



Universiteit
Leiden
The Netherlands

Exploring deep learning for data-efficient and generalizable visual recognition

He, K.

Citation

He, K. (2026, September 22). *Exploring deep learning for data-efficient and generalizable visual recognition*. Retrieved from <https://hdl.handle.net/1887/4310276>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4310276>

Note: To cite this publication please use the final published version (if applicable).

Chapter 2

Few-shot & meta-learning methods for image understanding: a survey

In the previous chapter we have introduced the background and research questions for this thesis. Improving the model’s generalization with limited data is one of the main challenges in few-shot learning (FSL), and is strongly related to our proposed approaches described in Chapter 3-6, we give a comprehensive survey of recent advances that address FSL performance beyond predicting accuracy.

State-of-the-art deep learning systems (e.g. ImageNet image classification) typically require very large training sets to achieve high accuracies. Therefore one of the grand challenges is called few-shot learning where only a few training samples are required for good performance. In this survey, we illuminate one of the key paradigms in few-shot learning called meta-learning. These meta-learning methods, by simulating the tasks which will be presented at inference through episodic training, can effectively employ previous prior knowledge to guide the learning of new tasks. In this chapter, we provide a comprehensive overview and key insights into the meta-learning approaches and categorize them into three branches according to their technical characteristics, namely metric-based, model-based and optimization-based meta-learning. Due to the major importance of the evaluation process, we also present an overview of current widely-used benchmarks, as well as performances of recent meta-learning methods on these datasets. Based on over 200 papers in this survey, we conclude with the major challenges and future directions of few-shot learning and meta-learning.

This chapter is based on the following publication [33]:

- **He, K.**, Pu, N., Lao, M., and Lew, M.S. “Few-shot and meta-learning methods for image understanding: a survey.” International Journal of Multimedia Information Retrieval, 2023.

2.1 Introduction

Image classification [38, 39] is an important application in computer vision [40, 41] and machine learning [42, 43]. With the continuous development of deep learning [44, 45, 46], recent years have witnessed great breakthroughs in this area [47, 48]. However, such success relies on a huge amount of data [49, 50] (usually in the order of millions), which is difficult and time-consuming to obtain in the real world. In order to reduce the data requirement, there has been growing interest in small-sample image classification [51, 52, 53], such as few-shot classification [54, 55, 56], which learns a classification rule from few (1-5) labeled samples.

A core challenge in few-shot image classification is to alleviate the susceptibility of models to overfitting under a few-data regime [57, 58, 59]. To address this problem, researchers have proposed several promising approaches, such as transfer learning [60, 61], meta-learning [1, 2, 62] and data augmentation [63, 64, 65]. In transfer learning, a model is first trained on a source domain where abundant source data is available. Then this trained model is fine-tuned [66, 67, 68] on another target domain with few labeled target samples. The learnt prior knowledge can be transferred from source tasks to target tasks during this process. Meta-learning, or learning to learn, has emerged as one of the prominent approaches for few-shot learning. It is proposed to train a meta-learner which can quickly generalize to new tasks with few examples [69, 70, 71, 72]. A meta-learning procedure also involves learning at two levels, within and across tasks. Meta-learning approaches simulate the tasks that will be presented at inference through episodic training [73, 74, 75], enabling the generalization ability of meta-learner within minor adaption steps. Data augmentation methods are often used as pre-processing in few-shot learning (FSL). In order to solve the problem of insufficient training data, they introduce various kinds of existing data variance for the model to capture. For image classification, one commonly-used method is deformation [76, 77, 78, 79], including horizontal flipping, cropping and rotation. Besides these, more advanced methods, such as generating training samples and pseudo labels [80, 81, 82], are also an important part of data augmentation,

In this chapter, we present a survey of recent meta-learning methods for few-shot image classification. Meta-learning focuses on learning prior knowledge from previous tasks which can bring efficient downstream learning to new tasks. This learning mechanism enables models to learn new concepts quickly where only few samples are available. Meta-learning deserves special attention as it is an essential part of few-shot image classification and it has also demonstrated outstanding performance on benchmark datasets [83, 84]. To be specific, in this survey we divide meta-learning into three categories according to the different mechanisms, namely metric-based, model-based and optimization-based methods [85, 86, 87, 88].

A number of surveys on FSL have been proposed. In 2018, Shu *et al.* [52] provided an early survey on small sample learning, discussing approaches for different

scenarios (zero-shot learning [89, 90, 91] and FSL) and tasks (image classification, visual question answering [92, 93, 94] and object detection [95, 96, 97]). Wang *et al.* [98] conducted a comprehensive review in 2021, which provides a formal definition of FSL and distinguishes it from other machine learning problems, exploring FSL from a fundamental viewpoint of error decomposition in supervised learning. Li *et al.* [99] published another comprehensive review on FSL in 2021, which is entirely focused on meta-learning and review literature [100, 101, 102, 103] over a long period in this area. There is another review on few-shot image classification [104] published in 2023, which is fully devoted to metric learning methods [105, 106, 107]. Compared with these surveys [52, 98, 99, 104], our review presents an up-to-date survey of meta-learning approaches for few-shot image classification and provides a thorough analysis of these different kinds of methods to better understand their individual strengths and limitations.

2.2 The framework of few-shot image classification

2.2.1 Notation and definitions

In this section, we first present a brief introduction to few-shot learning and meta-learning, then provide the notation and unified definitions of few-shot image classification [108, 109, 110].

Few-shot learning is a surprising research area that focuses on learning patterns from a set of data (base classes) and then adapting to a disjoint set (novel classes) with limited training samples. Few-shot image classification is the one with most attention and researches. As the most popular approach for few-shot learning, meta-learning organizes the learning process into two phases, called meta-training and meta-testing. During each phase, the meta-training set or meta-testing set is split into multiple episodes. Each episode samples from the task distribution and is further divided into a small training set and a testing set.

In the standard few-shot image classification setting, two distinct datasets are involved, namely the base dataset $D_{base} = \{(x_i, y_i); x_i \in X_{base}, y_i \in Y_{base}\}_{i=1}^{N_{base}}$ and the novel dataset $D_{novel} = \{(x_i, y_i); x_i \in X_{novel}, y_i \in Y_{novel}\}_{i=1}^{N_{novel}}$, where x_i represents the original feature vector of the i -th image and y_i is the corresponding class label; N_{base} and N_{novel} denote the total numbers of instances in D_{base} and D_{novel} , respectively. The base dataset is an auxiliary dataset that is used to train the classifier to learn some prior or shared knowledge and the novel dataset is used for the classifier to perform new classification tasks. Note that D_{base} and D_{novel} are disjoint, which means $Y_{base} \cap Y_{novel} = \emptyset$. In order to train and test the classifier, the D_{novel} is usually split into the support set D_S and the query set D_Q and they share the same label space.

