Robust Few-Shot Learning for User-Provided Data

Journal:	*IEEE Transactions on Neural Networks and Learning Systems*
Manuscript ID	TNNLS-2019-P-11825
Manuscript Type:	Regular Paper
Date Submitted by the Author:	22-Jul-2019
Complete List of Authors:	Lu, Jiang; Tsinghua University, Department of Automation Jin, Sheng; Tsinghua University, Department of Automation Liang, Jiang; Tencent, Cloud and Smart Industries Group Zhang, Changshui; Tsinghua University, Department of Automation
Keywords:	machine learning, few-shot learning, robustness, outlier suppression, authentication

Robust Few-Shot Learning for User-Provided Data

Jiang Lu, Sheng Jin, Jian Liang, and Changshui Zhang, *Fellow, IEEE*

***Abstract*—Few-shot learning (FSL) focuses on distilling transferrable knowledge from existing experience to cope with novel concepts for which only a few supervised data are available. Typical FSL settings consistently assume that the small-scale supervised set only contains clean data without any outlier interference. In many realistic applications, however, the supervised data are provided by users and unreadable, and they may be disturbed by noise. In this context, we introduce a novel research topic, robust few-shot learning (RFSL) with user-provided data, and handle two types of outliers, representation outlier (RO) and label outlier (LO). Moreover, we formulate a metric for robustness estimate and use it to extensively investigate the performance of currently advanced few-shot learning methods on RFSL problems. Furthermore, we propose robust attentive profile networks (RapNets) to achieve outlier suppression. Comprehensive evaluation results on benchmark datasets demonstrate the deficiency of current few-shot learning methods and the superiority of the proposed RapNets when dealing with RFSL problems, establishing a benchmark for follow-up studies.**

***Index Terms*—Machine learning, few-shot learning, robustness, outlier suppression, authentication.**

I. Introduction

DEEP learning has made impressive achievements in a broad spectrum of research fields including language [1], vision [2] and speech [3]. However, these successes heavily hinge on enormous training data and cumbersome manual annotation. Few-shot learning (FSL) remains challenging where only a few supervised data for novel concepts are available. It imitates the generalized learning ability of humans [4], [5], who can draw inferences about novel instances from only a few scraps of information (*e.g.* inferring various appearances of *zebra* given only one or a handful of *zebra* images). Concretely, in FSL we are provided with an auxiliary set $\mathcal{A} = \{(x_i, y_i)\}_{i=1}^{N_a}$ containing some classes with sufficient labeled data per class and a support set $\mathcal{S}'=\{(x'_i, y'_i)\}_{i=1}^{NK}$ containing N novel classes with only K labeled data per class (called support data). The goal is to correctly categorize one novel data x'_i (called query data) of a query set $\mathcal{Q}' = \{x'_i\}_{i=1}^{N_q}$ into one of the N novel classes. The class label space of $\mathcal{S}'$ and $\mathcal{Q}'$ are shared but does not overlap with that of $\mathcal{A}$. This typical FSL setting is known as N-way K-shot problem.

Despite the growing research interest in FSL and persistent improvement on benchmark performance [6]–[14], FSL is not robust to the noise. In many realistic applications, the support set $\mathcal{S}'$ are possibly unreadable and has suffered from outlier disturbance, especially when the support data in $\mathcal{S}'$ are provided by external users. A typical application scenario where the robustness needs to be considered is the few-shot authentication based on mobile devices using sensor information (*e.g.* accelerometer, gyroscope, gravimeter, *etc*)[1] [15], [16]. If regarding each user identity as an independent class, the authentication is naturally equivalent to a classification problem, where x (or x') denotes user data and y (or y') user identity label. In consideration of practicability and user experience, it is always expected to collect as few training data from new users as possible. In this context, FSL has valid realistic backgrounds for authentication: *auxiliary set $\mathcal{A}$ contains sufficient historical supervised data for old users attained by a long-term accumulation, support set $\mathcal{S}'$ involves the few data recently provided by N new users, and query set $\mathcal{Q}'$ includes online data to be authenticated amongst the N user identities*. An important concern about authentication is that the sensor information provided by user's mobile device (*i.e.* user data x') may be disturbed by noise injection during the phase of acquisition and transmission, or the data provided by user i may be annotated wrongly as user j ($j \neq i$).[2] Worse still, the sensor information are unreadable in most cases, which makes it impractical to find out these outliers manually.

Besides authentication, such outliers may widely exist in other real-world FSL problems. Suppose one zoological task is to automatically recognize several kinds of uncommon animals, and we have acquired only a few annotated images for them due to their rarity. However, the images have been potentially corrupted because of uncontrollable shooting environment and instrumental malfunction. Another application is the identification for several kinds of emerging products based on a few provided labeled product images. Nonetheless, some inevitable perfunctory errors or fraudulent behaviors in commercial markets may result in label-wrong support images. Unfortunately, the indistinguishability underlying the emerging concepts makes the potential outliers hard to be screened out by human empirical observation. Accordingly, this paper takes a step towards robust FSL and seeks to answer

This work was supported in part by the National Natural Science Foundation of China under Grant 61473167, in part by the Beijing Natural Science Foundation under Grant No.L172037, and in part by the German Research Foundation (DFG) through Project Cross Modal Learning under Grant NSFC 61621136008/DFG TRR-169. *(Corresponding author: Changshui Zhang.)*

J. Lu, S. Jin and C. Zhang are with the Institute for Artificial Intelligence, Tsinghua University (THUAI), Beijing 100084, China, and also with the State Key Laboratory of Intelligence Technologies and Systems, Beijing National Research Center for Information Science and Technologies (BNRist) , Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: lu-j13@mails.tsinghua.edu.cn; js17@mails.tsinghua.edu.cn; zcs@mail.tsinghua.edu.cn).

J. Liang is with the Wireless Security Products Department of the Cloud and Smart Industries Group, Tencent, Beijing 100080, China (e-mail: joshualiang@tencent.com).

[1] Digital information captured from multiple sensors on mobile devices are distinctive and used to recognize user identities based on the fact that different users have various use behaviors and operational habits.

[2] In mobile device authentication applications, the label-wrong data usually arises due to the automatic user identity annotation based on virtual account (*e.g.* WeChat, QQ). In other words, if user i uses the virtual account of user j to perform some operations on one mobile device, then the captured sensor information is naturally labeled into user j, even its true label is user i.

the following question: *if some outliers have been mingled into support set $\mathcal{S}'$, can current FSL models effectively alleviate the influence from such outliers and still maintain an acceptable few-shot classification capability on query set $\mathcal{Q}'$?* Intuitively, outliers severely deviate the distribution of normal data and thereby erect one obvious roadblock for FSL. Thus, we believe it is critical to devise a more robust framework to cope with the disturbance from outliers.

In this work, we introduce a novel and substantially meaningful research topic, *robust few-shot learning* (RFSL), in which outliers exist among the user-provided support set $\mathcal{S}'$. The goal is to classify the query data of $\mathcal{Q}'$ based on $\mathcal{S}'$. To better conduct this study, we focus on two types of outliers, *representation outlier* (RO) and *label outlier* (LO). RO involves noise interference or information breach on raw observed data, which may be caused by environmental conditions or instrumental errors. LO involves label-wrong data replacement, which may be caused by perfunctory errors or fraudulent behaviors. Moreover, we formulate mathematically a novel evaluation metric, *mean accuracy loss rate* (mALR), to estimate the robustness of FSL approaches. Furthermore, we propose the first practicable machine learning model tailored for RFSL, *robust attentive profile networks* (RapNets), which consist of *Embedding Module*, *Correlation Module* and *Attentive Module* and can be trained in an end-to-end manner. Our key insight is to unveil the heterogeneity and homogeneity within each group of congener support data by feature-level similarity assessment and then assign distinguishing emphases to different data by a sophisticated learnable attention mechanism. Afterwards the potential outliers can be suppressed implicitly and the feature-level attentive *profile* for each concept can be established. Different from other FSL models, we excavate the auxiliary set $\mathcal{A}$ more fully and devise the novel *biased meta episode* (BiME) to consolidate the discriminant ability of RapNets without introducing extra costs in data collection or annotation. Typical (*i.e.* no outliers) few-shot experimental results on benchmark datasets show that RapNets achieve a competitive performance compared with state-of-the-art FSL methods. More importantly, the comprehensive RFSL experiments under various few-shot settings and outlier disturbances demonstrate that RapNets are able to effectively suppress the potential outliers, yielding state-of-the-art robustness performance.

