against feature-level attacks such as FAMF. By augmenting the meta-training process with adversarially perturbed support and query sets, the model learns to construct more robust decision boundaries in the embedding space. However, standard adversarial training significantly increases computational cost and may reduce clean-task accuracy, necessitating the development of efficient adversarial training variants tailored to the episodic learning paradigm (Madry et al., 2018).

At the data level, robust aggregation and outlier detection mechanisms can mitigate poisoning and backdoor attacks. For meta-training, spectral analysis of gradient updates across tasks can identify and exclude poisoned episodes. For inference-time support sets, consistency checks based on geometric properties of the embedding space—such as unexpected prototype shifts or anomalous inter-sample distances—can flag potentially corrupted support examples before they influence predictions.

At the model level, architectural modifications can enhance intrinsic robustness. Ensemble approaches that aggregate predictions from multiple meta-learners with diverse initializations reduce the impact of any single compromised component. Additionally, regularization techniques that enforce smoothness in the embedding space—such as Lipschitz continuity constraints—limit the adversary's ability to induce large changes in model output through small perturbations.

At the system level, access control and monitoring mechanisms are essential for preventing model extraction and membership inference. Query logging, anomaly detection in query patterns, and differential privacy guarantees can collectively raise the cost of attacks while preserving legitimate user functionality. The trade-off between security and usability must be carefully managed, as excessive restrictions may undermine the rapid adaptation capability that defines FSL.

The relationship between attack vectors and corresponding defenses can be summarized as follows. Feature-level adversarial attacks such as FAMF are primarily countered through adversarial training at the input level, with embedding space regularization serving as a complementary secondary defense. Poisoning attacks targeting either the meta-training set or the support set require spectral signature detection as the primary defense, supplemented by outlier removal techniques. Backdoor attacks embedded during meta-training are addressed through neural cleanse methods, with fine-pruning of dormant neurons as an additional safeguard. Membership inference attacks at the inference stage are mitigated through differential privacy mechanisms, supported by output perturbation to obscure confidence patterns. Finally, model extraction attacks at the system level are countered through rate limiting and query logging, with watermarking techniques providing secondary protection for intellectual property.

In conclusion, the security of Few-Shot Learning systems cannot be ensured through any single defensive mechanism. Rather, a defense-in-depth approach that combines input sanitization, robust training, architectural hardening, and system-level monitoring is required to address the multifaceted threat landscape. As FSL continues to be deployed in safety-critical and privacy-sensitive applications, the integration of security considerations into the core design of FSL methodologies will become not merely advisable but essential.

6. Challenges and Future Directions in Few-Shot Learning

Despite significant advances in Few-Shot Learning (FSL), the field still faces several fundamental challenges that limit its full applicability in real-world scenarios. One of the most important limitations is the reliance on the conventional **M-way K-shot** setting, where models are trained under highly controlled conditions. In real-world problems, however, the number of classes and the number of samples per class are not predefined, which reduces model flexibility and adaptability. Moreover, meta-learning approaches often assume that training and testing tasks are independently and identically distributed according to a task distribution $p(T)$. This assumption leads to significant

performance degradation under cross-domain scenarios, such as transferring from natural images to text or audio data (Parnami & Lee, 2021).

Another key challenge is the difficulty of integrating knowledge from both seen and unseen classes within a unified framework. Many existing models are designed only to classify novel classes within a fixed support set, which limits their effectiveness in real-world environments where both old and new classes must be recognized simultaneously. In addition, data heterogeneity in domains such as audio, wireless signals, and text prevents the construction of standardized datasets, which are essential for stable meta-learning. Furthermore, the vulnerability of FSL models to adversarial attacks and feature-space manipulation raises serious concerns regarding robustness and trustworthiness (Kim et al., 2026).

To address these challenges, future research directions focus on several key areas. First, the integration of structured prior knowledge, such as knowledge graphs and ontologies, may reduce reliance on large-scale pretraining data. Second, the development of task-adaptive distance metrics, which can dynamically adjust instead of relying on fixed measures such as Euclidean distance, is considered a promising direction for improving model flexibility.

In addition, advanced meta-learning architectures, particularly hierarchical models, can enhance the ability to capture meta-knowledge and better handle task heterogeneity. The combination of different learning paradigms—including transfer learning, active learning, and reinforcement learning—may also lead to more powerful hybrid systems. Furthermore, multi-modal learning approaches that enable knowledge transfer across text, image, and audio modalities, as well as progress in zero-shot learning, represent important future research directions (Da Silva & Costa, 2019).

Finally, robustness and security remain critical concerns. The development of feature-level defense mechanisms, the use of adversarial training to improve robustness, and the adaptation of certified robustness concepts to FSL settings are essential strategies for ensuring reliability in safety-critical applications (Parnami & Lee, 2021; Kim et al., 2026).

