SCIENCE CHINA
Information Sciences

February 2021, Vol. 64 120104:1–120104:14
https://doi.org/10.1007/s11432-020-3156-7

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

Task-wise attention guided part complementary learning for few-shot image classification

Gong CHENG[1,2,3], Ruimin LI[1,2,3], Chunbo LANG[1,2,3] & Junwei HAN[2*]

[1]*Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen* 518057*, China;*
[2]*School of Automation, Northwestern Polytechnical University, Xi'an* 710072*, China;*
[3]*CETC Key Laboratory of Aerospace Information Applications, Shijiazhuang* 050081*, China*

Received 8 November 2020/Accepted 24 December 2020/Published online 20 January 2021

Abstract A general framework to tackle the problem of few-shot learning is meta-learning, which aims to train a well-generalized meta-learner (or backbone network) to learn a base-learner for each future task with small training data. Although a lot of work has produced relatively good results, there are still some challenges for few-shot image classification. First, meta-learning is a learning problem over a collection of tasks and the meta-learner is usually shared among all tasks. To achieve image classification of novel classes in different tasks, it is needed to learn a base-learner for each task. Under the circumstances, how to make the base-learner specialized, and thus respond to different inputs in an extremely task-wise manner for different tasks is a big challenge at present. Second, classification network usually inclines to identify local regions from the most discriminative object parts rather than the whole objects for recognition, thereby resulting in incomplete feature representations. To address the first challenge, we propose a task-wise attention (TWA) module to guide the base-learner to extract task-specific image features. To address the second challenge, under the guidance of TWA, we propose a part complementary learning (PCL) module to extract and fuse the features of multiple complementary parts of target objects, and thus we can obtain more specific and complete information. In addition, the proposed TWA module and PCL module can be embedded into a unified network for end-to-end training. Extensive experiments on two commonly-used benchmark datasets and comparison with state-of-the-art methods demonstrate the effectiveness of our proposed method.

Keywords few-shot learning, meta-learning, task-wise attention, part complementary learning

Citation Cheng G, Li R M, Lang C B, et al. Task-wise attention guided part complementary learning for few-shot image classification. Sci China Inf Sci, 2021, 64(2): 120104, https://doi.org/10.1007/s11432-020-3156-7

1 Introduction

Deep learning has made great achievements in many fields, such as object detection [1–6], image classification [7–10], semantic segmentation [11, 12], and so on [13, 14], but these methods all rely heavily on substantial labeled training data. How to undertake new tasks with a small amount of labeled data has become a new research hotspot [15, 16]. Few-shot learning (FSL) provides a solution to this problem and meta-learning is a general framework to achieve the goal of few-shot learning [17–20].

Typically, a meta-learning algorithm consists of two main components, namely, a meta-learner and a base-learner [20, 21]. The meta-learner is generally trained from a large number of similar few-shot tasks (also known as episodes), with the goal of being well generalized to new tasks, by maximizing the united generalization power of the learner on all tasks. The base-learner is trained for each specific task with only a few labeled training examples. In other words, the goal of meta-learning is to train a well-generalized meta-learner on a variety of tasks in order to facilitate the training of base-learners for each future task. Therefore, the key of meta-learning is to develop high-capacity yet trainable meta-learners and base-learners.

During the past few years, researchers have proposed many meta-learners, including metrics [22, 23] and optimizers [24–31], which have shown promising results for few-shot learning. For instance, some

* Corresponding author (email: junweihan2010@gmail.com)

studies such as prototypical networks [22] and matching networks [23] proposed to use distance metric learning algorithms as meta-learners for non-parametric learners like nearest neighbor classifiers and their variants. Distance metric-based methods work through matching the labeled examples of the support set and the unlabeled data of the query set, which work well for few-shot image classification problem and are widely used. However, a metric does not really train a learner and it actually affects its behavior through modifying the distances between the examples. Therefore, meta-learners based on distance metric are mainly suitable for non-parametric learners.

Recent studies move meta-learners towards deep neural networks. Some studies used a recurrent model such as long short-term memory (LSTM) as meta-learners owing to its capability of adaptively modeling optimization algorithms [24–26]. For example, Andrychowicz et al. [24] formulated LSTM as a stochastic gradient descent (SGD)-like optimizer to imitate the process of model update of the learner and obtained promising results when compared to hand-crafted optimization algorithms. Ravi et al. [25] proposed an LSTM-based meta-learner that is trained to learn a neural network classifier which can quickly converge on each specific task. The meta-learner could capture both long-term knowledge common across all the tasks and short-term knowledge within a specific task. LSTM-based meta-learners are versatile and show promising results; however, using such meta-learners to learn learners such as convolutional neural networks (CNNs) encounters high complexity. Besides, Finn et al. [27] proposed a simple yet effective model-agnostic meta-learning (MAML) algorithm. Li et al. [28] developed an SGD-like, conceptually simple, and easy-to-implement meta-learner, called meta-SGD, which can initialize and adapt all differentiable learners in just one step. Furthermore, in order to obtain an unbiased meta-learner and improve its generalization ability, Jamal and Qi [29] proposed an entropy-based task-agnostic meta-learning (TAML) algorithm. Zhou et al. [30] proposed to equip a meta-learner with a deep learning based concept generator to enable it to better learn in the new concept space. Sun et al. [31] presented a new few-shot learning approach named meta-transfer learning (MTL) which can learn to adapt a deep neural network to few-shot learning tasks.

The abovementioned methods mainly focus on the research of meta-learner, which is a slow learning process of task-agnostic meta-level model performing across a variety of tasks with the goal of capturing common knowledge of different tasks. In fact, the learning of base-learners acting with each specific task is also significantly crucial for few-shot learning. However, there are only few literatures specially focusing on the research of base-learners. For example, Lee et al. [32] proposed to leverage linear classifiers (e.g., support vector machine, ridge regression) as base-learners to tackle few-shot learning problems instead of nearest-neighbor classifiers in prototypical networks. Promising results with the consideration of dual formation and Karush-Kuhn-Tucker (KKT) conditions indicate that the regularized linear model produced better generalization under novel categories and reduced overfitting. Ref. [21] introduced a differentiable ridge regression base-learner, called R2-D2, which achieved a tradeoff between adaptation-free for novel class samples (e.g., metric-based approaches) and costly iterative techniques (e.g., LSTM-based approaches). Lifchitz et al. [33] broadened the well-trained base network by introducing extra convolution kernels at the top, called new neurons implants, striving to achieve rapid adaptation of base-learners on previously unseen classes with few available annotations.