The rest of this paper is organized as follows. We review related works in Section II. The problem definition and related notation are presented in Section III. The proposed RapNets are detailed in Section IV. We report the experimental results in Section V and conclude our work in Section VI.

II. Related Work

A. Few-Shot Learning

Compared to many machine learning regimes involving large-scale supervised data, the development of FSL is relatively tardy due to its intrinsic difficulty. As a seminal work towards FSL, variational bayesian framework [17], [18] represents object classes by probabilistic models to achieve one-shot image classification. With deep learning booming, more efforts are focused on metric-learning [19]. Specifically, siamese neural networks [6] formulate one-shot recognition as a similarity-based image matching task. Matching networks [7] leverage an attention mechanism over the embeddings of all support data to construct a differentiable nearest neighbor classifier. Prototypical networks [10] propose to learn the prototype of each class as the cluster center of the embeddings of this class's support data. Relation networks [13] learn a relation comparator to match query data with support data. And in [20], mean average precision (mAP) is directly optimized to extract as much information as possible from a few samples. Another line of FSL is based on meta-learning [21]. Typically, MAML [9] and meta-learner LSTM [8] attempt to learn a set of good initial weights whereby the recognizer can reach well generalization performance within a few weight update steps on the sparse support data. Recurrent neural networks with memories [22]–[24] are exploited to rapidly assimilate never-seen-before information. Other approaches seek to solve FSL from a different perspective. Graph neural networks [25] deal with FSL via a graph-based model. Feed-forward parameter prediction based on factorization [26] is developed to construct the one-shot learner. The category-agnostic mapping from activation to parameters [27] is exploited to directly predict the parameters of novel class's recognizer. However, existing FSL methods congenitally do not consider interference of outliers, and seem to lack resistance to outliers.

B. Outlier Detection, Data Denoising and Label Noise

Outlier detection [28]–[31], also known as anomaly detection, is an attractive topic covering a broad spectrum of research fields. Here we briefly introduce several landmark outlier detection methods. Elliptic envelope (EE) [32] assumes the data obeys Gaussian distribution and uses minimum covariance determinant to measure the outlier-ness of data. One-class SVMs [33] regard the outlier detection as a binary classification problem between normal data and outliers. Local outlier factor (LOF) [34] evaluates how isolated the data is from its neighborhoods. Isolation forest [35] constructs a tree structure to isolate outliers explicitly. However, these outlier detection methods rely on large amounts of data, which are impracticable for RFSL problems.

Data denoising [36]–[38] is a hotspot in the research field of signal analysis, which is an indispensable step for many practical applications, such as image recognition and medical diagnosis. Some effective denoising methods, like sparse and redundant representations over learned dictionaries [39], non-local means [40], denoising auto-encoder [41], multi-column stacked sparse denoising auto-encoder [42] and denoising convolutional neural networks [43], have been developed. Nevertheless, such denoising approaches are not applicable to handle ROs in RFSL problems since they are noise-specific and which data within support set are ROs is unknown.

The proposed LO is a kind of mislabeled data in support set, and it essentially introduces the label noise [44] into RFSL tasks. Similar to outlier detection methods, most of label noise-robust methods including ensemble [45]–[47], boosting [48] and random forests [49] are not suitable for RFSL due to its scarce support data.

C. Neural Attention

Neural attention has been widely adopted in many machine learning problems including image classification [50], [51], image captioning [52]–[54], machine translation [55] and speech recognition [56]. For FSL, matching networks [7] also introduce neural attention mechanisms. Specifically, the attention of matching networks is generated by normalizing the cosine distances between the embeddings of query data with that of each support data. It is overlaid on the labels of all support data to infer the label of query data. In contrast, our attention is produced by the attentive module using correlation features within congener support data as input. It weighs the embeddings of these congener support data to form the attentive profile for the corresponding identity.

III. Definition and Notation

A. Robust Few-Shot Learning

A brief description for FSL and RFSL has been given in Section I. Formally, we define the RFSL problem as follows.

Problem 1. *(Robust Few-Shot Learning) Let x (or x') be the data and y (or y') the class label, $\mathcal{Y}_o$ and $\mathcal{Y}_n$ be the class label space of old and new concepts, respectively, and $\mathcal{Y}_o \cap \mathcal{Y}_n = \emptyset$. Given a clean auxiliary set $\mathcal{A} = \{(x_i, y_i)\}_{i=1}^{N_a}$ ($y_i \in \mathcal{Y}_o$) which contains sufficient supervised data for each old concept, and a support set $\mathcal{S}' = \{(x'_i, y'_i)\}_{i=1}^{NK}$ ($y'_i \in \mathcal{Y}_n$) which describes N new concepts, the goal is to train a model over $\mathcal{A} \cup \mathcal{S}'$, and then utilize this model to categorize the newcoming query data of $\mathcal{Q}' = \{x'_i\}_{i=1}^{N_q}$ amongst the N new concepts. In $\mathcal{S}'$, each new concept is described by K supervised data (K is usually very small) among which C outliers exist ($0 \leq C \ll K$). A hypothesis is presented that the user-provided $\mathcal{S}'$ is too cryptic or unreadable to be cleansed manually. For notation convenience, we also entitle the problem as the N-way K-shot C-outlier problem (NwKsCo).*

The assumption that auxiliary set $\mathcal{A}$ is clean is based on the logical justification: if $\mathcal{A}$ contains noise-polluted instances, since $\mathcal{A}$ is a off-line dataset containing sufficient supervised data attained by a long-term accumulation, there are enough time to sift the data in $\mathcal{A}$, or alternatively we can leverage the feasible data cleansing methods [44], [57] to filter $\mathcal{A}$.

B. Representation Outlier and Label Outlier

We distinguish two types of outliers for RFSL problems: representation outlier (RO) and label outlier (LO), which suffer observational and semantic disturbance, respectively.

Specifically, RO is the generic name of outliers whose corresponding observed data (or features) has suffered noise interference or information breach, and it figuratively simulates the cases where a few raw support data are corrupted due to harsh environmental conditions or instrumental malfunction. In this work, we technically use four common noises to generate ROs: gaussian noise (GN), speckle noise (SN), poisson noise (PN) and salt-and-pepper noise (SPN). In a real sense, GN always appears in low-lighting conditions, while SN occurs in diffuse reflections of light, and PN is the discrete electronic noise, while SPN is the bit errors in message transmission. The reason that we define the data polluted by observational noise as outliers is based on the empirical evidence [58]–[60] that the noise-polluted data can lead to an elusive representation embedding as well as error semantic prediction.

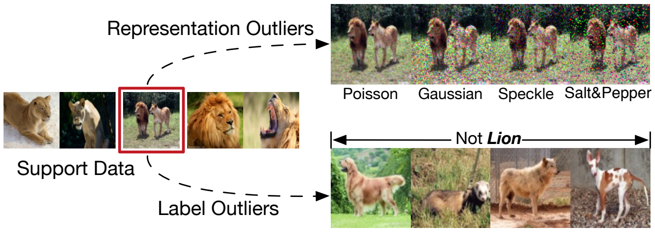


Fig. 1. Examples of ROs and LOs in one congener support set.

Comparably, LO represents label-wrong data whose real class label disaccords with its observed label that others in the congener support set belong to, and it simulates the cases in which one or several irrelevant data have blended in with the clean congener support data due to perfunctory errors or fraudulent behaviors. In other words, LO is an essentially label noise [44]. Specifically, we construct a LO through randomly replacing one label-true data by one data from other concept classes.

For a more intuitive comprehension, we present a visualized example in Fig. 1. Five images are considered to be from the same concept: *lion*, but actually one of them is corrupted.