Definition 1. The few-shot image classification task aims to learn a classifier from D_{base} and D_S to correctly classify the samples in D_Q . It is generally termed as an N -way K -shot problem, where N and K denote the number of classes and instances in D_S , respectively. If $K = 1$, it becomes a one-shot image classification task; and if $K = 0$, then the task is called zero-shot classification.

Definition 2. A few-shot image classification task is called cross-domain few-shot image classification when the base dataset and the novel dataset are from two different domains, i.e., $X_{base} \neq X_{novel}$.

2.2.2 Datasets

In this section, we briefly introduce several well-known datasets for few-shot image classification. According to different data types, we categorize them into simple image dataset (Omniglot [111]), complex image dataset (MiniImageNet [14], TieredImageNet [62], CIFAR-FS [112] and FC100 [23]) and special image dataset (CUB-200 [113, 114]). Among these datasets, CIFAR-FS and FC100 are considered more difficult as the resolution of images from the two datasets is 32×32 . It's more challenging for models to extract useful information from low-resolution images. Statistics of these datasets and popular experimental settings are summarized below. We also present some sample images from these benchmark datasets in Figure 2.1.

Omniglot is one of the most frequently used benchmarks for evaluating few-shot image classification algorithms. It contains 1623 handwritten characters collected from 50 different alphabets. Each character consists of 20 samples, drawn by different human subjects. This dataset is usually augmented by the rotations in multiples of 90 degrees, 1200 characters are used for training and the rest for evaluation.

MiniImageNet and TieredImageNet are two mini versions of the large ImageNet dataset [13]. MiniImageNet is composed of 60,000 color images from 100 classes, with 600 images in each class. Following the widely-used splitting protocol proposed by Revi and Larochelle [115], 64 classes are used for training, 16 classes for validation and 20 classes for evaluation. TieredImageNet is another larger subset of ImageNet with a hierarchical structure. It contains 779,165 images from 34 high-level categories (or 608 classes), which are further split into 20 base categories (351 classes), 6 validation categories (97 classes), and 8 novel categories (160 classes).

CIFAR-FS and FC100 are two widely-used datasets derived from CIFAR-100 [12]. CIFAR-FS is constructed from 100 classes with 600 images per class. The 100 classes are split into 64, 16 and 20 classes for training, validation, and evaluation, respectively. FC100 also contains 100 classes, which are further divided into 20 super-categories, with five classes in each super-categories. FC100 is split into 12 base, 4 validation, and 4 novel super-categories.

CUB-200 is a fine-grained dataset consisting of 200 bird species. The CUB-200 dataset has two versions, while the initial version was proposed in 2010 [114] which



Figure 2.1: Sample images of benchmark datasets widely-used for few-shot image classification.

includes 6033 images and is extended to 11,788 images in 2011 [113]. The CUB-200-2010 dataset is often split into 130 base, 20 validation, and 50 novel classes [116]; while the CUB-200-2011 dataset is divided into 100 classes for training, 50 classes for validation, and 50 classes for testing [55].

MiniImageNet $\rightarrow$ CUB is a dataset designed for cross-domain few-shot image classification [21, 117, 118, 119]. MiniImageNet plays the role of the base dataset, while 50 classes of CUB-200-2011 are used for validation, and the remaining 50 classes serve for evaluation.

2.2.3 Evaluation process of few-shot image classification

In this section, we present a general procedure [22, 120, 121, 122, 123] to evaluate a classifier’s performance on N -way K -shot image classification problems in Algorithm 1. The whole evaluation process is composed of E episodes. In each episode, we first randomly select N classes from the novel label space with K samples in each class to form a support set D_S , and M examples from the remaining samples of those N classes to compose a query set D_Q . A final classifier can be obtained based on the base dataset D_{base} and support set D_S , which is used to predict labels of samples in

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

D_Q . We use $acc^{(e)}$ to denote the classification accuracy in the e -th episode, and the performance of a learning algorithm can be measured by the averaged classification accuracy over all episodes.

Algorithm 1: Evaluation process of N -way K -shot image classification

Input: D_{base} ; D_{novel} ; the total number of episodes E and a classifier f .

. **for** $e = 1, \dots, E$ **do**

 Randomly select N classes from D_{novel} ;

 Randomly select K samples from each class to compose the support set $D_S^{(e)}$;

 Randomly select M samples from the remaining samples of each class to compose the query set $D_Q^{(e)} = \{(X^{(e)}, Y^{(e)})\}$;

 Use the classifier f to predict labels of the samples from D_Q :

$\hat{Y}^{(e)} = f(X^{(e)} | D_{base}, D_S^{(e)})$;

 Record accuracy $acc^{(e)} = \frac{1}{M} \sum_{j=1}^M 1[\hat{Y}^{(e)} = Y^{(e)}]$;

end

Output: mean accuracy $\frac{1}{E} \sum_{e=1}^E acc^{(e)}$.

2.3 Paradigms of meta-learning for few-shot image classification

The goal of meta-learning for few-shot image classification [124, 125, 126, 127, 128] is to enable models, especially deep neural networks, to perform well on new tasks when only few samples are available. With the rapid development of few-shot learning [129, 130, 131], a number of meta-learning approaches [132, 133, 134] have been proposed. In this section, we provide a comprehensive overview of recent meta-learning studies and their advances. In order to let beginners better understand, we follow the main trend and still categorize meta-learning into metric-based, model-based and optimization-based methods. Besides, we also present other few-shot learning methods to make a comparison. Figure 2.2 shows an overview of few-shot image classification.

2.3.1 Metric-based meta-learning

Metric-based meta-learning methods [135, 136, 137, 138] aim to learn a distance metric, which can effectively measure the similarity among samples, ensuring it is optimal for new learning tasks. For few-shot image classification problems, the learned metric should follow the rule that samples from the same (or different) class should have a small (or large) distance.