These innovative design and considerable improvements for base-learners stimulated our research interest. In addition to optimizing the network classifier, how to capture specific but discriminative information under different task requirements is equally vital for the few-shot learning scenarios. Towards this end, a task-specific feature extraction module named task-wise attention (TWA) was proposed in this work. TWA learns to find the most representative features associated with current task by reinforcing or suppressing a portion of the knowledge provided by meta-learner. Under the guidance of TWA, the learning process of base-learners becomes more efficient and specialized, demonstrating their unique nature of "adaptation". Whereas, such a solution is insufficient for the scenarios with few data attainable (e.g., 1 shot or 5 shots), because base-learners may fail to analyze the critical classification information from a global and comprehensive perspective, ultimately tending to overfit during the meta-training phase. For this purpose, we designed a part complementary learning (PCL) module in the form of dual-branch network, incorporating and enriching the formerly extracted features. PCL is not limited to recognition from the most discriminative object parts, which often incline to be local and incomplete. On the contrary, the proposed module efficiently leverages complementary information generated by the new branch to assist classification process from a global perspective, exhibiting the "generalization" characteristic of base-learners. Compared to conventional meta-learning algorithms, our proposed model (called TP-Net) can exploit and explore the feature map in-depth, and further, achieve a tradeoff between rapid

"adaptation" and inductive "generalization".

In summary, our study aims to contribute to the growing area of few-shot learning by exploring the modification of meta-learning algorithm, which can be stated out from three aspects:

• First, we focus on the research of base-learners in meta-learning framework, which is crucial but rarely investigated.

• Second, TWA and PCL modules are integrated into base models, achieving fast adaptation and inductive generalization during meta-testing phase.

• Third, our proposed end-to-end meta-learning network, called TPNet, demonstrates superior performance compared to the state-of-the-art approaches on benchmark few-shot datasets, e.g., miniImageNet and CUB-200.

2 Related work

For the past few years, meta-learning algorithms have received widespread attention from scholars due to their universality and effectiveness in the few-shot learning scenarios. By leveraging the base datasets with sufficient samples, a capable meta-learner can be developed in the meta-training phase, and then provides prior knowledge (also known as meta-knowledge) for base-learners, specifically guiding the adaptation process under different task requirements. In general, most representative few-shot learning approaches based on meta-learning can be organized into four categories as follows.

Model-based meta-learning methods aim to quickly update the parameters on a small number of samples through the design of model structure, directly establishing the mapping function of input variables and predicted results. Santoro et al. [26] proposed to utilize memory-augmented neural network to tackle few-shot learning task, exhibiting high-capable of acquiring data knowledge and accurate prediction based on small samples from new categories. The rapid generalization characteristic of meta network [34] derives from its unique "fast weight" mechanism, which is established with the gradient provided during training phase as a medium. In general, the proposed meta-model can be conceptually summarized into two parts, namely meta-learner and base-learner. Meta-learner learns to seek commonalities among various meta-tasks, and then store the extracted information under the guidance of memory strategy; while base-learner adapts quickly to new tasks with the gradient information (i.e., meta-information) anteriorly generated by meta-learner, conclusively producing predictive outputs.

Metric-based meta-learning approaches [22, 23, 35–39] try to learn a mapping function from source domain to embedding space, where a fundamental principle of "clustering" is established. In other words, images belonging to the same category are closer together while instances of different categories are separated. Consequently, the differences between sample features can be intuitively reflected in the form of distance in embedded space, and this unique property is expected to be equally applicable to unseen categories. Sung et al. [35] proposed an efficient feature extraction network to capture the information of query samples and labeled support samples, then measured the difference of feature embeddings through a well-trained relation module. Snell et al. [22] proposed to obtain a prototype representation for each category, which mathematically refers to the mean of labeled support data in embedding space, and then calculated the Euclidean distance between query samples and each prototype to determine similarity measurement. Wang et al. [36] proposed that the attention information of images can be extracted from category tags. The authors designed a novel network architecture for generating attention maps based on semantic embedding classification tags and utilized these attention maps to create discriminative features for classification.

Optimization-based meta-learning techniques [24–32] take into account the inadaptability of conventional gradient descent algorithm to few-shot learning scenarios. To this end, this type of approach raises the level of significant factors that affect the optimal parameter determination, such as initialization and optimizer. The purpose of [27] is to learn good optimization or good initialization during meta-training phase, resulting valid generalization of the base-learner on novel database. The goal of [25] is to capture the generalization ability of optimization algorithm through several iterative steps, so that meta-learner can guide the base-learner to converge to the approximate optimal solution on each task. TAML algorithm [29] proposed to meta-train an unbiased initial model by avoiding overperforming on some specific tasks or directly minimizing the inequality of measures. Even if facing a new task, starting from the determined initial parameters, an optimal model can still be rapidly iterated with only a few training samples. MetaOptNet [32] proposed to utilize linear predictors as base-learner to obtain

representations and proved that feature embeddings can be universally generalized under the linear classification rules for novel categories. Meta-SGD [28] introduced a meta-learning algorithm similar to SGD, which can be regarded as an extension of [27], not only learning initialization parameters of base-learners, but also learning update strategies (such as learning rate and update direction).

Hallucination-based meta-learning. At present, the biggest challenge in few-shot learning scenarios is consistently the shortage of training samples, and therefore how to acquire more informative and meaningful labeled data is the core of issue. The hallucination based approaches [40–43] use a small number of labeled samples to generate more hallucination data, with the goal of achieving a robust and powerful network. Zhang et al. [40] pretrained a saliency network to segment foreground and background of available image samples and generated additional data by combining the appropriate foreground-background pairs. Alfassy et al. [41] proposed to perform labeled-set operations (intersection, union, and subtraction) on multi-label samples in the embedded space. For example, by solving the difference set between the images containing man and sheep, and the images containing sheep in embedded space, the representation of images containing only man can be obtained. The authors utilized these operations to generate more hallucination data for training phase, and thus significantly improved classification performance. Chen et al. [42] proposed to relate new concepts to the existing ones in semantic space, and leveraged the relationships to generate new additional samples by interpolating among concepts, further facilitating learning.

3 The proposed method

In this part, we explain the proposed task-wise attention guided part complementary learning method (TPNet) at length. In Subsection 3.1, we first describe the problem definitions and data notations used in this study. The introduction of meta-learner is presented in Subsection 3.2. We proceed with explaining the TWA module and the PCL module in Subsections 3.3 and 3.4. The implementation details of our model are given in Subsection 3.5. Furthermore, the architecture of the proposed TPNet at different stages is illustrated respectively.

3.1 Problem definition and notations

Given two subsets $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ of dataset $\mathcal{C}$, we have the following properties: $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \emptyset$ and $\mathcal{C}_{\text{base}} \cup \mathcal{C}_{\text{novel}} = \mathcal{C}$.

As for the former subset, let $\mathcal{C}_{\text{base}}$ denote the base set that contains a certain number of different image classes with all labeled samples per class, and thus meta-learner can acquire transferable knowledge under the supervision of $\mathcal{C}_{\text{base}}$.

As for the latter subset, the authors randomly select several tasks (also known as episodes) from $\mathcal{C}_{\text{novel}}$ to constitute the sets of novel class, and each task involves the samples from n categories. Further, $(k+m)$ samples are extracted from each category to form the support set and query set, defined as $\mathcal{I}_{\text{su}} = \{(x_i^j, y_i)|i = 0, 1, \ldots, n; j = 0, 1, \ldots, k\}$ and $\mathcal{I}_{\text{qu}} = \{(x_i^j, y_i)|i = 0, 1, \ldots, n; j = 0, 1, \ldots, m\}$ respectively, where x denotes a data point and y is the corresponding label.