C. Mean Accuracy Loss Rate

For an independent RFSL problem (NwKsCo), we use $acc(N, K, C)$ to measure the performance:

$$acc(N, K, C) = N_t / N_q * 100\%, \tag{1}$$

where $N_t \leq N_q$ denotes the number of correctly classified query data among $\mathcal{Q}' = \{x'_i\}_{i=1}^{N_q}$. To obtain a more robust estimate of performance and show the model capacity more objectively, we take the averaged performance $\mathrm{Acc}(N, K, C)$ over $n_{\mathcal{P}}$ independent NwKsCo problems as the metric to evaluate the classification accuracy of current model:

$$\mathrm{Acc}(N, K, C) = \frac{1}{n_{\mathcal{P}}} \sum_{j=1}^{n_{\mathcal{P}}} acc_j(N, K, C), \tag{2}$$

where subscript j associates the j-th NwKsCo problem. Especially, $\mathrm{Acc}(N, K, 0)$ degenerates into the usual accuracy metric for the typical FSL settings without any outliers.

Besides the accuracy, another capacity being concerned is the robustness. Since we are the first to propose this concept in terms of FSL, no off-the-shelf evaluation metric is available. Therefore, we design a novel metric for robustness estimate, called *mean Accuracy Loss Rate* (mALR):

$$\mathrm{mALR}_\alpha = \frac{1}{\lfloor \alpha K \rfloor} \sum_{C=1}^{\lfloor \alpha K \rfloor} \frac{\mathrm{Acc}(N, K, 0) - \mathrm{Acc}(N, K, C)}{\mathrm{Acc}(N, K, 0)}, \tag{3}$$

where $\mathrm{Acc}(N, K, C)$ is the accuracy performance given by Eq. (2), $\alpha \in (0, 1)$ is the predetermined maximum mixing ratio of outliers and $\lfloor \cdot \rfloor$ is the rounding down operator. The

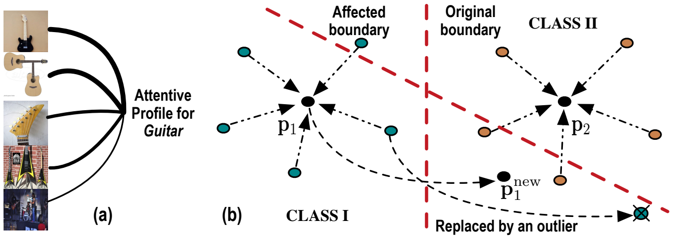


Fig. 2. (a) Our motivation of attentive profile for RFSL problem. The object saliency of *Guitar* is taken into consideration to assign reasonable attention weights to different support images. Note the thickness of curves indicates the weights given to corresponding *Guitar* pictures. (b) Example of prototypical networks when one outlier is existing. Once the outlier is mingled into support set of CLASS I to replace the normal data, the classification boundary of CLASS I and CLASS II is seriously affected.

mALR_α indicates the relative performance loss when a certain amount of outliers are mixed into the originally clean support set. Obviously, the smaller value of mALR_α demonstrates the better robustness. In our work, we consistently evaluate the robustness via $\mathrm{mALR}_{0.2}$ and $\mathrm{mALR}_{0.4}$ (*i.e.* $C = 1, 2$ when $K = 5$) since $\alpha \geq 0.5$ dose not conform to reality very well.

IV. Robust Attentive Profile Networks

A. Motivation

It is foreseeable that the optimization process of meta-learning based FSL methods, like MAML [9], will drift into a misunderstanding if the support data were mingled with outliers since these methods pin their hope on fine-tuning FSL learner on the support data. Metric-based method, prototypical networks [10], seems to be more robust to outliers because of its *average* mechanism over the congener support data, but it tends to fail in some classification dilemmas when outliers occur as Fig. 2(b). In particular, prototypical networks represent each class by an averaged embedding, known as *prototype*, over the embedding features of the congener support data belonging to that class, however, the average operation can hardly highlight the distinct importance of different support data. Thus, we believe one more reasonable strategy is to put distinguishing emphases on different support data in light of some key characteristics within data, like object saliency, semantic ambiguity and data *outlier-ness*, rather than absolute uniformity and then form an attentive *profile* representation for the corresponding concept (Fig. 2(a)). For a query data, we make inference by estimating the similarity between its representation embedding with each concept-specific profile.

B. Model Architecture

Episodic training strategy has been extensively employed by many FSL methods [7], [10], [13] to mimic the ultimate N-way K-shot evaluation settings. An episode is actually an independent FSL problem, which is constructed by randomly sampling N synthetic "novel" concepts with K support data and some query data per concept class from the auxiliary set $\mathcal{A}$ (the class label of query data is actually known). To distinguish from the true $\mathcal{S}'$ and $\mathcal{Q}'$, the synthetic support set and query set in training episodes are denoted as $\mathcal{S}=\{(x_i, y_i)\}_{i=1}^{NK}$ and $\mathcal{Q}=\{(x_i, y_i)\}_{i=1}^{N_q}$, respectively. Generally, FSL model is trained on multiple episodic problems constructed on $\mathcal{A}$ to acquire meta knowledge, and then transferred to the novel problem which is described by $\mathcal{S}'$ and $\mathcal{Q}'$. For the proposed RFSL, we also adopt the episodic training strategy to enhance the model's generalization capacity on novel RFSL problems. Differently, we devise an improved version, called biased meta episodes (BiME), to consolidate the discriminant ability of RapNets to outliers amongst a handful congener support data, which will be detailed in Section IV-C.

Our RapNets are inspired by prototypical networks [10], and it consists of three main modules, embedding module f_ϕ, correlation module g and attentive module h_φ, as depicted in Fig. 3. The embedding module f_ϕ is responsible for generating representation embeddings for support and query data, and the correlation module g computes the correlation features between embeddings of congener support data. Afterwards, these correlation features are leveraged as the input of the attentive module h_φ to generate a set of attention weights for the congener support data. Finally, the attention weights are overlaid upon the congener support data to form an attentive profile for the corresponding concept, whereby query data can be classified by the nearest-neighbor fashion in the embedding space.

1) Embedding Module: This module is designed in light of the data form. If the observed data are images, convolutional neural network (CNN) is chosen as the embedding module f_ϕ, and if these data are feature vectors, f_ϕ can be multilayer perceptron (MLP), where ϕ denotes the learnable parameter set of CNN or MLP. What the embedding module do is to map the observed data (or feature) x into a representation embedding $f_\phi(x)$.

2) Correlation Module: For the synthetic support subset $\mathcal{S}^n = \{(x_i^n, y_i^n)\}_{i=1}^{K} \subset \mathcal{S}$ in an episode problem, whose data are deemed to be from the n-th class, $n = 1, \cdots, N$, we can obtain their corresponding embeddings $F^n = \{f_\phi(x_i^n)\}_{i=1}^{K}$. In order to reveal the heterogeneity and homogeneity of each support data x_i^n within the K congener support data from the n-th class, we develop the nonparametric correlation module g to construct the correlation feature for each support data x_i^n:

$$g(x_i^n|\mathcal{S}^n) = \begin{bmatrix} \cos\left(f_\phi(x_i^n), f_\phi(x_1^n)\right) \\ \cdots \\ \cos\left(f_\phi(x_i^n), f_\phi(x_K^n)\right) \end{bmatrix}, \quad i = 1, \cdots, K. \tag{4}$$

In other words, the embedding-level cosine distance is used to portray the extent of relevance between two congener support data. The merits of using the correlation feature $g(x_i^n|\mathcal{S}^n)$ instead of the raw embedding $f_\phi(x_i^n)$ is that the correlation feature $g(x_i^n|\mathcal{S}^n)$ represents the higher-level semantic abstraction of the raw embedding $f_\phi(x_i^n)$ with reduced dimension, making the distinct importance or outliers underlying congener support data easier to be revealed and learned by the attentive module (see the results of ablation model Ab-1 in Section V-C).