7. Conclusion

7.1. Summary and Key Contributions

Few-Shot Learning (FSL) has emerged as one of the most important research directions in machine learning, particularly in scenarios where data scarcity is a fundamental limitation. This review systematically organized the FSL literature into three main paradigms: model fine-tuning, data augmentation,

and transfer learning. Within the transfer learning paradigm, meta-learning approaches—including metric-based, optimization-based, and model-based methods—were comparatively analyzed, highlighting their strengths and limitations (Parnami & Lee, 2021).

A key contribution of this study is the emphasis on security and robustness alongside predictive performance. Results from feature-level adversarial attacks such as FAMF (Feature-level Adversarial Attack) demonstrate that while metric-based models perform well at the input level, they remain vulnerable in the feature representation space. This indicates that FSL systems should not only focus on improving accuracy but must also explicitly consider robustness, stability, and trustworthiness in their design (Kim et al., 2026).

Moreover, several structural challenges remain unresolved, including the limitations of the M-way K-shot framework, dependence on fixed task distributions, difficulty in integrating seen and unseen classes, and poor generalization to non-visual domains. These challenges suggest that FSL is still an evolving field requiring fundamental innovations. Promising future directions include the use of prior knowledge, development of adaptive distance metrics, integration of multiple learning paradigms, and design of robust models resistant to adversarial attacks (Song et al., 2023).

In conclusion, Few-Shot Learning is not merely a technical machine learning approach but a strategic research area that will play a crucial role in the future of artificial intelligence. In a world where data scarcity is a universal constraint, successful advancement of FSL can enable broader deployment of AI systems in real-world, low-data, and complex environments, ultimately contributing to more accessible and equitable intelligent technologies (Song et al., 2023).

7.2. Foundation Models and the Evolution of Few-Shot Learning

The emergence of large-scale foundation models—such as GPT-4, CLIP, DINO, and Segment Anything Model (SAM)—has fundamentally altered the landscape of Few-Shot Learning. These models, pretrained on internet-scale datasets using self-supervised or contrastive learning objectives, acquire highly generalizable representations that can be adapted to novel tasks with minimal or no task-specific training. This capability, often termed emergent few-shot performance, challenges the traditional boundaries of FSL research.

Foundation models approach the few-shot problem through a different mechanism than classical meta-learning. Rather than learning an explicit adaptation algorithm through episodic training, these models leverage the vast diversity of their pretraining data to implicitly encode a broad spectrum of

visual, linguistic, and conceptual relationships. When presented with a novel task defined by a few examples, the model can leverage these pre-encoded relationships to make accurate predictions without gradient-based fine-tuning. For instance, CLIP's joint embedding of images and text enables zero-shot and few-shot classification through natural language prompts, bypassing the need for task-specific architectures (Radford et al., 2021).

The implications of this paradigm shift for FSL are profound. First, the performance gap between foundation models and specialized meta-learning methods has narrowed considerably, with models like GPT-4 achieving competitive few-shot results across diverse domains without domain-specific architectural engineering. Second, the distinction between pretraining and adaptation is becoming increasingly blurred, as foundation models can be prompted or conditioned on task descriptions rather than requiring explicit support sets.

However, foundation models also introduce new challenges for FSL. Their massive scale makes them computationally prohibitive for resource-constrained environments, contradicting one of the original motivations for FSL—efficient learning. Furthermore, their generalization capabilities, while impressive, are not guaranteed and can fail systematically on out-of-distribution tasks or domains underrepresented in pretraining data. The black-box nature of these models also complicates the application of the security analyses presented in Section 5, as adversarial vulnerabilities may exist in latent spaces that are neither interpretable nor directly accessible.

Despite these challenges, the trajectory of FSL research is increasingly converging with foundation model development. Future FSL systems will likely adopt hybrid architectures that combine the efficiency and interpretability of classical meta-learning with the representational power of foundation models, achieving the best of both paradigms.

7.3. In-Context Learning as Emergent Few-Shot Capability

A particularly significant development arising from foundation models is in-context learning (ICL), a phenomenon where large language and vision models learn to perform new tasks simply from examples provided within the input context, without any parameter updates. First systematically observed in GPT-3, in-context learning has since been demonstrated across modalities and represents a radical departure from conventional gradient-based adaptation (Brown et al., 2020).

In the context of Few-Shot Learning, in-context learning can be understood as an extreme form of few-shot adaptation where the learning occurs entirely

at inference time through attention mechanisms. The model does not update its weights; instead, it reconfigures its internal computation based on the contextual relationships among the provided examples and the query. This mechanism bears a conceptual resemblance to metric-based meta-learning, where classification is performed by comparing the query to support examples in an embedding space. However, in-context learning operates in a vastly higher-dimensional and more flexible representational space, enabled by the scale of the underlying model.

The relationship between in-context learning and classical FSL raises important theoretical questions. Research has shown that transformer-based in-context learning can implement gradient descent algorithmically within its forward pass, effectively simulating the adaptation process of optimization-based meta-learners without explicit parameter updates. This suggests a deep structural connection between the two paradigms, with in-context learning representing a more implicit and scalable realization of meta-learning principles (von Oswald et al., 2023).