Through the preceding steps, a series of n-way k-shot tasks are determined for experiment. The primary purpose of few-shot image recognition is to identify all samples in the query set $\mathcal{I}_{\text{qu}}$ according to the support set $\mathcal{I}_{\text{su}}$, realizing rapid adaption and accurate prediction.

In this paper, the application of few-shot learning in classification task can be divided into two phases: meta-training phase and meta-testing phase. To be more specific, in the meta-training phase, we use the entire base class samples of $\mathcal{C}_{\text{base}}$ to train a well-generalized meta-learner (or backbone network) that provides meta-knowledge for subsequent recognition tasks of unseen categories. During the meta-testing phase, we utilize the support set $\mathcal{I}_{\text{su}}$ with only few data to adapt and generalize the base-learners, and finally evaluate the efficiency of proposed model on query set $\mathcal{I}_{\text{qu}}$.

3.2 Meta-knowledge acquisition

It is of crucial importance to provide base-learners with extensive but valuable meta-knowledge, which is closely related to the training process of meta-learner.

Generally speaking, the major objective of feature extraction network in the low-level layer is to extract common features, such as color, edge, and texture, while in the high-level layer is commonly the

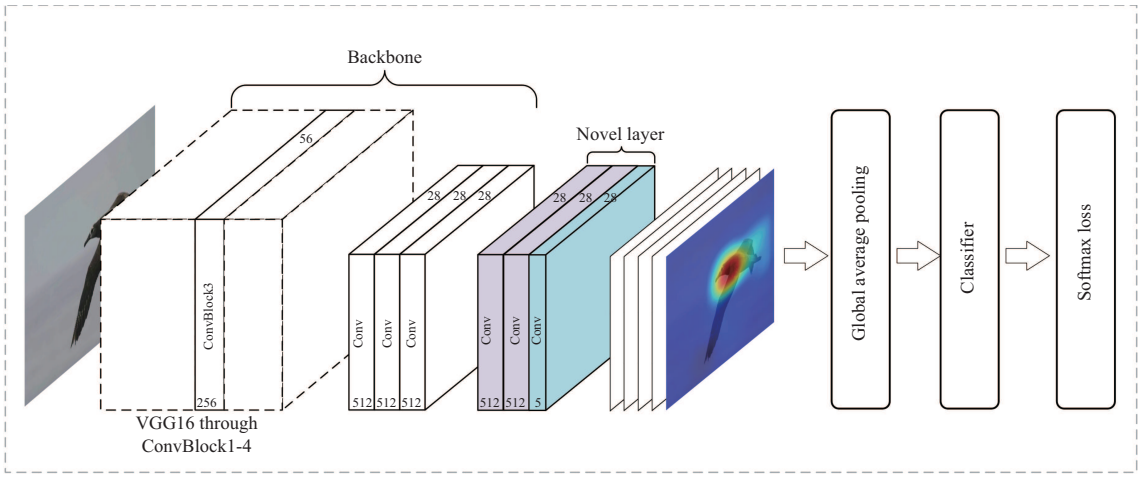

Figure 1 (Color online) Illustration of the classification network utilized in our method during meta-training phase. The network framework includes backbone network, novel layer, global average pooling layer, and a classifier. "ConvBlock3" represents the first three blocks of VGG Net. The number at the top (e.g., "28") denotes the spatial size of feature map, and that at the bottom (e.g., "512") denotes the number of channels.

category relevant information. Therefore, the authors hope that meta-learners in our work can learn transferable knowledge and appropriately retain the ability to capture general representations as they identify new categories, thereby facilitating the extraction of high-quality features. Additionally, the knowledge information acquired from support samples is expected to be correlated with the category information, making the learning content more specific and meaningful.

To achieve the desired purpose above, an efficient feature extraction network is designed that can be considered as a variant of VGG Net. The authors add a "novel layer" to the first five blocks of VGG Net-D (see [8]), which contains three convolutional operations with different kernel size. The kernel size of first two layers is set to 3, and that of the last layer is set to 1. Global average pooling (GAP) and softmax layers are then added on the top of the model.

Note that the first five blocks of VGG Net-D are served as backbone network f_θ in our study, which is pre-trained on the base set $\mathcal{C}_{\text{base}}$ for learning prior knowledge (see Figure 1). The parameters of backbone network are fixed during the meta-testing phase to ensure the effective utilization of acquired meta-knowledge.

3.3 Task-wise attention module

The main objective of meta-learning approach is to efficiently utilize the prior knowledge provided by meta-learner, achieving rapid adaption and inductive generalization in the face of new tasks with only few samples available. However, there are significant differences among meta-tasks, and the ability of meta-learner to extract features under different requirements is limited. For this purpose, a task-specific feature extraction module named TWA is introduced in this work, which learns to find the most representative features associated with current task by reinforcing or suppressing a portion of the knowledge provided by meta-learner.

As demonstrated by the dotted box in Figure 2, $\mathcal{F}_m$ is the output of backbone network with size $C \times H \times W$. We first squeeze the information of each feature map through GAP layer, which can be formulated as

$$s_k = \frac{\sum_{i,j} S_{i,j}^k}{H \times W}, \tag{1}$$

where $S_{i,j}^k$ denotes the element of the kth feature map at the ith row and the jth column. The overall description with C channels is also obtained by $s = [s_1, s_2, \ldots, s_C]$, which includes the global representation of image.

Two fully connected (FC) layers are then added to capture task-specific discriminative information. The number of output channels of the second FC layer is set to be the same as the number of categories n, realizing correspondence between features and categories. The above operations can be summarized as follows:

$$u_{\text{A}} = \text{W}_2(\text{W}_1(s)), \tag{2}$$

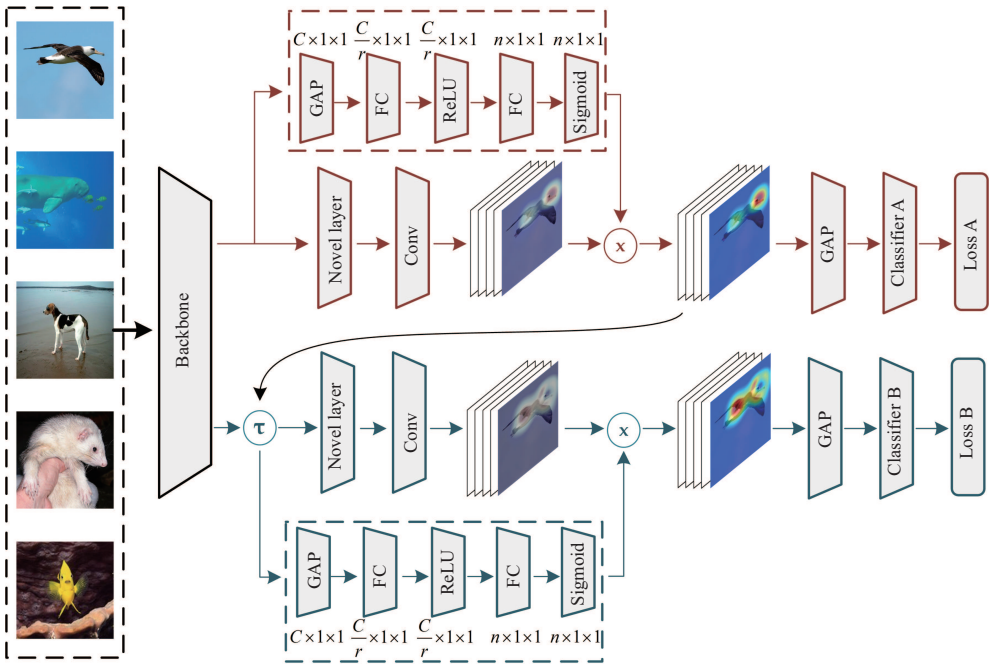

Figure 2 (Color online) Illustration of the proposed TPNet for 5-way 1-shot task in training phase. As shown, the TPNet framework consists of a task-wise attention module for increasing the sensitivity of network to discriminative information and a part complementary learning module for learning complementary descriptions. τ indicates the "erasing" operation of PCL module according to a pre-defined threshold.

where $\mathrm{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathrm{W}_2 \in \mathbb{R}^{n \times \frac{C}{r}}$ refer to the weight matrices of FC layers. The first FC layer is followed by a ReLU activation function.

In order to prevent excessive enhancement and suppression, the sigmoid function σ is adopted to normalize the weight (or response) of each feature map

$$u'_{\mathrm{A}} = \sigma(u_{\mathrm{A}}). \tag{3}$$

Finally, TWA module is applied to the branch A through channel-wise multiplication between u'_{A} and feature maps $\mathcal{F}_{\mathrm{ha}}$ extracted by the novel layer:

$$\mathcal{F}'_{\mathrm{ha}} = u'_{\mathrm{A}} \odot \mathcal{F}_{\mathrm{ha}}, \tag{4}$$

where $\mathcal{F}_{\mathrm{ha}} \in \mathbb{R}^{n \times H \times W}$ and $u'_{\mathrm{A}} \in \mathbb{R}^{n \times 1 \times 1}$. $\odot$ is the element-wise product.

As for branch B, the acquisition of channel weight u'_{B} is the same as that of branch A. The difference between these two branches lies in the input feature representations of TWA module, which is elaborated in Subsection 3.4. Similarly, the task-specific features extracted by the B branch can be formulated by

$$\mathcal{F}'_{\mathrm{hb}} = u'_{\mathrm{B}} \odot \mathcal{F}_{\mathrm{hb}}, \tag{5}$$

where $\mathcal{F}_{\mathrm{hb}}$ denotes the feature maps processed by the novel layer, sharing the same size as $\mathcal{F}_{\mathrm{ha}}$.

To sum up, TWA module improves the ability of base-learners to capture discriminative information or suppress irrelevant information. The connection between extracted features and categories is successfully established, facilitating the subsequent refinement and classification stages by the high-level semantic information acquired.

3.4 Part complementary learning module

Generally speaking, classification networks tend to identify local regions from the most discriminative object parts rather than from a global perspective for recognition, which leads to incomplete feature representations. To this end, the authors designed a PCL module in the form of dual-branch network, incorporating and enriching the formerly extracted features.

In the training phase, the PCL module consists of two branches A and B (see Figure 2), capturing the discriminative and complementary information respectively. Let $\mathcal{F}_m$ denote the feature extraction of

backbone network, which is served as the input of branch A. Then, the proposed novel layer is applied to $\mathcal{F}_m$ to obtain the task-specific representations $\mathcal{F}'_{\rm ha}$ with n channels. Followed by GAP and softmax layers, the classification loss of branch A is subsequently determined.

As described in Subsection 3.3, the feature representation has a certain corresponding relationship with the category information (n-dimensional vs. n-category) after passing through the novel layer, and the larger response value of which contains richer information of target category. Therefore, the authors extract the channel with the largest response among n feature representations as the object activation map. A pre-defined threshold τ is then served as the criterion to determine unique and discriminative regions in target feature map. Finally, "erasing" the corresponding area in $\mathcal{F}_m$, and take the processed feature maps $\mathcal{F}'_m$ as the input of branch B. It is worth noting that f_θ and $\mathcal{F}'_{\rm ha}$ have the same spatial dimensions, reflecting the feasibility of "erasing" operation. Given the input feature maps $\mathcal{F}'_m$ of branch B, the complementary information of other significant regions $\mathcal{F}'_{\rm hb}$ can be acquired through the novel layer designed, just like branch A. Thus, the sum of classification loss during the training phase can be calculated by the following equation:

$$\text{Loss} = \text{Loss}_{\rm A} + \lambda \cdot \text{Loss}_{\rm B}, \tag{6}$$

where

$$\begin{aligned} \text{Loss}_{\rm A} &= \mathcal{L}(f_\alpha(\mathcal{F}_m), y_i), \\ \text{Loss}_{\rm B} &= \mathcal{L}(f_\beta(\mathcal{F}_m \odot \text{mask}), y_i); \end{aligned}$$

f_α and f_β represent feature extraction of branches A and B respectively; λ is the loss weight of branch B; mask denotes the indicator of "erasing" operation that is determined by τ. The first half of (6) represents the classification loss of branch A, and the second half is the classification loss of branch B. In Subsection 4.3, we conduct a comprehensive ablation experiments for the values of parameters λ and τ.

In the testing phase, as shown in Figure 3, a max fusion module is added at the top of the network with the goal of obtaining complete and adequate representations, which integrate the information of branches A and B according to the following formula:

$$\mathcal{F}_h = \max(\mathcal{F}'_{\rm ha}, \mathcal{F}'_{\rm hb}). \tag{7}$$

The fused features increase the weight (or response) of discriminative regions to some extent, reflecting the crucial task-specific information from a global perspective.

3.5 Implementation details

In the meta-training phase, we apply the same hyper-parameter settings for the miniImageNet dataset and CUB dataset. The network of CUB dataset is fine-tuned on the pre-trained weights of ILSVRC [44] and that of miniImageNet is trained from scratch. In our experiments, all input images of miniImageNet dataset and CUB dataset are resized to 224×224 pixels for training and testing. The framework of classification model adopted at this stage is shown in Figure 1.

In the meta-testing phase, a support set $\mathcal{I}_{\rm su}$ and a query set $\mathcal{I}_{\rm qu}$ are fed to the base-learner for each experiment. We randomly sample n novel classes, and then select k support instances and 15 query instances from each class. The performance of proposed TPNet is evaluated under both 5-way 1-shot and 5-way 5-shot settings, and the results presented are the average of 500 independent trials.

Note that we utilize the SGD optimizer with Nesterov momentum for parameter updating, and the learning rate adopted for training in each dataset is different, more specifically, 0.01 for the miniImageNet dataset and 0.001 for the CUB dataset. Our source code is available at the website[1)].