3) Attentive Module: The aforementioned correlation feature set $G^n = \{g(x_i^n|\mathcal{S}^n)\}_{i=1}^{K}$ corresponding to the n-th class is fed into the attentive module h_φ to generate a set of attention scores $h_\varphi(G^n) = \{a(x_i^n|\mathcal{S}^n)\}_{i=1}^{K}$ for $\mathcal{S}^n$. Here, we design a parametric attentive module h_φ by a learnable bidirectional

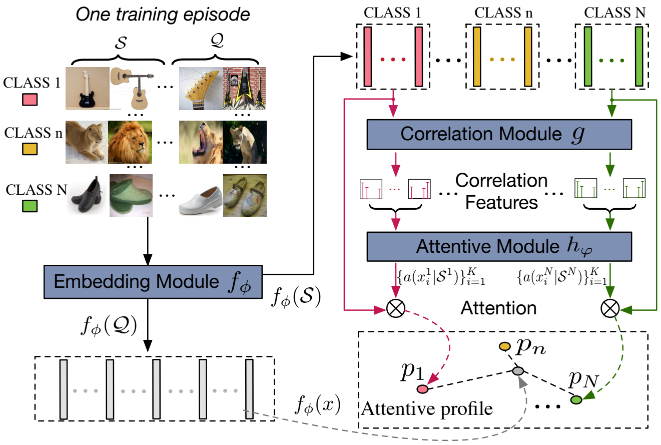


Fig. 3. Architecture of the proposed RapNets.

long-short term memory (BiLSTM) [61] and a fully-connected neural network (FCN), which aims at taking into account the interconnection between congener correlation features and the possible temporality between user-provided congener support data in many application cases. Concretely, each correlation feature $g(x_i^n|\mathcal{S}^n)$ contains the relevance information between x_i^n and other $K-1$ congener support data, which simultaneously scatters across other $K-1$ correlation features. The properties of interconnection and possible temporality make it more reasonable to take into consideration the full context of S^n to generate the ultimate attention, and BiLSTM can provide a good fit to the requirement. Thus, we regard each congener support set $\mathcal{S}^n$ as a sequence and use the BiLSTM to encode the correlation feature sequence $G^n = \{g(x_i^n|\mathcal{S}^n)\}_{i=1}^K$ so as to achieve the full context embedding:

$$\begin{aligned} \overrightarrow{z}_i, \overrightarrow{c}_i &= \mathrm{BiLSTM}(g(x_i^n|\mathcal{S}^n), \overrightarrow{z}_{i-1}, \overrightarrow{c}_{i-1}|\varphi_{lstm}) \\ \overleftarrow{z}_i, \overleftarrow{c}_i &= \mathrm{BiLSTM}(g(x_i^n|\mathcal{S}^n), \overleftarrow{z}_{i+1}, \overleftarrow{c}_{i+1}|\varphi_{lstm}), \end{aligned} \tag{5}$$

where z and c denote the hidden and cell state vectors inside BiLSTM, respectively, and φ_{lstm} is the learnable parameter set of BiLSTM. Then, the output of BiLSTM are concatenated together to be fed into the FCN followed by a K-way softmax layer which produces the probabilistic attention $\{a(x_i^n|\mathcal{S}^n)\}_{i=1}^K$. Its formalized forward process is:

$$\begin{aligned} &v^n = \mathrm{FCN}(\mathcal{C}(\mathcal{C}(\overrightarrow{z}_1, \overleftarrow{z}_1), \cdots, \mathcal{C}(\overrightarrow{z}_K, \overleftarrow{z}_K))|\varphi_{fcn}) \\ &\{a(x_i^n|\mathcal{S}^n)\}_{i=1}^K = \mathrm{Softmax}(v^n), \end{aligned} \tag{6}$$

where $n = 1, \cdots, N$ and $\mathcal{C}$ denotes the concatenation operation on vectors, and φ_{fcn} the parameter set of FCN. BiLSTM and FCN composes the whole attentive module h_φ.

Certainly, we can directly utilize a FCN to take $G^n = \{g(x_i^n|\mathcal{S}^n)\}_{i=1}^K$ as input and then generate the attention:

$$\begin{aligned} &v^n = \mathrm{FCN}(\mathcal{C}(g(x_1^n|\mathcal{S}^n), \ldots, g(x_K^n|\mathcal{S}^n))|\varphi'_{fcn}) \\ &\{a(x_i^n|\mathcal{S}^n)\}_{k=1}^K = \mathrm{Softmax}(v^n), \end{aligned} \tag{7}$$

where φ'_{fcn} denotes the parameter set of this FCN. However, this strategy omits the coupling between congener correlation features as well as the temporality of user-provided data, which leads to the attenuated performance (see the results of ablation model AB-2 in Section V-C). Alternatively, we can adopt a straightforward and nonparametric attentive mechanism by assessing the distance between each member to their congener cluster center:

$$\begin{aligned} c_n &= \frac{1}{K} \sum_{i=1}^K f_\phi(x_i^n), \\ a(x_i^n|\mathcal{S}^n) &= \frac{\exp(-d(f_\phi(x_i^n), c_n))}{\sum_{j=1}^K \exp(-d(f_\phi(x_j^n), c_n))}, i = 1, \ldots, K, \end{aligned} \tag{8}$$

where $d(\cdot,\cdot)$ denotes the squared Euclidean distance. This nonparametric-attention variant has also been discussed in Section V-C (see the results of ablation model AB-3).

The aforementioned generated attention $\{a(x_i^n|\mathcal{S}^n)\}_{i=1}^K$ is then overlaid upon the raw embeddings of support data in $\mathcal{S}^n$ to form an attentive profile for the n-th concept class:

$$p_n = \sum_{i=1}^K a(x_i^n|\mathcal{S}^n) \cdot f_\phi(x_i^n), \quad n = 1, \ldots, N. \tag{9}$$

For a query data x from the synthetic query set $\mathcal{Q}$, its predicted label probabilistic distribution over the N classes can be naturally formulated as follows:

$$P(y = n|x) = \frac{\exp(-d(f_\phi(x), p_n))}{\sum_{m=1}^N \exp(-d(f_\phi(x), p_m))}, \tag{10}$$

where $d(\cdot,\cdot)$ is the squared Euclidean distance. In inference phase, we categorize the query data x' of $\mathcal{Q}'$ by

$$\hat{y}' = \arg\min_n d(f_\phi(x'), p_n). \tag{11}$$

C. Biased Meta Episode

To explicitly enforce stronger discriminant power against outliers, we capitalize on auxiliary set $\mathcal{A}$ and propose BiME to train our RapNets. Firstly, we sample the clean episode (CE) problem from $\mathcal{A}$, $(\mathcal{S}, \mathcal{Q}) \sim \mathcal{A}$. Then, we build a BiME based on this CE. For a minimum cost, here we construct the BiME only using the label noise (*i.e.* LO). Concretely, a BiME is achieved by deliberately injecting some label-noise samples into the originally clean $\mathcal{S}$, where the label-noise samples come from non-target concept classes.[3] It is worth noting that this building process of BiME does not introduce any extra annotation cost compared with previous episodic training strategy. The underlying idea of BiME is to proactively build some biases into the training episode problems to facilitate the skepticism of RapNets to potential outliers.

D. Overall Training

We use the negative log-probability for each synthetic query data $(x, y) \in \mathcal{Q}$ as the classification loss function:

$$\mathcal{L}_c(\mathcal{S}, \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{(x,y)\in\mathcal{Q}} \big[-\log P(y|x) \big], \tag{12}$$

where $|\mathcal{Q}| = N_q$ is the number of query data, and $\log P(y|x)$ is given by Eq. (10). For BiME, its bias is controlled by us,

[3] In real-world applications, it is impractical to train with all potential noise types beforehand. We keep the models unaware of the existence of ROs during training and evaluate them with ROs during testing, to simulate the real scenes when the models meet unseen noise or outlier types.

and actually we know which ones in $\mathcal{S}$ are noise data during training within a BiME. Thus, we devise the following outlier suppression loss acted upon the generated attention scores of the corresponding support data:

$$\mathcal{L}_o(\mathcal{S}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{(x^n,y^n)\in\mathcal{S}^n}\Big[-\mathrm{I}_n(x^n)\log\big(1-a(x^n|\mathcal{S}^n)\big)\Big], \tag{13}$$

where the indicator function $\mathrm{I}_n(x^n)=1$ when x^n is a noise data and 0 otherwise, and $a(x^n|\mathcal{S}^n)$ is computed by Eq. (5) and (6). Certainly, we can also use clean episodes (CEs) which contain no outliers to train RapNets, or use BiME but the loss in Eq. (13) is abandoned. These two variants are analysed in Section V-C (see the results of ablation model Ab-4 and Ab-5).