Siamese Network is one of the most widely-used metric-based methods for one-shot image classification. The term "Siamese" was first proposed for signature verification

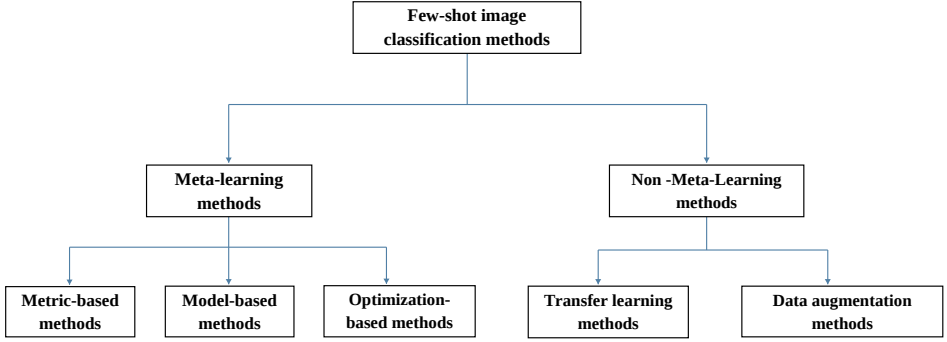


Figure 2.2: An overview of few-shot image classification approach.

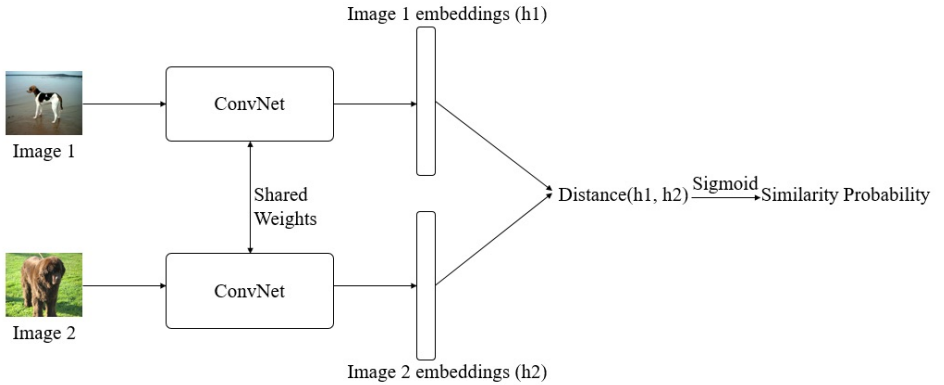


Figure 2.3: Siamese Network Architecture.

[139] and the principal structure of the Siamese Network was introduced for the fingerprint similarity estimation problem [140]. In 2015, Koch *et al.* [141] adopted a pair of identical VGG-styled [39] convolutional layers with shared weights to extract high-level features from two input images and calculate the weighted L_1 distance between the two feature vectors. The network finally outputs a score, representing the probability that the two images belong to the same class. The architecture of Siamese Network is shown in Figure 2.3. Wang *et al.* [142] proposed an Attention Based Siamese Network, which exploits an attention kernel function to measure the similarity between two feature vectors. To bridge the gap between one-shot image recognition [143, 144, 145] and regular classification, Lungu *et al.* [146] proposed a Multi-Resolution Siamese Network, which mixes different kernel size streams into one layer and adopts a hybrid training mechanism.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

As another powerful metric-based meta-learning method, Matching Network [14] uses different networks to encode support and query images. For support images embedding, a bidirectional Long-Short Term Memory (LSTM) [147] is used in the context of the support set D_S ; for query images embedding, an LSTM with an attention kernel is taken to enable the dependency on D_S , where the attention kernel [148, 149] is used to compute cosine similarities between support and query images and then normalize the similarities through a softmax function. Matching Network’s output is defined as a sum of the labels (one-hot encoded) of support images weighted by the attention kernel. In 2019, Mai *et al.* [150] proposed an Attentive Matching Network (AMN), introducing a feature-level attention mechanism to pay more attention to the features that can better reflect the inter-class differences and a complementary cosine loss function for optimization.

The initial Prototypical Network was proposed by Snell *et al.* [1] based on the hypothesis that there exists an embedding space where each class can be represented by a unique prototype, and all samples are supposed to cluster around their corresponding prototypes. Figure 2.4 shows the architecture of a Prototypical Network. A simple convolutional neural network with 4 layers is exploited to extract features, and the prototype of each class is defined as the mean value of feature embeddings from the support samples belonging to that class. The squared Euclidean Distance is employed as a distance metric, calculating the distance between query embeddings and each class prototype. Build on this, Li *et al.* [151] proposed a Covariance Metric Network (CovaMNet), using the covariance matrix of embedding vectors to represent the class prototype and also apply a covariance-based metric to measure the similarity between the query sample and the class prototype. Wang and Zhai [152] proposed a Prototypical Siamese Network (PSN), adding a prototype module in Siamese Network to obtain high-quality prototype representations of each class.

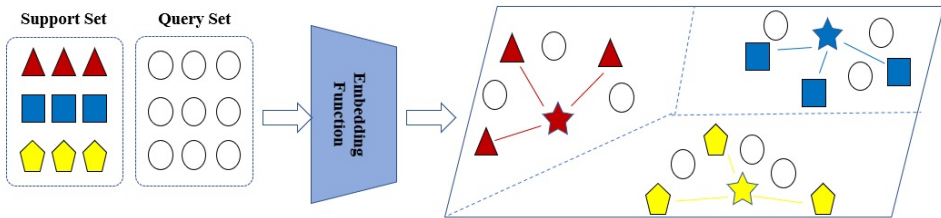


Figure 2.4: Prototypical Network Architecture.

Relation Network [24] is the first study that employs a neural network to estimate a similarity score of feature embeddings rather than manual computation. This model consists of two main components: an embedding module and a relation module. The embedding module is composed of convolutional blocks, mapping input images into an embedding space; and the relation module builds on two convolutional blocks and two fully-connected layers, calculating a relation score between each query and