4 Experiments

In this section, we introduce in detail two public few-shot learning datasets, and conduct a series of experiments on them to evaluate the proposed TPNet. The experimental results are then compared with current state-of-the-art results under both 5-way 1-shot and 5-way 5-shot settings. Finally, comprehensive ablation experiments are performed to investigate the efficiency of TWA and PCL modules.

1) https://github.com/lrmek/TPNet.

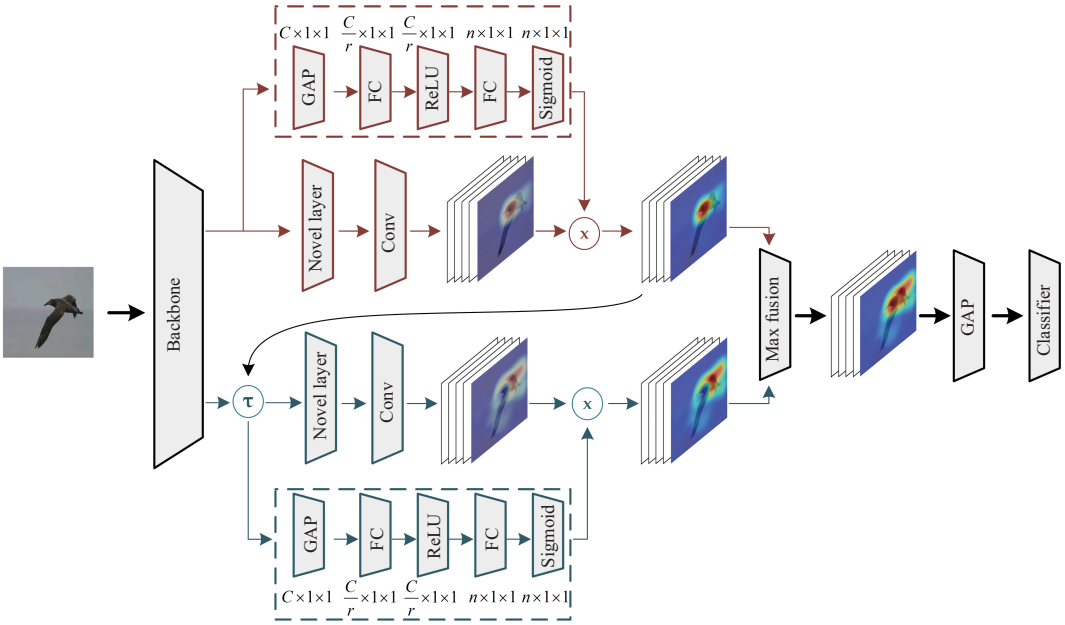

Figure 3 (Color online) Illustration of the proposed TPNet for 5-way 1-shot task in testing phase. As shown, part complementary learning module is composed of two branches for learning discriminative and complementary features, respectively. The max fusion module can obtain multiple representative features by integrating the information of two branches. τ indicates the "erasing" operation of PCL module according to a pre-defined threshold.

4.1 Datasets

miniImageNet. This dataset [23] is originally used for the task of few-shot image recognition, including 100 classes with a total number of 60000 images. Following the splits used in [25], we select 64, 16, and 20 classes for training, validation, and testing respectively.

CUB-200. This dataset [45] is originally used for the task of fine-grained image classification, including 200 bird species with a total number of 11788 images. Following the evaluation protocol of [46], we randomly select 100, 50, and 50 classes for training, validation, and testing respectively.

4.2 Comparison with state-of-the-arts

In this subsection, we conduct a series of experiments on the above datasets according to the n-way k-shot classification task. Both 5-way 1-shot and 5-way 5-shot experimental settings are taken into consideration. The comparison results with state-of-the-art approaches for each dataset are presented in Tables 1 and 2, respectively.

As for the miniImageNet dataset, the proposed TPNet is compared with numerous types of few-shot learning techniques, such as prototypical net [22] (metric-based), MAML [27] (optimization-based), meta net [34] (model-based), and "spot and learn" [43] (hallucination-based). It can be found from Table 1 that TPNet exhibits superior performance over other state-of-the-art approaches under both 1-shot and 5-shot settings, especially in the latter case is more competitive. The reason for this phenomenon is that the available information for each category becomes richer as the number of support set increases. Compared with other few-shot learning approaches, the proposed TPNet exhibits a more efficient utilization of the information provided by support set. By capturing discriminative and complementary task-related information, higher recognition accuracy is eventually achieved under 5-shot setting. Specifically speaking, our approach gains 8.19%, 7.2%, and 3.71% improvements over DN4 net, Saliency hallucination, and MTL respectively, demonstrating the superiority in few-shot classification tasks. Compared with [22,23], our approach adopts a deeper backbone network (VGG Net) for feature extraction instead of only 4-layer shallow network, facilitating the acquisition of high-level information. However, the utilization of deep backbone network also leads to the increased risk of overfitting, which is the primary reason why deprecated by previous work. In the proposed TPNet, the above drawbacks are overcome from two aspects. On one side of the spectrum, the non-parametric GAP layer in backbone network and TWA module is utilized to replace the FC layer and capture task-specific representations respectively, which significantly

Table 1 Few-shot classification accuracy on miniImagNet dataset with 95% confidence intervals[a)]

Model	1-shot accuracy (%)	5-shot accuracy (%)
Meta net [34]	49.21 ± 0.96	–
Matching net [23]	46.6	60.0
Prototypical net [22]	49.42 ± 0.78	68.20 ± 0.66
Relation net [35]	50.44 ± 0.82	65.32 ± 0.70
DN4 net [38]	51.24 ± 0.74	71.02 ± 0.64
EGNN+transduction [47]	–	76.37
MAML [27]	48.70 ± 1.84	63.11 ± 0.92
MTL [31]	61.2 ± 1.8	75.5 ± 0.8
LR-D2 [21]	51.9 ± 0.2	68.7 ± 0.2
Spot and learn [43]	51.03 ± 0.78	67.96 ± 0.71
Saliency hallucination [40]	57.45 ± 0.88	72.01 ± 0.67
TPNet	**59.31** ± 0.99	**79.21** ± 0.64

a) Both 5-way 1-shot and 5-way 5-shot experimental settings are taken into consideration. The best results are presented in boldface. '–' indicates not reported.