Therefore, the total objective function for one BiME can be concluded as:

$$\mathcal{L}_{\mathrm{BiME}}(\mathcal{S},\mathcal{Q}) = \mathcal{L}_c(\mathcal{S},\mathcal{Q}) + \lambda\mathcal{L}_o(\mathcal{S}), \tag{14}$$

where λ is a trade-off between classification loss and outlier suppression loss. Like previous FSL methods [7], [10], [13], RapNets are trained on multiple independent BiMEs in an end-to-end manner by gradient back-propagation thanks to the differentiability of all modules:

$$\phi,\varphi = \arg\min_{\phi,\varphi} E_{(\mathcal{S},\mathcal{Q})\sim\mathcal{A}}\big[\mathcal{L}_{\mathrm{BiME}}(\mathcal{S},\mathcal{Q})\big]. \tag{15}$$

V. Experiments

In this section, we extensively investigate the performance of currently cutting-edge FSL methods as well as the proposed RapNets on both typical FSL and RFSL settings. Additionally, we provide the comprehensive ablation study of RapNets to demonstrate the contribution of each component. We also present the generated attention distribution to enhance the intuitive understanding to our methods. Our source codes will be released on https://github.com/LuJiangTHU/RapNets-for-Robust-Few-shot-Learning. All the experiments are implemented with Pytorch and carried out on an NVIDIA Tesla P40 GPU by Adam [62].

A. Typical Few-Shot Learning Problems

Because our proposed attention is acted upon at least two support data, *i.e.* $K \geq 2$, we compare RapNets with other methods only on the typical 5-shot FSL task settings.

1) Datasets: We perform typical FSL experiments on two benchmark datasets, Omniglot [4], [63] and miniImageNet [7], [8]. (i) Omniglot contains 1623 handwritten characters of 50 alphabets. Each character contains 20 character images drawn by 20 different writers. Following [7], [10] we resize the gray-scale images into 28×28 and augment each character with 90°, 180° and 270° rotations to form 4 different classes for each character. The 1200 characters with rotations (4800 classes) are used for training and the rest 423 characters with rotations (1692 classes) for evaluating. (ii) miniImageNet is a subset of ILSCRC-12 [64] and originally built in [7], consisting of 100 classes with 600 RGB images of size 84×84 per class. We follow the splits proposed by [8]: 64 classes for training, 16 classes for validation and 20 classes for test. Unlike the works [27] leveraging $64+16$ classes for training, in our experiments the 16 validation classes are only used to monitor the model's generalization performance.

TABLE I
Typical FSL on Omniglot. Results are reported by averaging accuracies over 1000 random tests. "Fine-tune" indicates whether or not the model is trained with support data.

Approaches	Fine-tune	5w5s0o	20w5s0o
Siamese Nets [6]	N	98.4%	96.5%
Siamese Nets [6]	Y	98.4%	97.0%
Matching Nets (FCE) [7]	N	98.9%	98.5%
Matching Nets (FCE) [7]	Y	98.7%	98.7%
Neural Statistician [65]	N	99.5%	98.1%
MAML [9]	Y	**99.9%**	98.9%
Prototypical Nets [10]	N	99.7%	98.9%
GNN [25]	N	99.7%	99.0%
Relation Nets [13]	N	99.8%	**99.1%**
RapNets (ours)	N	99.8%	**99.1%**

TABLE II
Typical FSL on miniImageNet using C64E as embedding architecture. Results are reported by averaging accuracies over 600 random tests with 95% confidence interval.

Approaches	Fine-tune	5w5s0o
Matching Nets (FCE) [7]	N	$55.31 \pm 0.73\%$
Meta-Leaner LSTM [8]	Y	$60.60 \pm 0.71\%$
MAML [9]	Y	$63.11 \pm 0.92\%$
Prototypical Nets [10]	N	$68.20 \pm 0.66\%$
Map-inspired Nets [20]	N	$63.94 \pm 0.72\%$
GNN [25]	N	$66.41 \pm 0.63\%$
Acts2Params(C64E) [27]	N	$67.87 \pm 0.20\%$
Relation Nets [13]	N	$65.32 \pm 0.70\%$
RapNets (ours)	N	$\mathbf{70.89 \pm 0.64\%}$

2) Setups: The embedding architectures f_ϕ we used keep the same with that of [7], [9], [10], [13], *i.e.* one 4 block based ConvNet model (called C64E here), and each of block comprised 64-channel 3×3 convolution, batch normalization, ReLU nonlinearity and 2×2 max-pooling. C64E leads to 64- and 1600-dimensional embedding vectors for data of Omniglot and miniImageNet, respectively. The attentive module h_φ consists of a 2-layer BiLSTM and a 2-layer FC followed by a K-way softmax layer. For Omniglot, we train RapNets with each BiME containing 60 classes and 5 support data per class for both 5-way and 20-way tasks. For miniImageNet, we train RapNets with each BiME containing 10 classes with 5 support data for each class. Each BiME randomly contains $C \leq 1$ label-noise sample per class in $\mathcal{S}$ and trade-off λ is set to 0.1.

3) Results: Performance comparisons for typical FSL on Omniglot are shown in Table. I. We achieve the state-of-the-art performance under the 20-way 5-shot setting and quite a competitive accuracy for 5-way 5-shot despite our RapNets have not been fine-tuned on support data. Table. II shows the typical FSL results on miniImageNet, in which we only list the performance of method [13] and [14] under the same embedding framework to make fair comparisons. The results indicate that our RapNets outperform the prototypical networks [10] and achieve the second best performance under

TABLE III

RFSL RESULTS ON OMNIGLOT. RESULTS ARE REPORTED USING $\mathrm{mALR}_{0.2}$ AND $\mathrm{mALR}_{0.4}$ WITH $\mathrm{Acc}(N, K, C)$ EVALUATED BY AVERAGING OVER 1000 RANDOM N-WAY K-SHOT C-OUTLIER PROBLEMS ON THE TEST SET. RESULTS IN BLUE COLOR ARE THE AVERAGED mALR_{α} OVER FIVE SETTINGS.

	Approaches	*ROs*						*LOs*					
		5w5s	**10w5s**	**20w5s**	**50w5s**	**100w5s**	***Avg.***	**5w5s**	**10w5s**	**20w5s**	**50w5s**	**100w5s**	***Avg.***
$\mathrm{mALR}_{0.2}$	MATCHING NETS [7]	3.72%	5.94%	5.92%	9.24%	11.21%	7.21%	2.45%	3.01%	5.02%	6.02%	8.91%	5.08%
	MATCHING NETS(FCE) [7]	3.29%	4.15%	5.13%	5.56%	7.45%	5.12%	2.20%	2.12%	2.90%	4.14%	4.67%	3.21%
	MAML [9]	0.83%	1.03%	2.38%	3.95%	6.81%	3.00%	1.23%	1.35%	3.25%	5.53%	9.98%	4.27%
	PROTOTYPIC NETS [10]	0.66%	0.80%	1.59%	3.56%	4.66%	2.25%	0.29%	0.43%	0.90%	1.49%	2.45%	1.11%
	RELATION NETS [13]	0.32%	0.44%	0.81%	2.24%	3.81%	1.52%	0.28%	0.34%	0.69%	1.35%	2.14%	0.96%
	RAPNETS (OURS)	**0.25%**	**0.35%**	**0.60%**	**0.63%**	**1.27%**	**0.62%**	**0.26%**	**0.23%**	**0.47%**	**0.63%**	**0.55%**	**0.43%**
$\mathrm{mALR}_{0.4}$	MATCHING NETS [7]	3.75%	6.40%	5.51%	12.65%	18.50%	9.36%	2.93%	3.39%	4.33%	8.68%	12.09%	6.28%
	MATCHING NETS(FCE) [7]	2.20%	4.59%	6.52%	9.55%	13.05%	7.18%	1.54%	2.64%	3.88%	7.24%	9.54%	4.97%
	MAML [9]	0.91%	1.68%	3.47%	7.49%	10.52%	4.81%	1.84%	3.08%	4.27%	10.92%	17.50%	7.52%
	PROTOTYPIC NETS [10]	2.56%	4.30%	6.53%	11.49%	13.85%	7.75%	0.73%	1.49%	2.52%	4.53%	6.84%	3.22%
	RELATION NETS [13]	0.47%	1.03%	2.16%	5.77%	8.56%	3.60%	0.61%	1.08%	2.17%	4.36%	7.16%	3.08%
	RAPNETS (OURS)	**0.45%**	**0.50%**	**0.81%**	**1.35%**	**1.66%**	**0.95%**	**0.09%**	**0.25%**	**0.48%**	**0.74%**	**1.18%**	**0.55%**