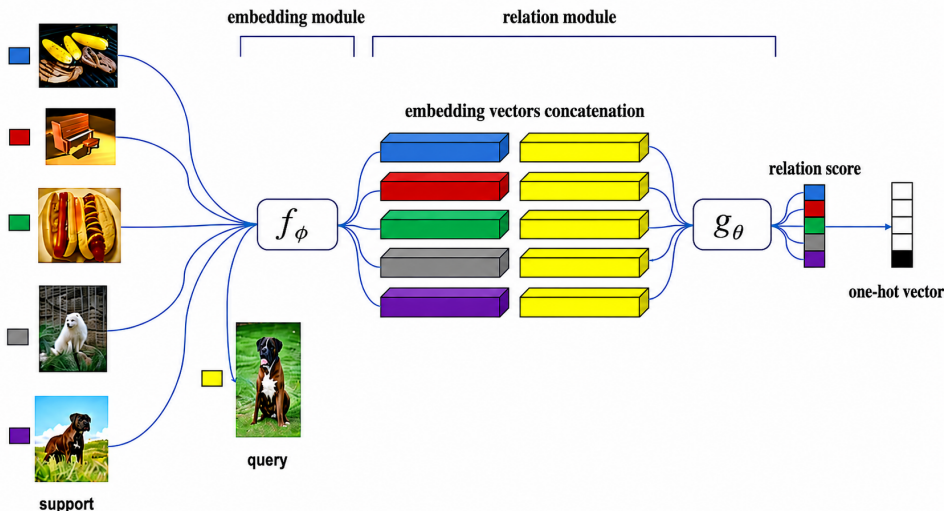


Figure 2.5: Relation Network Architecture. We use different colors to represent samples from different classes, where f_ϕ and g_θ denote embedding module and relation module, respectively

support image (or a class prototype when the number of support samples is more than one). Note that the feature embeddings of support and query images need to be concatenated together before they are fed into the relation module. The architecture of Relation Network is presented in Figure 2.5. In order to obtain discriminative features for fine-grained image classification [153, 154, 155], the subsequent work [156] proposed a Bi-Similarity Network (BSNet), which combines an extra cosine module with the existing similarity measure as a new relation module, generating a more compact feature space by forcing features to adapt to the new relation module.

In order to get optimal matching image regions, Zhang *et al.* [37] proposed a Deep-EMD algorithm, which adopts the Earth Mover’s Distance (EMD) [157, 158, 159] as a distance metric to calculate the similarity. They introduce a cross-reference mechanism to produce the weights of elements in the EMD formulation and embed the EMD layer into the network for end-to-end training. Motivated by this, Xie *et al.* [160] proposed a deep Brownian Distance Covariance (DeepBDC) approach, which applies the BDC metric for few-shot learning. To learn discriminative feature representations, Afrasiyabi *et al.* [161] proposed a Mixture-based Feature Space Learning (MixtFSL) approach, learning both the feature representations and the mixture model via an online manner. Different from those few-shot classification methods that extract a single feature vector from each image, Afrasiyabi *et al.* [162] held the view that a set-based representation can build a richer and more robust representation of images from base classes. To do so, they proposed a Matching Feature

Sets method which embeds self-attention modules in between convolutional blocks and introduces set-to-set metrics for evaluation. We summarize those introduced metric-based meta-learning approaches in Table 2.1.

2.3.2 Model-based meta-learning

With the goal of fast learning, model-based methods [163, 164] mainly focus on model architectures, adjusting model parameters based on presented tasks. There are several frequently used architectures in model-based methods, such as Convolutional Neural Networks (CNNs) [165], Recurrent Neural Networks (RNNs) [166, 167], Long Short-Term Memory (LSTM) [168], etc. According to the model architecture types, these model-based methods are further separated into memory-based, rapid-adaptation-based and miscellaneous models.

Memory-Augmented Neural Network(MANN) is a famous memory-based method proposed by Santoro *et al.* [169], which aims at improving task adaptation by utilizing the Neural Turing Machine (NTM) [170, 171, 172]. NTM is a neural network that integrates an external memory component during its learning process, enabling it has access to retrieve previously stored information. To be specific, NTM consists of a controller, interacting with an external memory module via a number of read and write heads. The NTM scheme is shown in Figure 2.6. In MANN, a new addressing mechanism namely least recently used access (LRUA) [169] is proposed, writing memories to either the least used memory location or the most recently used memory location. Through the stored information of a coupled representation-class label in the external memory, MANN can access them for later classification. Tran *et al.* [173] proposed a Memory Augmented Matching Network (MAMN), which combines MANN and Matching Network. In MAMN, to reduce the biased on class prototypes caused by data distribution skew, weighted class prototypes are introduced by incorporating the distances of class-wise samples. As another memory-based meta-learning method, Memory Matching Network (MM-Net) [174] incorporates the memory module extracted from Key-Value Memory Network [175] into Matching Network. Different from traditional one-shot learning methods, MM-Net encodes and generalizes the whole support set into memory slots and can generate a unified model regardless of the number of shots and categories.

Meta Network (MetaNet) [176] is a model that designed with specific architecture and training process for rapid adaption across tasks. Meta Network contains a base learner, a meta-learner and an external memory. It performs a generic knowledge acquisition in a meta space and shifts its inductive biases via fast parameterization for rapid generalization. Conditional Shifted Neurons (CSNs) [177] is a generic neural mechanism designed for fast adaption, which is able to extract conditional information and generate conditional shifts for prediction during the meta-learning process. Compared with previous works [169, 176, 178], CSNs is more efficient computationally as the number of neurons is usually much smaller than that of weight

2.3 Paradigms of meta-learning for few-shot image classification

Table 2.1: A summary of presented metric-based meta-learning approaches.

Approach	Author	Year	Brief Description
Siamese Network [141]	Koch <i>et al.</i>	2015	Exploit a pair of identical networks with shared weights to extract features and calculate the distance between feature vectors.
Attention Based Siamese Network [142]	Wang <i>et al.</i>	2018	Use an attention kernel function to measure the similarity.
Multi-Resolution Siamese Network [146]	Lungu <i>et al.</i>	2020	Mix different kernel size streams into one layer.
Matching Network [14]	Vinyals <i>et al.</i>	2016	Use different networks to encode support and query images.
AMN [150]	Mai <i>et al.</i>	2019	Introduce a feature-level attention mechanism.
Prototypical Network [1]	Snell <i>et al.</i>	2017	Propose class prototypes and calculate the distances between query samples and class prototypes.
CovaMNet [151]	Li <i>et al.</i>	2019	Apply the covariance matrix of embedding vectors to represent the class prototype and use a covariance-based metric to measure the similarity.
PSN [152]	Wang and Zhai	2020	Add a prototype module in Siamese Network.
Relation Network [24]	Sung <i>et al.</i>	2018	Employ a neural network to estimate a similarity score of feature embeddings.
BSNet [156]	Li <i>et al.</i>	2021	Combine an extra cosine module with the existing similarity measure as a new relation module.
DeepEMD [37]	Zhang <i>et al.</i>	2020	Adopt the Earth Mover’s Distance (EMD) as a distance metric to calculate the similarity.
DeepBDC [160]	Xie <i>et al.</i>	2022	Apply the Brownian Distance Covariance (BDC) metric.
MixtFSL [161]	Afrasiyabi <i>et al.</i>	2021	Learn the feature representations and the mixture model jointly in an online manner.
Matching Feature Sets [162]	Afrasiyabi <i>et al.</i>	2022	Embed self-attention modules between convolutional blocks and introduce set-to-set metrics.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