For practical FSL applications, in-context learning offers several advantages. It eliminates the need for episodic training and task design, reducing the engineering overhead associated with classical meta-learning. It also enables seamless integration of multimodal information, as contextual examples can include text descriptions, images, and structured data simultaneously. However, the effectiveness of in-context learning is highly sensitive to prompt design—the selection, ordering, and formatting of examples significantly influence performance, a phenomenon known as prompt sensitivity or prompt brittleness.

Moreover, in-context learning inherits the security vulnerabilities discussed in Section 5, albeit in modified forms. Adversarial perturbations can be applied to contextual examples to manipulate model predictions, representing a new variant of feature-level attack. Poisoning attacks can target the examples retrieved from external knowledge bases to populate the context, and membership inference risks persist regarding whether specific examples were included in the model's pretraining data.

Looking forward, the integration of in-context learning with classical FSL methodologies represents a promising research direction. Hybrid approaches that use meta-learning to optimize prompt templates or example selection strategies for in-context learning could combine the efficiency of explicit adaptation with the representational power of large foundation models. Such integration would mark a significant step toward truly human-like few-shot learning systems that can rapidly acquire new concepts from minimal examples while maintaining robustness, interpretability, and security.

References

Cao, T., Law, M. T., & Fidler, S. (2020). *A theoretical analysis of the number of shots in few-shot learning*. arXiv preprint arXiv:1909.11722.

Da Silva, F. L., & Costa, A. H. R. (2019). *A survey on transfer learning for multiagent reinforcement learning systems*. *Journal of Artificial Intelligence Research, 64*, 645–703.

Finn, C., Abbeel, P., & Levine, S. (2017). *Model-agnostic meta-learning for fast adaptation of deep networks*. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. *International Conference on Learning Representations (ICLR)*.

Howard, J., & Ruder, S. (2018). *Universal language model fine-tuning for text classification*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Kim, G. N., Lee, H. J., Jeong, I. W., Shin, J. M., & Choi, S. H. (2026). *EAMF: Feature-level adversarial attack on metric-based few-shot learning models*. *IEEE Access*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. *International Conference on Learning Representations (ICLR)*.

Munkhdalai, T., & Yu, H. (2017). *Meta networks*. In *International Conference on Machine Learning*.

Parnami, A., & Lee, M. (2021). *Learning from few examples: A summary of approaches to few-shot learning*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Payandeh, A., Baghaei, K. T., Fayyazsanavi, P., Ramezani, S. B., Chen, Z., & Rahimi, S. (2023). *Deep representation learning: Fundamentals, technologies, applications, and open challenges*. *IEEE Access, 11*, 137621–137659.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., ... & Wu, Q. J. (2022). *A review of generalized zero-shot learning methods*. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4051–4070.

Schwartz, E., Karlinsky, L., Feris, R., Giryes, R., & Bronstein, A. (2022). *Baby steps towards few-shot learning with multiple semantics*. *Pattern Recognition Letters, 160*, 142–147.

Shu, J., Xu, Z., & Meng, D. (2018). *Small sample learning in big data era*. arXiv preprint arXiv:1808.04572.

Snell, J., Swersky, K., & Zemel, R. (2017). *Prototypical networks for few-shot learning*. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). *A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities*. *ACM Computing Surveys, 55*(13s), 1–40.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). *Learning to compare: Relation network for few-shot learning*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). *Matching networks for one shot learning*. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xian, Y., Korbar, B., Douze, M., Schiele, B., Akata, Z., & Torresani, L. (2020). *Generalized many-way few-shot video classification*. In *European Conference on Computer Vision* (pp. 111–127). Springer International Publishing.

Xu, X., Kamath, S., Butt, M. A., & Raducanu, B. (2025, October). *An h-space based adversarial attack for protection against few-shot personalization*. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 4904–4913).

Zhang, C., Hu, M., Li, W., & Wang, L. (2025). *Adversarial attacks and defenses on text-to-image diffusion models: A survey*. *Information Fusion, 114*, 102701, 1–15.

Zhao, J., Kong, L., & Lv, J. (2025). *An overview of deep neural networks for few-shot learning*. *Big Data Mining and Analytics, 8*(1), 145–188.

Zheng, B., Liang, C., & Wu, X. (2025). *Targeted attack improves protection against unauthorized diffusion customization*. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901.

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7, 47230–47244.

Liu, Y., Ma, X., Bailey, J., & Lu, F. (2020). Reflection backdoor: A natural backdoor attack on deep neural networks. In European Conference on Computer Vision (ECCV) (pp. 182–199).

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML) (pp. 8748–8763).

Shafahi, A., Huang, W. R., Najibi, M., et al. (2018). Poison frogs! Targeted clean-label poisoning attacks on neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 31.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In IEEE Symposium on Security and Privacy (SP) (pp. 3–18).

Tramer, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. In USENIX Security Symposium (pp. 601–618).

von Oswald, J., Niklasson, E., Randazzo, E., et al. (2023). Transformers learn in-context by gradient descent. In International Conference on Machine Learning (ICML) (pp. 35151–35174).