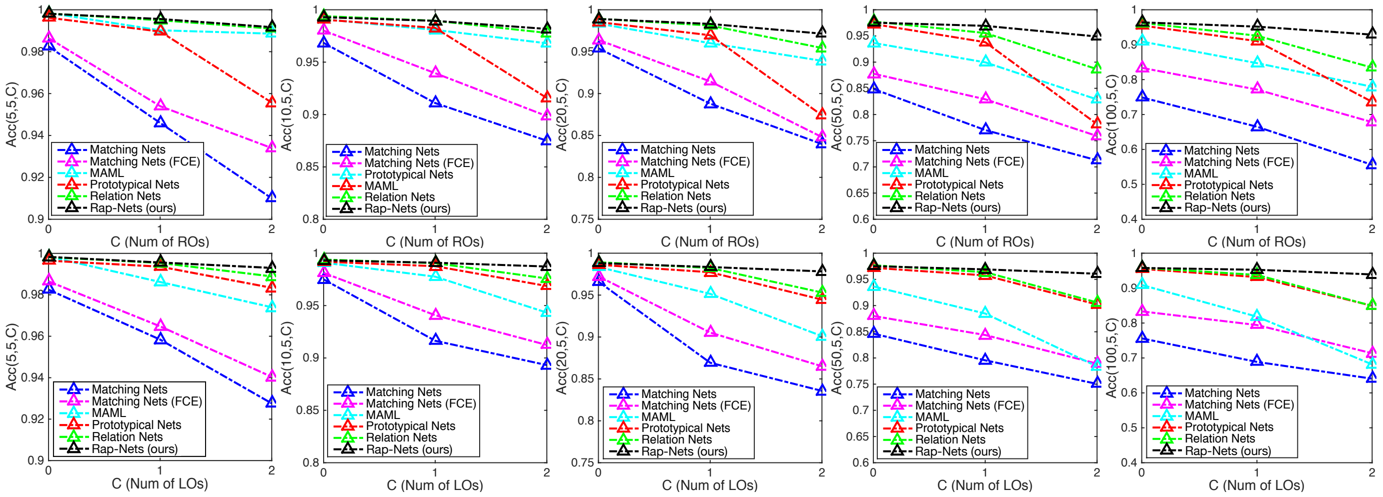


Fig. 4. Performance comparison of N-way K-shot C-outlier problems on Omniglot dataset. Each point denotes the $\mathrm{Acc}(N, K, C)$ value of one corresponding FSL method. The $\mathrm{Acc}(N, K, C)$ is computed over 1000 random N-way K-shot C-outlier problems on the test set. Each subgraph exhibits the change trends of $\mathrm{Acc}(N, K, C)$ w.r.t the number of LOs or ROs (*i.e.* C) under the corresponding fixed N and K setting.

typical 5-way 5-shot task setting. It is worthy noting that the experimental performance of other comparison methods that reported in Table. I and II are cited from their original papers.

B. Robust Few-Shot Learning Problems

In this section, we describe the RFSL experiments and compare our RapNets with several typical mainstream FSL methods including matching networks [7], MAML [9], prototypical networks [10], GNN [25], relation networks [13] and Acts2Params [27].

1) Datasets: We carry out RFSL experiments on Omniglot [4], [63] and miniImageNet [7], [8]. In particular, Omniglot and miniImageNet follow the same splits with those of Sec. V-A except that more N-way K-shot settings are exploited (detailed in the second row of Table. III and IV) to build a more comprehensive benchmark for RFSL tasks. As discussed in Sec. III-B, ROs are generated by randomly adding one type of noises (GN, SN, PN, and SPN) into raw support image, but LOs are produced via replacing raw image with another irrelevant image from other test classes except for the selected N classes in current test RFSL problem.

2) Setups: C64E is adopted as the embedding architecture for all methods for fairness. For miniImageNet, we modify the C64E of MAML into C32E (32 filters each Conv layer) to reduce overfitting, as done by [8], [9]. We reimplement these compared approaches in accordance with their corresponding original papers in order to reach the equivalent performance w.r.t typical FSL settings. For the added settings, like $10/50/100$-way 5-shot in Omniglot and 10-way 5-shot in miniImageNet, we adjust the training related hyperparameters of these methods to improve their performance on no outlier settings (*i.e.* $\mathrm{Acc}(N, K, 0)$) as far as possible. The attentive module of RapNets is realized using the same model architecture in Sec. V-A, and we train RapNets with BiMEs each of which randomly contains $C \leq \alpha K$ label-noise sample(s) per class in $\mathcal{S}$ ($\alpha = 0.2$ or 0.4). More implementation details are given in the Appendix.

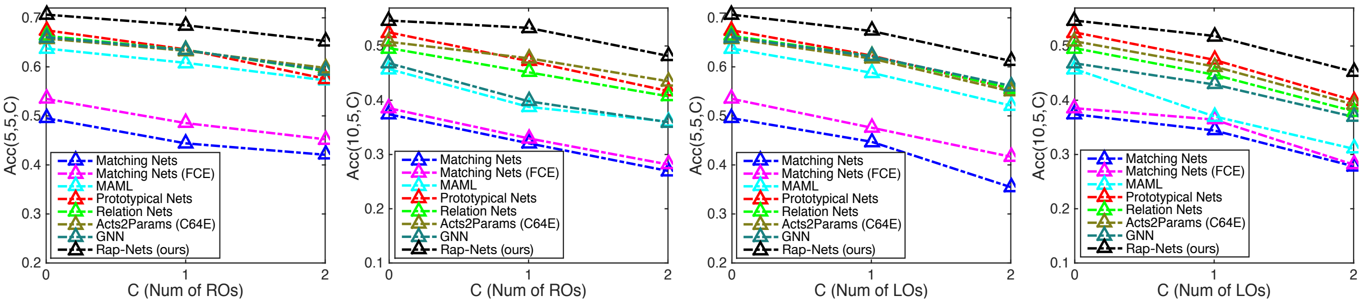


Fig. 5. Performance comparison of N-way K-shot C-outlier problems on miniImageNet dataset. Each $\mathrm{Acc}(N,K,C)$ is evaluated over 600 random N-way K-shot C-outlier problems on the test set. Each subgraph exhibits the change trends of $\mathrm{Acc}(N,K,C)$ w.r.t the number of LOs or ROs (*i.e.* C) under the corresponding fixed N and K setting.

3) Evaluation: Each $\mathrm{Acc}(N,K,C)$ is reported by averaging the classification accuracies over $n_{\mathcal{P}}$ random $N\mathrm{w}K\mathrm{s}C\mathrm{o}$ test problems, where $n_{\mathcal{P}}=1000$ for Omniglot, and $n_{\mathcal{P}}=600$ for miniImageNet. Each test problem contains 5 query data per class for Omniglot and 15 query data per class for miniImageNet. The effects from ROs or LOs are separately studied, and the ultimate robustness of methods are evaluated by $\mathrm{mALR}_{0.2}$ and $\mathrm{mALR}_{0.4}$ as described in Sec. III-C. In addition to mALR_{α}, we also compare the absolute classification accuracy $\mathrm{Acc}(N,K,C)$ between different methods, since the improvement of robustness naturally cannot be achieved at the cost of sacrificing classification accuracy.

4) Results on Omniglot: Five different RFSL settings on Omniglot including 5/10/20/50/100-way 5-shot with C=0, 1 or 2 are extensively investigated. Quantitative comparison results about mALR_{α} w.r.t ROs and LOs are shown in Table. III. The proposed RapNets achieve very low $\mathrm{mALR}_{0.2}$ and $\mathrm{mALR}_{0.4}$ on both ROs and LOs configures, significantly outperforming other strong baseline FSL methods. Additionally, with the number of task classes increasing from 5 to 100, the mALR_{α} of other FSL methods present the trend of malignant growth, illustrating the fact that they are easier to be seriously affected by outliers when dealing with FSL tasks of larger scales. Instead, the mALR_{α} of RapNets are able to retain a slight increase as the task scale growing and always stays low. More vivid qualitative and quantitative results are shown in Fig. 4, which describes the relationships of $\mathrm{Acc}(N,K,C)$ w.r.t the number of ROs or LOs under different FSL methods and various RFSL settings. We observe in Fig. 4 that the dashed line of RapNets (*black*) evidently lies above others and maintains a relatively low fall regardless of the outlier types and RFSL settings. Although the relation networks' performance (*green*) on no outlier settings (*i.e.* $\mathrm{Acc}(N,K,0)$ in Fig. 4) compares favorably with our RapNets, it inclines to a more rapid accuracy loss than ours when ROs or LOs mingled into support data. In brief, it can be reasonably concluded that our RapNets possess stronger robustness as well as higher classification accuracy than other methods on Omniglot.