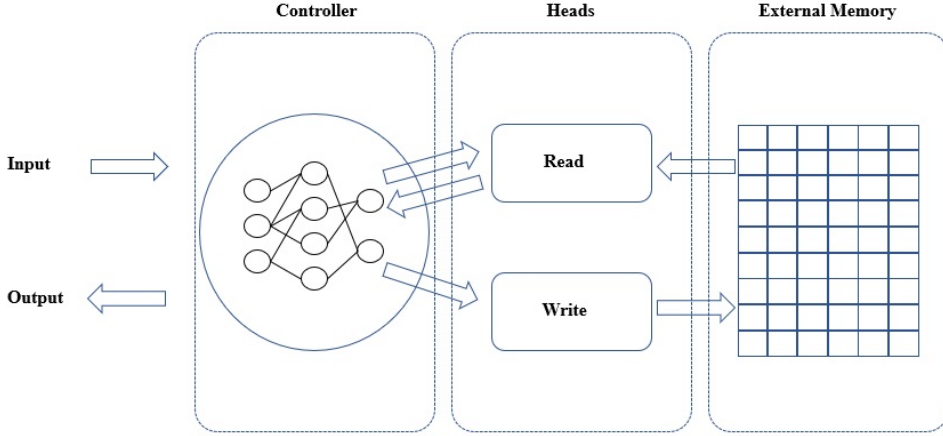


Figure 2.6: Neural Turing Machine scheme.

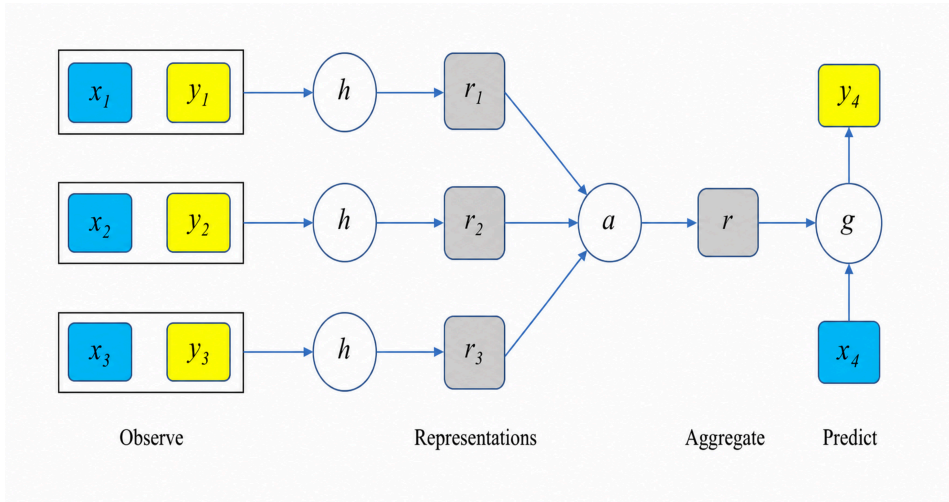


Figure 2.7: Conditional Neural Processes scheme. The encoder h maps each observed context pair (x_i, y_i) to an individual representation r_i . The permutation-invariant aggregation operator a combines these representations into a global context representation r . The decoder g then uses r together with the target input x_4 to predict the corresponding target output y_4 .

parameters. Moreover, CSNs can be integrated into various neural architectures, including CNNs and RNNs. Similar to MetaNet, CSNs contains a base learner, a meta learner and a memory module. During the description time, the meta learner extracts and employs conditional information to generate memory values for samples within a task; at the prediction phase, the meta learner generates query keys of query images by a key function for the purpose of getting the value of conditional shift.

Simple Neural Attentive Learner (SNAIL) [179] is a general model-based meta-learning architecture that incorporates temporal convolution and soft attention mechanism. The temporal convolution acts as high bandwidth memory access, and the soft attention enables access to specific pieces of information. This combination enables models to better leverage information from past experiences. Similar to SNAIL, Garnelo *et al.* [180] proposed Conditional Neural Processes (CNPs) which consists of a meta-learner and task-learner. The meta-learner generates a memory value by aggregating representations of the support set, and the task-learner makes predictions by processing the aggregated representations. Figure 2.7 shows the CNPs scheme. We also make a short summary of those model-based meta-learning approaches and present it in Table 2.2.

Table 2.2: A summary of presented model-based meta-learning approaches.

Approach	Author	Year	Brief Description
MANN [169]	Santoro <i>et al.</i>	2016	Employ a modified NTM to quickly assimilate new data into memory.
MAMN [173]	Tran <i>et al.</i>	2019	Combine MANN and Matching Network.
MM-Net [174]	Cai <i>et al.</i>	2018	Incorporate the memory module extracted from Key-Value Memory Network into Matching Network.
MetaNet [176]	Munkhdalai and Yu	2017	Perform a generic knowledge acquisition in a meta space and shift its inductive biases via fast parameterization.
CSNs [177]	Munkhdalai <i>et al.</i>	2018	Extract conditional information and generate conditional shifts for prediction during the meta-learning process.
SNAIL [179]	Mishra <i>et al.</i>	2018	Incorporate temporal convolution and soft attention mechanism.
CNPs [180]	Garnelo <i>et al.</i>	2018	Make predictions based on concise representations of seen classes.

2.3.3 Optimization-based meta-learning

Optimization-based meta-learning methods are an important vital branch in the field of few-shot image classification [181, 182, 183, 184, 185]. Basically, this kind of algorithm attempts to obtain a better initialization model or gradient descent direction by leveraging the meta-learning architecture, and optimizes the initialization parameters through episodic training, enabling an optimization procedure to work on a small number of training samples. Optimization-based methods generally contain a task-specific learner trained for a given task and a meta-learner trained on distributions of tasks.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

In 2017, Finn *et al.* [2] proposed Model-Agnostic Meta-Learning (MAML), the first algorithm for learning an initialization. The key idea of MAML is to enable the model's parameters to adapt fast to new unseen tasks through a gradient-based learning rule. During the meta-training phase, MAML attempts to update the task-specific parameters and the global initialization jointly in an iterative manner. The MAML scheme is presented in Figure 2.8. The main contribution of MAML is its compatibility in different application domains, not only in classification, but also in regression [186, 187, 188] and reinforcement learning [189, 190, 191]. To address the limitation of neural networks that are trained with gradient-based optimization on few-shot learning tasks [84, 192, 193], Ravi and Larochelle [115] proposed an LSTM-based meta-learner to learn both the exact task-specific optimization of a classifier, as well as good initialization values for the parameters of task-specific learner.