Table 2 Few-shot classification accuracy on CUB dataset with 95% confidence intervals[a)]

Model	1-shot accuracy (%)	5-shot accuracy (%)
Matching net* [23]	61.16 ± 0.89	72.86 ± 0.70
Prototypical net* [22]	51.31 ± 0.91	70.77 ± 0.69
Relation net* [35]	62.45 ± 0.98	76.11 ± 0.69
DN4-DA net [38]	53.15 ± 0.84	81.90 ± 0.60
MAML* [27]	55.92 ± 0.95	72.09 ± 0.76
Baseline++ [48]	60.53 ± 0.83	79.34 ± 0.61
DeepEMD [49]	75.65 ± 0.83	88.69 ± 0.50
FEAT [50]	68.87 ± 0.22	82.90 ± 0.15
DPGN [51]	75.71 ± 0.47	91.48 ± 0.33
MACO [46]	60.76	74.96
Multiple-semantics [52]	76.1	82.9
TPNet	**77.30** ± 0.86	**94.20** ± 0.34

a) Both 5-way 1-shot and 5-way 5-shot experimental settings are taken into consideration. The best results are presented in boldface. '*' indicates the results reported by [48].

reduces the network parameters and prevents over-fitting in few-shot scenarios. On the other side of the spectrum, the "erasing" operation in PCL module can be regarded as a "Dropout" strategy with remarkable learning ability, which generates a mask according to the given threshold τ, thereby inactivating some neurons of the feature maps extracted by backbone network to realize generalization. To sum up, our proposed TPNet adapts to the deeper backbone network structure by virtue of its unique properties, and leverages the captured high-level discriminative features to assist recognition task, thus achieving advanced performance on current dataset.

In general, the recognition task on fine-grained datasets is more complex than that on standard datasets due to the characteristic of smaller inter-class variation and larger intra-class variation. Therefore, we also conduct fine-grained few-shot classification experiments on CUB dataset to comprehensively evaluate the robustness and efficiency. Seven state-of-the-art approaches are selected for comparison, namely matching net* [23], prototypical net* [22], relation net* [35], MACO [46], DN4-DA net [38], MAML* [27], and Baseline++ [48]. Especially, matching net*, prototypical net*, relation net*, MAML*, and Baseline++ are conducted by Chen et al. [48] with conv-4 backbone and data augmentation operation, where * denotes the re-implementation of original approach on CUB dataset, and "conv-4" represents a four-layer convolutional network. The backbone network of DN4-DA net denoted as Conv-64f is composed of four convolutional blocks, each of which has convolutional layer, batch normalization layer, and leaky ReLU layer. It can be seen from Table 2 that our TPNet has achieved higher recognition accuracy compared with the performance of above-mentioned approaches in both 1-shot and 5-shot classification tasks. Especially under the 5-way 1-shot setting, the proposed model gains 24.15%, 16.54%, and 16.77% improvements over DN4-DA net, MACO, and Baseline++, respectively. The utilization of TWA and PCL modules facilitates base-learners to obtain task-specific discriminative information and prevent incomplete feature representation, thus exhibiting strong robustness on fine-grained recognition tasks with smaller inter-class

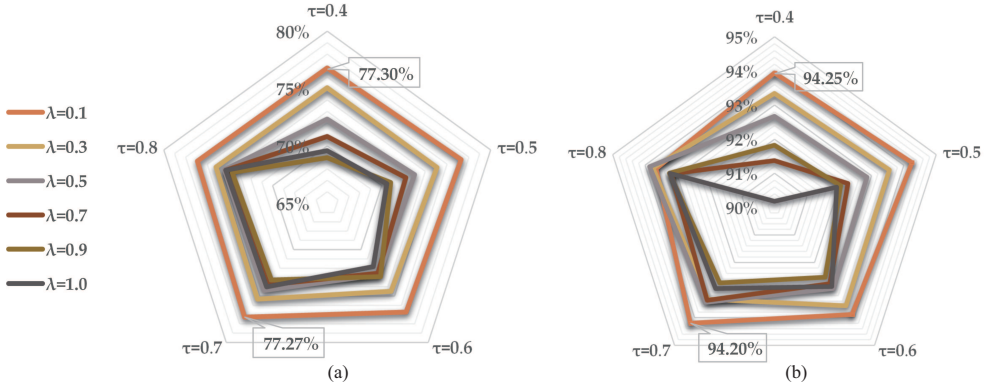

Figure 4 (Color online) The influence of parameters λ and τ on CUB dataset. (a) corresponds to the 5-way 1-shot setting, while (b) corresponds to the 5-way 5-shot setting. It is obvious that when $\lambda = 0.1$ and $\tau = 0.5$, the proposed TPNet can achieve the best performance in both 5-way 1-shot and 5-way 5-shot settings.

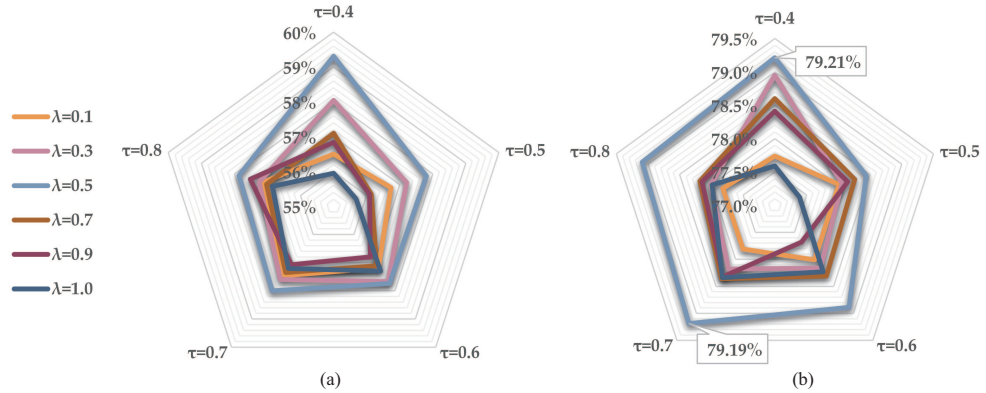

Figure 5 (Color online) The influence of parameters λ and τ on miniImageNet dataset. (a) corresponds to the 5-way 1-shot setting, while (b) corresponds to the 5-way 5-shot setting. It is obvious that when $\lambda = 0.5$ and $\tau = 0.4$, the proposed TPNet can achieve the best performance in both 5-way 1-shot and 5-way 5-shot settings.

and larger intra-class variations.

4.3 Ablation study

4.3.1 *The influence of parameters λ and τ*

In PCL module, the complementary information of interest is obtained via setting the corresponding value of the input feature maps of branch B to zero according to the given threshold τ. In addition, we introduce a weight parameter λ for branch B in the loss function to coordinate the contribution of discriminative feature and complementary feature to classification accuracy. The values of these two hyper-parameters are both explored due to their impact on performance, and the experimental results under different settings are presented in Figures 4 and 5.

Compared with CUB dataset with bird species only, the images in miniImageNet dataset have different object categories and contain complex background information. Therefore, a larger loss weight λ of the second branch is adopted on miniImageNet dataset than that on CUB dataset, so that the complementary features of target object in complex background can be deeply explored. Furthermore, due to the large difference of samples in various datasets, the criteria for distinguishing the significance of their features are not the same, which ultimately leads to distinct determinations of the optimal threshold τ.