TABLE IV
RFSL RESULTS ON MINIIMAGENET. RESULTS ARE REPORTED USING $\mathrm{mALR}_{0.2}$ (%) WITH $\mathrm{Acc}(N,K,C)$ EVALUATED BY AVERAGING OVER 600 RANDOM N-WAY K-SHOT C-OUTLIER TASKS ON THE TEST SET.

	Approaches	*ROs*		*LOs*	
		5w5s	**10w5s**	**5w5s**	**10w5s**
$\mathrm{mALR}_{0.2}$	Matching Nets [7]	10.36%	14.25%	9.69%	7.94%
	Matching Nets(FCE) [7]	9.24%	14.43%	10.98%	5.45%
	MAML [9]	4.58%	15.24%	7.75%	19.04%
	Prototypical Nets [10]	5.78%	10.09%	7.63%	9.64%
	Ralation Nets [13]	4.60%	8.97%	6.26%	9.86%
	Acts2Params(C64E) [27]	3.82%	5.95%	6.20%	9.12%
	GNN [25]	4.00%	14.95%	5.92%	8.33%
	RapNets (ours)	**3.07%**	**2.47%**	**4.73%**	**5.19%**
$\mathrm{mALR}_{0.4}$	Matching Nets [7]	12.71%	21.07%	19.04%	16.67%
	Matching Nets(FCE) [7]	12.32%	20.66%	16.52%	16.22%
	MAML [9]	7.38%	18.19%	12.99%	25.58%
	Prototypical Nets [10]	10.22%	15.33%	12.66%	16.87%
	Ralation Nets [13]	7.57%	13.42%	11.20%	16.65%
	Acts2Params(C64E) [27]	6.46%	10.28%	11.30%	15.99%
	GNN [25]	7.11%	19.07%	10.46%	14.71%
	RapNets (ours)	**5.34%**	**7.19%**	**9.04%**	**11.21%**

5) Results on miniImageNet: We present RFSL results on miniImageNet in Table. IV and Fig. 5, where all methods are compared under two few-shot settings w.r.t ROs and LOs, *i.e.* 5/10-way 5-shot. Our RapNets outperform these advanced FSL methods in robustness on miniImageNet, such as matching networks, MAML, prototypical networks, relation networks, Acts2Params and GNN. It should be emphasized that the reason why mALR_{α} values of all FSL methods on this dataset are generally higher than those on Omniglot is blamed on the complexity and ambiguity of the real-world images in miniImageNet. Compared to experiments on Omniglot, these difficulties underlying raw real-world data cause an essentially lower $\mathrm{Acc}(N,K,0)$ (the state-of-the-art $\mathrm{Acc}(5,5,0)$ on this dataset with C64E embedding architecture is only about 70% as shown in Table. II) and a larger decline as ROs or LOs added to support data. Similar to Fig. 4, we observe in Fig. 5 that our model also consistently overmatches other methods in the aspect of $\mathrm{Acc}(N,K,C)$, maintaining a relatively high level of accuracy even when one or two outliers are added.

C. Ablation Study

To demonstrate the effectiveness of each component of RapNets, we provide an ablation study on RFSL problems. We consider four variants, RapNets without correlation module (RapNets w/o COR), RapNets without full context embedding (RapNets w/o FCE), RapNets with nonparametric attentive

TABLE V
Ablation results of RapNets on RFSL tasks. Results are reported using $\mathrm{mALR}_{0.4}$.

	Ablations	*Omniglot*						*miniImageNet*	
		5w5s	**10w5s**	**20w5s**	**50w5s**	**100w5s**	***Avg.***	**5w5s**	**10w5s**
ROs	Ab-1: RapNets w/o COR	0.49%	0.56%	0.85%	1.45%	1.83%	1.04%	5.76%	7.99%
	Ab-2: RapNets w/o FCE	0.54%	0.98%	1.27%	2.78%	4.39%	1.99%	6.11%	9.28%
	Ab-3: RapNets with NAM	0.63%	1.39%	2.15%	4.98%	6.83%	3.20%	8.09%	11.75%
	Ab-4: RapNets with CE	0.51%	1.25%	1.82%	4.41%	6.51%	2.90%	6.25%	10.14%
	Ab-5: RapNets w/o OSL	0.49%	1.02%	1.49%	2.33%	3.23%	1.71%	6.09%	8.26%
	Standard RapNets	**0.45%**	**0.50%**	**0.81%**	**1.35%**	**1.66%**	**0.95%**	**5.34%**	**7.19%**
LOs	Ab-1: RapNets w/o COR	0.25%	0.34%	0.53%	0.82%	1.26%	0.64%	10.30%	13.08%
	Ab-2: RapNets w/o FCE	0.41%	0.98%	1.17%	2.46%	3.89%	1.78%	11.03%	14.08%
	Ab-3: RapNets with NAM	0.65%	1.18%	2.07%	3.88%	6.24%	2.80%	12.43%	15.58%
	Ab-4: RapNets w CE	0.56%	1.07%	1.64%	3.41%	5.64%	2.46%	11.36%	13.91%
	Ab-5: RapNets w/o OSL	0.32%	0.54%	0.87%	1.65%	2.44%	1.16%	10.88%	13.45%
	Standard RapNets	**0.09%**	**0.25%**	**0.48%**	**0.74%**	**1.18%**	**0.55%**	**9.04%**	**11.21%**

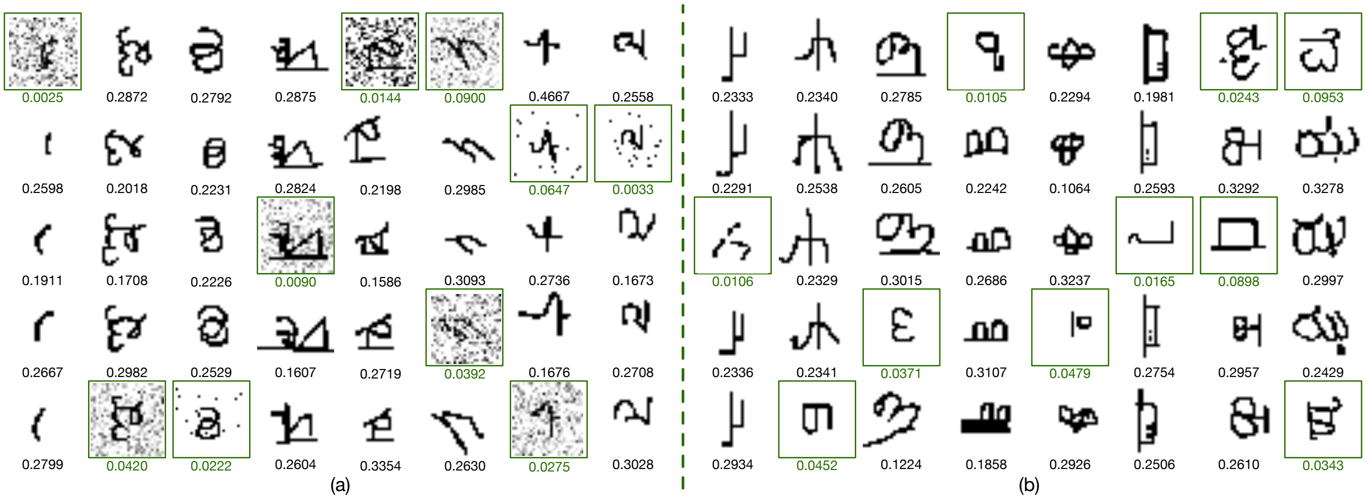


Fig. 6. Attention visualization on the test set of Omniglot. Each column represents one congener support set. (a) RO support sets, (b) LO support sets. Outliers are marked with green boxes. Numbers under images are attention scores generated by our RapNets.

module (RapNets with NAM), RapNets with clean episodes (RapNets with CE) and RapNets without outlier suppression loss (RapNets w/o OSL).