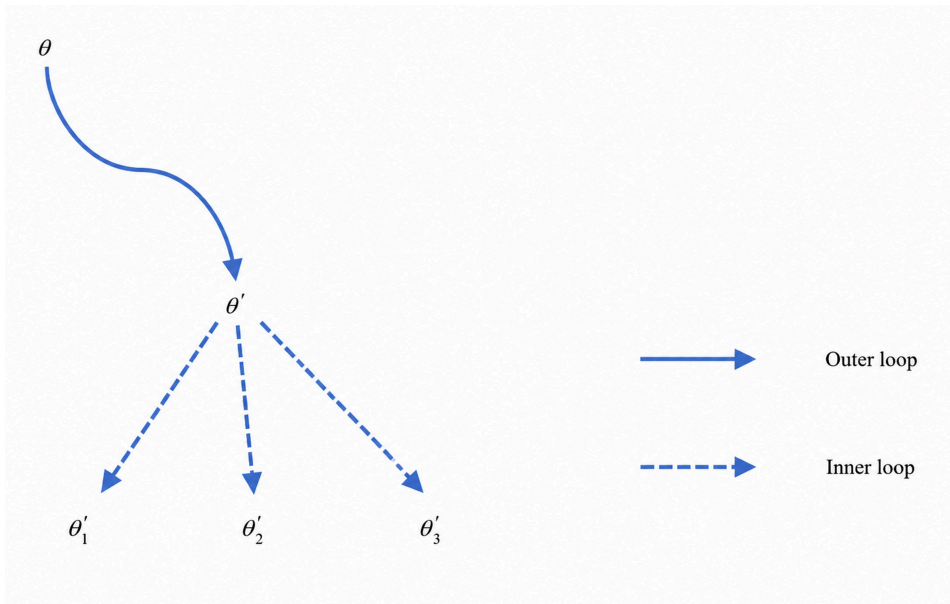


Figure 2.8: Model-Agnostic Meta-Learning scheme. The outer loop updates the shared model parameters from θ to θ' . The inner loop adapts the model separately to each sampled task, producing task-specific parameters θ'_i .

By taking ideas from Prototypical Network and MAML, Triantafillou *et al.* [194] proposed Proto-MAML, incorporating the advantages of both the former's simple inductive bias and the latter's flexible adaptation mechanism. As an extension to MAML, CAVIA [195] divides the model into parameters and task-specific context parameters which are shared across tasks. Compared with MAML, CAVIA is less prone to meta-overfitting and easier to parallelize. To address the issue that meta-learning models would be too biased toward existing tasks and lead to poor generalization, Jamal and Qi [29] proposed a Task Agnostic Meta-Learning (TAML) algorithm, where two approaches are exploited to train a model unbiased over tasks.

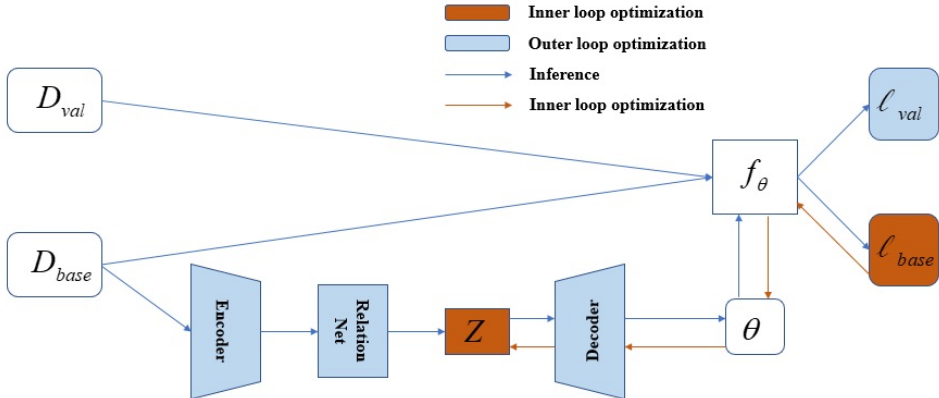


Figure 2.9: Latent Embedding Optimization scheme. Here, z denotes the latent embedding learned from D_{base} , θ denotes the learnable model parameters, and f_{θ} denotes the prediction function parameterized by θ . The model is optimized through an inner loop based on the base loss ℓ_{base} and an outer loop based on the validation loss ℓ_{val} .

In order to improve generalization performance, BaiK *et al.* [196] proposed a novel framework called Meta-Learning with Task-Adaptive Loss Function (MeTAL). Particularly, MeTAL learns a task-adaptive loss function through two meta-learners and can be applied to different MAML variants.

Wang *et al.* [197] introduced a new approach called task-aware feature embeddings for low-shot learning (TAFE-Net) which mainly concentrates on tuning task-specific feature embedding through the generic embedding of a meta-learner. TAFE-Net is composed of a meta-learner and a prediction network, where the task-aware feature embedding is obtained by utilizing the meta-learner to develop task-specific feature layers of the prediction network. Sun *et al.* [198] introduced a meta-transfer learner (MTL) method, which focuses on generating task-specific feature extractors by leveraging both meta-learning and transfer learning. In MTL, scaling and shifting operations are introduced on pre-trained feature embeddings to freeze the feature extractor. Besides, similar fine-tuning steps are taken in MTL as those in previous work [55]. This work also proposed a novel hard task meta-batch process that put more focus on hard tasks through sampling extra instances from the classes that the classifier failed.

Considering difficulties that exist in optimization on high-dimensional parameter spaces such as those faced by MAML [2], Rusu *et al.* [199] proposed an innovative algorithm called Latent Embedding Optimization (LEO) that learns a low-dimensional latent representation of model parameters and performs optimization-based meta-learning in this space. Similar to MAML, LEO also consists of an inner loop training where the task-specific values are learned and an outer loop training where global shared initializations are updated. To instantiate low-dimensional latent embedding

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

of model’s parameters, samples pass through a combination of an encoder and a relation network. The encoder is used to generate hidden codes from the support set. Then, these hidden codes are concatenated pairwise and fed into a relation network, leading to a probability distribution over latent codes in a lower dimension. Finally, the decoder produces task-specific initial parameters which are differentiable to backpropagate for adaptation. The LEO scheme is shown in Figure 2.9. We present a short summary of optimization-based meta-learning approaches in Table 2.3.