4.3.2 *The influence of TWA and PCL modules*

In order to investigate the influence of the proposed TWA and PCL modules, a series of ablation experiments are conducted based on the baseline model, and the results are shown in Tables 3 and 4. To be more specific, the authors first introduce the baseline and various modules in detail, then compare the

Table 3 Comparison of the proposed TPNet model under various configurations on miniImageNet with 95% confidence intervals[a)]

Model	PCL	EB	TWA	1-shot accuracy (%)	5-shot accuracy (%)
0	✘	✘	✘	56.75 ± 0.89	77.22 ± 0.66
1	✔	✘	✘	56.87 ± 0.92	78.25 ± 0.64
2	✔	✔	✘	56.92 ± 0.90	78.62 ± 0.65
3	✔	✘	✔	**59.31** ± 0.99	**79.21** ± 0.64
4	✔	✔	✔	58.59 ± 0.91	78.86 ± 0.64

a) Both 5-way 1-shot and 5-way 5-shot experimental settings are taken into consideration, and the best results are presented in boldface. "Model 0" is the baseline model. We can find that the model achieves optimal performance under the 3rd configuration.

Table 4 Comparison of the proposed TPNet model under various configurations on CUB with 95% confidence intervals[a)]

Model	PCL	EB	TWA	1-shot accuracy (%)	5-shot accuracy (%)
0	✘	✘	✘	74.81 ± 0.88	92.61 ± 0.35
1	✔	✘	✘	75.61 ± 0.90	93.60 ± 0.36
2	✔	✔	✘	75.69 ± 0.90	93.55 ± 0.36
3	✔	✘	✔	**77.30** ± 0.86	**94.20** ± 0.34
4	✔	✔	✔	76.40 ± 0.86	93.85 ± 0.38

a) Both 5-way 1-shot and 5-way 5-shot experimental settings are taken into consideration, and the best results are presented in boldface. "Model 0" is the baseline model. We can find that the model achieves optimal performance under the 3rd configuration.

classification results of algorithms under different configurations, and finally analyze the effect of each component.

For the baseline model, we remove the layers after "ConvBlock5" of VGG Net, then follow the setup of [53] to add two convolutional layers with 512 filters (kernel size 3×3, stride 1, pad 1) and one convolutional layer with 512 filters (kernel size 1×1, stride 1, pad 1). The framework of baseline model is shown in Figure 1, which can only obtain incomplete feature representation without the guidance of TWA and PCL modules. As described in Subsection 3.4, the discriminative and complementary information is expected to be captured by branches A and B respectively, and the "erasing" operation is performed only once during this process. Therefore, in the ablation experiment, we also add an extra branch (denoted as "EB") to the PCL module to explore whether the synergy of multiple complementary branches would benefit the classification results.

From the experimental results given in Tables 3 and 4, we can observe that the introduction of PCL module improves classification accuracy to a certain extent. For example, on the CUB dataset, "Model 1" achieved 0.8% and 0.99% improvements under the settings of 5-way 1-shot and 5-way 5-shot respectively, indicating the crucial impact of complementary features on recognition task. However, when PCL module was extended to three branches, we found that multiple additional information brings no significant improvement on performance, even degradation when combined with TWA module, as can be seen from "Model 2" and "Model 4". The reason for above phenomenon is that under the guidance of attention mechanism, the ability of network to capture and analyze the most crucial information of current task is significantly improved, so the addition of more branches may lead to over-representation instead. Therefore, dual-branch network structure can be considered as the optimal choice for the base-learners, which can balance the significant information and additional information, ultimately realizing high-precision identification.

Furthermore, comparing the experimental results of "Model 1" and "Model 3", it can be found that by enhancing the adaptability of base-learners to few-shot learning tasks, the recognition accuracy on miniImageNet dataset is improved by 2.44% (1-shot) and 0.96% (5-shot), and that on CUB dataset is improved by 1.69% and 0.6%, respectively. These promising results more intuitively illustrate the necessity of capturing task-specific image features and the effectiveness of TWA strategy.

4.3.3 *Visualization*

In order to more intuitively illustrate the impact of our proposed TWA and PCL modules on classification results, the authors investigated the visualization features in each case, as shown in Figure 6. Note that the performance is evaluated by the coverage of target region and the corresponding color shade. It can be found that branches A and B extract the discriminative and complementary features respectively, which are consistent with our original intention of design.

Taking the sample in the 4th column as an example, the body characteristics of bird extracted by only

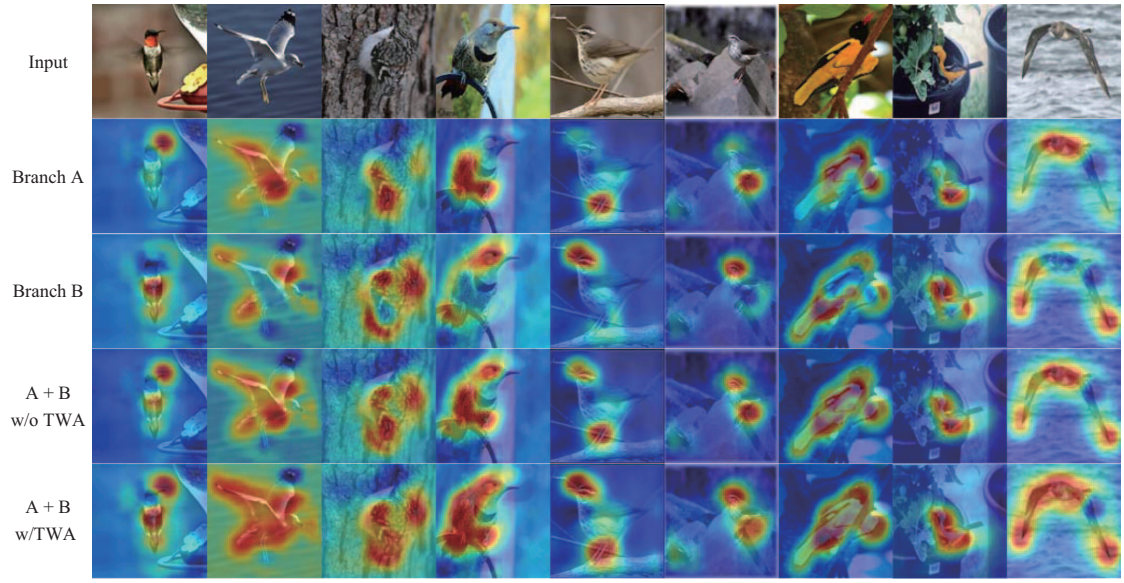

Figure 6 (Color online) Visualization of the proposed method. Red regions represent the positive region in favor of classification result, while blue regions are negative regions that reduce the recognition confidence. From top to bottom: input images; branch A; branch B; the feature fusion for branches A and B, without TWA module; the feature fusion for branches A and B, with TWA module. In the 5th row, comprehensive feature representations are acquired through the information fusion of two branches and the TWA module introduced. (Best viewed in color.)

one branch have certain discriminability, but are not complete (see 2nd row); while after adding PCL strategy, the visualization results further covered the characteristics of head and tail well (see 3rd row), thus increasing the comprehensiveness of sample representation. In addition, by comparing the results of the 4th and 5th rows, it can be noticed that the introduction of TWA strategy facilitates base-learners to explore the feature maps in depth, and the positive regions contribute more to the classification results.