Table 2.3: A summary of presented optimization-based meta-learning approaches.

Approach	Author	Year	Brief Description
MAML [2]	Finn <i>et al.</i>	2017	Enable a model’s parameters can adapt fast to new unseen tasks through the gradient-based learning rule.
LSTM-based meta-learner [115]	Ravi and Larochelle	2017	Learn both the exact task-specific optimization of a classifier and good initialization values for the parameters of task-specific learner.
Proto-MAML [194]	Triantafillou <i>et al.</i>	2020	Combine Prototypical Network and MAML.
CAVIA [195]	Zintgraf <i>et al.</i>	2019	Divide the model into parameters and task-specific context parameters which are shared across tasks.
TAML [29]	Jamal and Qi	2019	Train a model unbiased over tasks.
MeTAL [196]	Baik <i>et al.</i>	2021	Learn a task-adaptive loss function through two meta-learners.
TAFE-Net [197]	Wang <i>et al.</i>	2019	Tune task-specific feature embedding through the generic embedding of a meta-learner.
MTL [198]	Sun <i>et al.</i>	2019	Generate task-specific feature extractors by leveraging both meta-learning and transfer learning.
LEO [199]	Rusu <i>et al.</i>	2019	Learn a low-dimensional latent representation of model parameters.

2.3.4 Other methods

Transfer learning involves leveraging knowledge learned from a related task to enhance learning in a new task [200, 201, 202, 203, 204]. In the few-shot image classification scenario, transferring knowledge from another network is a viable option when original data is too limited to train a deep neural network from scratch. Compared with meta-learning, the learning experience involved in transfer learning is much narrower. To address few-shot hyperspectral image classification problems, Qu *et al.* [205] applied the transfer learning scheme to extract learned intrinsic

representations from the same kind of objects in different domains. Tai *et al.* [206] proposed a novel few-shot transfer learning approach for synthetic aperture radar image classification, which uses a connection-free attention module to transfer features from a source network to a target network. Sun and Yang [207] proposed Trans-transfer Learning, a two-phase learning method for few-shot fine-grained visual categorization problems. In some cases, knowledge transfer may also fail when the source domain and target domain are not related to each other, even causing negative transfer. To address this problem, Liu *et al.* [208] proposed an Analogical Transfer Learning (ATL), following the analogy strategy to effectively control the occurrence of negative transfer.

Considering the fundamental problem in few-shot image classification that models are prone to overfitting caused by few training samples, many researchers proposed a number of data augmentation approaches [209, 210, 211] to improve sample diversity and prevent overfitting during training. Goodfellow *et al.* [212] proposed the well-known Generative Adversarial Nets (GAN), which contains a generator for generating similar images and a discriminator for distinguishing. Based on GAN, Mehrotra and Dukkipati [213] proposed to generate samples for specific tasks, enabling these generated samples more suitable for few-shot learning. Zhang *et al.* [214] proposed MetaGAN. To help the classifier learn a clearer decision boundary, MetaGAN involves GAN and part of the classification network during the training process. Li *et al.* [26] proposed Adversarial Feature Hallucination Network (AFHN), using conditional Wasserstein Generative Adversarial Network (cWGAN) to generate samples.

We present experimental results of recent meta-learning methods in Table 2.4 and Table 2.5. Table 2.4 shows performances of different approaches on Omniglot. Omniglot is a handwritten dataset with multiple handwriting styles, languages and stroke types, this diversity makes Omniglot suitable for training deep learning algorithms. From Table 2.4 we can find that most meta-learning approaches obtain over 98% accuracies on Omniglot. Table 2.5 shows experimental results on Mini-ImageNet and TieredImageNet. These two datasets contain images with different objects, scenes and lighting conditions, which can improve the model’s robustness. However, the limitations in dataset size and image quality may affect the model’s performance. Table 2.5 shows that DeepBDC [160] and Matching Feature Sets [162] achieved best results on both datasets.

2.4 Major Challenges and future directions

Although meta-learning methods have achieved promising performance in few-shot image classification, there remain some vital challenges that ought to be dealt with in the future. These existing issues and suggested future research directions are outlined here.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

Table 2.4: Accuracy results on Omniglot dataset reported in original papers, with mean accuracy (%) and 95% confidence interval. i: Metric-based; ii: Model-based; iii: Optimization-based

Model	Category	Omniglot 5-way	
		1-shot	5-shot
Siamese Network [141]	i	97.30	98.40
Attention Based Siamese Network [142]	i	99.60	99.80
Multi-Resolution Siamese Network [146]	i	-	-
Matching Network [14]	i	98.10	98.90
AMN [150]	i	99.44 ± 0.09	99.86 ± 0.06
Prototypical Network [1]	i	98.80	99.70
CovaMNet [151]	i	-	-
PSN [152]	i	-	-
Relation Network [24]	i	99.60 ± 0.20	99.80 ± 0.10
BSNet [156]	i	-	-
DeepEMD [37]	i	-	-
DeepBDC [160]	i	-	-
MixtFSL [161]	i	-	-
Matching Feature Sets [162]	i	-	-
MANN [169]	ii	82.80	94.90
MAMN [173]	ii	98.90	99.70
MM-Net [174]	ii	99.28 ± 0.08	99.77 ± 0.04
MetaNet [176]	ii	98.45	-
CSNs [177]	ii	98.42 ± 0.21	99.37 ± 0.28
SNAIL [179]	ii	99.07 ± 0.16	99.78 ± 0.09
CNPs [180]	ii	95.30	98.50
MAML [2]	iii	98.70 ± 0.40	99.90 ± 0.10
LSTM-based meta-learner [115]	iii	-	-
Proto-MAML [194]	iii	-	-
CAVIA [195]	iii	-	-
TAML [29]	iii	99.47 ± 0.25	99.83 ± 0.09
MeTAL [196]	iii	-	-
TAFE-Net [197]	iii	-	-
MTL [198]	iii	-	-
LEO [199]	iii	-	-

Table 2.5: Accuracy results on MiniImageNet and TieredImageNet datasets reported in original papers, with mean accuracy (%) and 95% confidence interval. i: Metric-based; ii: Model-based; iii: Optimization-based

Model	Category	MiniImageNet 5-way		TieredImageNet 5-way	
		1-shot	5-shot	1-shot	5-shot
Siamese Net-work [141]	i	-	-	-	-
Attention Based Siamese Network [142]	i	51.20	69.70	-	-
Multi-Resolution Siamese Net-work [146]	i	-	-	-	-
Matching Net-work [14]	i	46.60	60.00	-	-
AMN [150]	i	54.97 ± 0.77	71.84 ± 0.67	-	-
Prototypical Network [1]	i	49.42 ± 0.78	68.20 ± 0.66	-	-
CovaMNet [151]	i	51.19 ± 0.76	67.65 ± 0.63	-	-
PSN [152]	i	48.70	69.40	-	-
Relation Net-work [24]	i	50.44 ± 0.82	65.32 ± 0.70	-	-
BSNet [156]	i	-	-	-	-
DeepEMD [37]	i	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
DeepBDC [160]	i	67.83 ± 0.43	85.45 ± 0.29	73.82 ± 0.47	89.00 ± 0.30
MixtFSL [161]	i	64.31 ± 0.79	81.66 ± 0.60	70.97 ± 1.03	86.16 ± 0.67
Matching Feature Sets [162]	i	68.32 ± 0.62	82.71 ± 0.46	73.63 ± 0.88	87.59 ± 0.57
MANN [169]	ii	-	-	-	-
MAMN [173]	ii	49.80	66.50	-	-
MM-Net [174]	ii	53.37 ± 0.48	66.97 ± 0.35	-	-
MetaNet [176]	ii	49.21 ± 0.96	-	-	-
CSNs [177]	ii	56.88 ± 0.62	71.94 ± 0.57	-	-
SNAIL [179]	ii	55.71 ± 0.99	68.88 ± 0.92	-	-
CNPs [180]	ii	-	-	-	-
MAML [2]	iii	48.70 ± 1.84	63.11 ± 0.92	-	-
LSTM-based meta-learner [115]	iii	43.44 ± 0.77	60.60 ± 0.71	-	-
Proto-MAML [194]	iii	-	-	-	-
CAVIA [195]	iii	51.82 ± 0.65	65.85 ± 0.55	-	-
TAML [29]	iii	51.73 ± 1.88	66.05 ± 0.85	-	-
MeTAL [196]	iii	66.61 ± 0.28	81.43 ± 0.25	70.29 ± 0.40	86.17 ± 0.35
TAFE-Net [197]	iii	-	-	-	-
MTL [198]	iii	61.20 ± 1.80	75.50 ± 0.80	-	-
LEO [199]	iii	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09

2.4.1 Limitations and challenges

- **Data availability and computational complexity.** In image classification, a large dataset typically has a thousand (or more) categories. Meta-learning approaches also require a large amount of data and computational resources, but in few-shot scenarios, it is quite challenging to collect sufficient data. For deep testing of meta-learning we may need thousands of large datasets! These may also be very difficult and slow to process.
- **Model selection.** There is not a one-size-fits-all model so selecting an appropriate model is important. Model selection is more crucial in few-shot image classification scenarios as the model is prone to overfitting the training data. The model may perform well on the base set while lacking generalization on new tasks.
- **Transferability.** Meta-learning models can transfer learned knowledge between various tasks. The success of transferability depends on the similarity between the tasks. Sometimes new tasks may have significant differences from old ones, making it difficult to transfer learned knowledge effectively, such as cross-domain tasks.
- **Task dependence.** Most meta-learning approaches are designed to work for a specific set of tasks or domains. They may not perform well on new tasks or domains that are significantly different from the ones used during training. Improving meta-learning’s generalization ability can be a hard task.
- **Interpretability.** Interpretability is a critical aspect of neural approaches that refers to the ability to understand how a model works. Unfortunately, neural approaches can be extremely challenging to interpret, and thus difficult to understand how it learns to learn and make predictions or decisions. This issue can make it arduous to debug, diagnose and improve models’ performances.

2.4.2 Future directions

- **Enhancing generalized feature learning.** To address the main challenge in few-shot learning, learning from a handful of samples [6, 215, 216], meta-learning employs shared knowledge from previously experienced tasks for unseen tasks. However, in most existing meta-learning methods, researchers attempt to learn discriminative features via attention mechanisms, multi-task learning, data augmentation, and so on. One major research direction is developing new approaches for learning features that generalize better to new domains; and evaluation measures for assessment and selection of the learned features.

- **Practice of episodic training strategy.** In order to realize fast adaption to new tasks with limited samples, episodic training requires that each training episode should have the same number of classes and examples as the evaluation episode. But, this setting is prone to catastrophic forgetting [217, 218, 219] and leads to model underfitting in base classes. A number of approaches have been proposed to address this issue, further improving model performance on both base and novel classes remains a vital direction for future work.
- **Improving stability.** Despite the continuous improvement of meta-learning in few-shot image classification, one existing issue is that some meta-learning methods obtain state-of-the-art performance on special datasets, but perform not well on other benchmarks. For example, a metric-based meta-learning method named global class representation (GCR) [220] achieved great performance on Omniglot, but cannot compete with other non-metric-based methods on miniImageNet. Further exploration of stable models [221, 222] will be very valuable.
- **Cross-domain and multimodal meta-learning.** In principle, the base dataset D_{base} and novel dataset D_{novel} in few-shot learning can be from different domains [223, 224]. However, most model performances will decline when the difference between D_{base} and D_{novel} . Developing meta-learning methods on cross-domain performance can be one future research direction. Multimodal deep learning has also brought great opportunities to few-shot learning [225, 226, 227]. For example, Peng *et al.* [228] proposed a Knowledge Transfer Network (KTN), which combines semantic features and image features for few-shot image classification tasks. Therefore, how to design a more appropriate multimodal fusion method is a research trend in few-shot image classification.

2.5 Chapter Conclusions

This chapter presents a survey comprised of over 200 papers on recent few-shot learning and meta-learning research for image understanding. Based on the research literature, we introduce the general approaches for few-shot learning and then turn to one of the key approaches called meta-learning. We separate existing meta-learning methods into three important categories: metric-based, model-based and optimization-based methods. We introduce both classical and state-of-the-art approaches in each category, and summarize the state-of-the-art. We also present the state-of-the-art performance of the literature approaches on well-known datasets. According to our study, we conclude with limitations, challenges and weaknesses for meta-learning and present promising directions of meta-learning from the perspectives of generalization, effectiveness and applicability.