5 Conclusion

In this paper, we presented a robust and flexible meta-learning framework for few-shot image classification task. Task-wise attention and part complementary learning modules are integrated into the base-learner, with the goal of realizing fast adaptation and inductive generalization during meta-testing phase. More specifically, the former module improves the ability to capture task-specific image features through attention mechanism, while the latter module facilitates the acquisition of comprehensive representations through additional information provided by complementary branch. A series of ablation experiments and comparative experiments have verified the effectiveness of proposed TPNet, that achieves state-of-the-art performance on several challenging datasets.

In future work, we will attempt to explore the impact of meta-knowledge acquired through various backbone networks (e.g., residual networks) on recognition performance. The post-processing methods of information extracted by each branch will also be further investigated, such as utilizing "concatenate + Conv" operation to fuse feature representations in a softer manner.

Acknowledgements This work was supported by Science, Technology and Innovation Commission of Shenzhen Municipality (Grant No. JCYJ20180306171131643) and National Natural Science Foundation of China (Grant No. 61772425).

References

1 Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of Advances in Neural Information Processing Systems, 2015. 91–99
2 Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2117–2125
3 Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 779–788
4 Cheng G, Zhou P C, Han J W. RIFD-CNN: rotation-invariant and fisher discriminative convolutional neural networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2884–2893
5 Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Proceedings of European Conference on Computer Vision, 2016. 21–37

6 Cheng G, Han J, Zhou P, et al. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE Trans Image Process, 2019, 28: 265–278
7 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
8 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
9 Cheng G, Yang C Y, Yao X W, et al. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE Trans Geosci Remote Sens, 2018, 56: 2811–2821
10 Cheng G, Gao D C, Liu Y, et al. Multi-scale and discriminative part detectors based features for multi-label image classification. In: Proceedings of International Joint Conference on Artificial Intelligence, 2018. 649–655
11 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3431–3440
12 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision, 2015. 1520–1528
13 Wang N, Ma S H, Li J Y, et al. Multistage attention network for image inpainting. Pattern Recogn, 2020, 106: 107448
14 Song L C, Wang C, Zhang L F, et al. Unsupervised domain adaptive re-identification: theory and practice. Pattern Recogn, 2020, 102: 107173
15 Wei X S, Wang P, Liu L Q, et al. Piecewise classifier mappings: learning fine-grained learners for novel categories with few examples. IEEE Trans Image Process, 2019, 28: 6116–6125
16 Ji Z, Chai X L, Yu Y L, et al. Improved prototypical networks for few-shot learning. Pattern Recogn Lett, 2020, 140: 81–87
17 Ji Z, Sun Y X, Yu Y L, et al. Attribute-guided network for cross-modal zero-shot hashing. IEEE Trans Neur Netw Lear Syst, 2020, 31: 321–330
18 Wang Y Q, Yao Q M, Kwok J T, et al. Generalizing from a few examples: a survey on few-shot learning. 2019. ArXiv:1904.05046
19 Ji Z, Yan J T, Wang Q, et al. Triple discriminator generative adversarial network for zero-shot image classification. Sci China Inf Sci, 2021, 64: 120101
20 Vilalta R, Drissi Y. A perspective view and survey of meta-learning. Artif Intell Rev, 2002, 18: 77–95
21 Bertinetto L, Henriques J F, Torr P H, et al. Meta-learning with differentiable closed-form solvers. In: Proceedings of International Conference on Learning Representations, 2019. 1–15
22 Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 4077–4087
23 Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 3630–3638
24 Andrychowicz M, Denil M, Gomez S, et al. Learning to learn by gradient descent by gradient descent. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 3981–3989
25 Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proceedings of International Conference on Learning Representations, 2017. 1–11
26 Santoro A, Bartunov S, Botvinick M, et al. Meta-learning with memory-augmented neural networks. In: Proceedings of the 33rd International Conference on Machine Learning, 2016. 1842–1850
27 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017. 1126–1135
28 Li Z G, Zhou F W, Chen F, et al. Meta-SGD: learning to learn quickly for few-shot learning. 2017. ArXiv:1707.09835
29 Jamal M, Qi G J. Task agnostic meta-learning for few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 11719–11727
30 Zhou F W, Wu B, Li Z G. Deep meta-learning: learning to learn in the concept space. 2018. ArXiv:1802.03596
31 Sun Q R, Liu Y Y, Chua T, et al. Meta-transfer learning for few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 403–412
32 Lee K, Maji S, Ravichandran A, et al. Meta-learning with differentiable convex optimization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 10657–10665
33 Lifchitz Y, Avrithis Y, Picard S, et al. Dense classification and implanting for few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 9258–9267
34 Munkhdalai T, Yu H. Meta networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017. 2554–2563
35 Sung F, Yang Y X, Zhang L, et al. Learning to compare: relation network for few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1199–1208
36 Wang P, Liu L Q, Shen C H, et al. Multi-attention network for one shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2721–2729
37 Li W B, Xu J L, Huo J, et al. Distribution consistency based covariance metric networks for few-shot learning. Assoc Adv Artif Intell, 2019, 33: 8642–8649
38 Li W B, Wang L, Xu J L, et al. Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7260–7268
39 Li H Y, Eigen D, Dodge S, et al. Finding task-relevant features for few-shot learning by category traversal. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1–10
40 Zhang H G, Zhang J, Koniusz P. Few-shot learning via saliency-guided hallucination of samples. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 2770–2779
41 Alfassy A, Karlinsky L, Aides A, et al. LaSO: label-set operations networks for multi-label few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6548–6557
42 Chen Z T, Fu Y W, Wang Y X, et al. Image deformation meta-networks for one-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8680–8689
43 Chu W H, Li Y J, Chang J C, et al. Spot and learn: a maximum-entropy patch sampler for few-shot image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6251–6260
44 Bearman A, Russakovsky O, Ferrari V, et al. What's the point: semantic segmentation with point supervision. In: Proceedings of the 14th European Conference on Computer Vision, 2016. 549–565
45 Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset. 2011. https://authors.library.caltech.edu/27452/

46 Hilliard N, Phillips L, Howland S, et al. Few-shot learning with metric-agnostic conditional embeddings. 2018. ArXiv:1802.04376
47 Kim J, Kim T, Kim S, et al. Edge-labeling graph neural network for few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 11–20
48 Chen W Y, Liu Y C, Kira Z, et al. A closer look at few-shot classification. 2019. ArXiv:1904.04232
49 Zhang C, Cai Y J, Lin G S, et al. DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 12203–12213
50 Ye H J, Hu H X, Zhan D C, et al. Learning embedding adaptation for few-shot learning. 2018. ArXiv:1812.03664
51 Yang L, Li L L, Zhang Z L, et al. DPGN: distribution propagation graph network for few-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 13390–13399
52 Schwartz E, Karlinsky L, Feris R, et al. Baby steps towards few-shot learning with multiple semantics. 2019. ArXiv:1906.01905
53 Zhang X L, Wei Y C, Feng J S, et al. Adversarial complementary learning for weakly supervised object localization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1325–1334