1) Ab-1 (RapNets w/o COR): This variant is built by removing the correlation module in standard RapNets. The embedding vectors of all congener support data are directly fed into the attentive module to generate the attention. The architecture of BiLSTM and FCN in attentive module was adjusted according to the dimension of embedding vectors.

2) Ab-2 (RapNets w/o FCE): This variant omits the full context embedding by BiLSTM, but directly utilizes a FCN to absorb the correlation features and then generate attention scores, as detailed in Eq. (7).

3) Ab-3 (RapNets with NAM): This variant uses the straightforward nonparametric attentive mechanism to generate the attention weights, as showed in Eq. (8). Other components or implementation details keep unchanged.

4) Ab-4 (RapNets with CE): This variant is trained through clean episodic problems which contain no outlier. Its BiLSTM and FCN keep the same as that of standard RapNets.

5) Ab-5 (RapNets w/o OSL): This variant is trained through BiMEs, but the outlier suppression loss in Eq. (13) is closed. In other words, there is no difference between RapNets w/o OSL with standard RapNets except that the trade-off hyperparameter λ of RapNets w/o OSL is set to 0.

With the results in Table. V, single-variable comparisons are available to show each component's contribution. We can conclude that the critical design elements in the RapNets, including the correlation module, the full context embedding achieved by BiLSTM, the parametric attentive mechanism, the BiME and the outlier suppression loss, all benefit the model's robustness.

D. Exploring the Generated Attention

For better comprehension of the RapNets, we perform the attention visualization for Omniglot and miniImageNet in Fig. 6 and 7, respectively. Notably, the images shown in Fig. 6 and 7 are chosen from the test sets, which means the RapNets have not seen these classes data before, and hence the attention results are able to reflect the actual learning efficacy of the attentive module. As shown in Fig. 6, (a) shows the attention

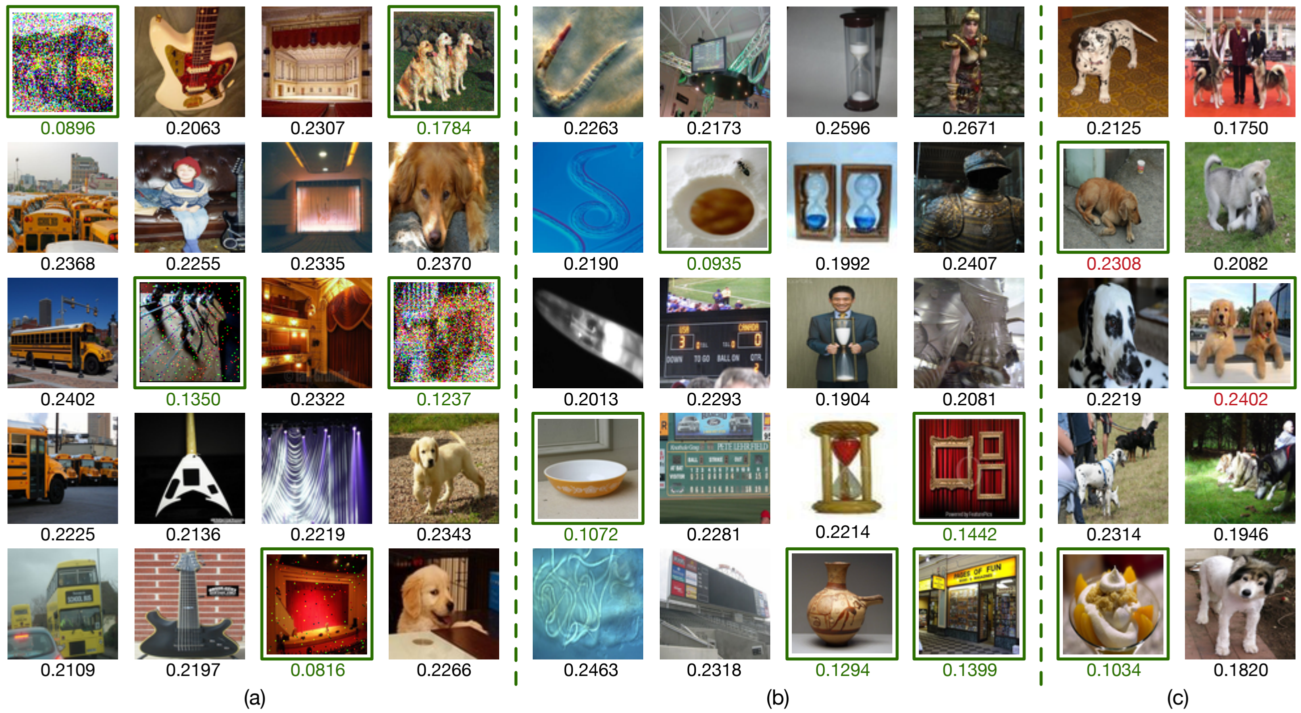


Fig. 7. Attention visualization on the test set of miniImageNet. Each column represents one congener support set. (a) RO support sets, (b) LO support sets, (c) Failure modalities. Outliers are marked with green boxes. Numbers under images are attention scores generated by our RapNets.

output of 8 random RO support sets (each RO contains one of GN, SN, PN, and SPN), and (b) illustrates the attention outputs of 8 random LO support sets. We observe that RapNets can accurately assign small attention weights for outliers no matter how slight the noise is, *e.g.* the third column in Fig. 6(a), or how similar the LO is with other clean support samples, *e.g.* the first LO character in the last column of Fig. 6(b). Although our RapNets have not seen any kinds of ROs during training, it successfully points out these ROs when testing. Similar observations can be drawn from Fig. 7. Compared to character images of Omniglot, the real-world images of miniImageNet have higher complexity and ambiguity, which reasonably leads to a less distinctive attentive probability distribution over congener support data. Even so, RapNets are still able to point out the outliers from only five support data in most cases. However, as depicted in Fig. 7(c), some failure modalities result in undesirable attention assignments, *e.g.* the serious inter-class similarity among various *dogs*. In summary, it shows that our RapNets have indeed gained the transferred attentive comprehension ability to suppress outliers.

VI. Conclusion

In this paper, we introduce a novel and meaningful reseach topic, robust few-shot learning (RFSL). This problem is extremely important and can not be neglected especially when dealing with user-provided data. We formulate the problem definition of RFSL and encapsulate two types of outliers, representation outlier (RO) and label outlier (LO). Additionally, we propose a mathematical metric, mean accuracy loss rate (mALR), to estimate the robustness to potential outliers. On these bases, we investigate the robustness of currently advanced few-shot learning methods, and propose the robust attentive profile network (RapNets) to learn to assign reasonable emphases on different supervised data by an attentive model and then achieve effective outlier suppression. Unlike previous efforts on unbiased episodic training strategy, we devise the biased meta episode (BiME) to proactively facilitate the skepticism of our RapNets against potential outliers. Extensive experimental results on two well-known benchmarks demonstrate that our RapNets not only reach the competitive performance on typical few-shot learning problems, but also achieve significantly stronger robustness to outliers than other mainstream few-shot learning methods. The ablation experiments validate that some momentous components of RapNets significantly contribute to the final robustness performance.

Appendix

Implementation Details

1) Architecture of Attentive Module: The attentive module h_φ consists of two parameterized networks, *i.e.* the 2-layer BiLSTM and 2-layer FCN, and two nonparametric operations, *i.e.* the concatenation and softmax. Let K be the number of data in one congener support set, the BiLSTM takes as inputs the sequence of K correlation features of one congener support data. The input and hidden size of the 2-layer BiLSTM are set to K and $(20, 20)$ respectively (because the dimension of correlation feature is K as shown in Eq. (4). The concatenation operation is acted upon the sequence of BiLSTM's outputs to produce a vector of size $40K$. The input and hidden size of 2-layer FCN are accordingly set to $40K$ and $(20, K)$. Specifically, the first layer of FCN is followed by a batch normalization layer and a ReLU nonlinearity, and the second

TABLE VI

CUTTING-EDGE FSL METHODS AND THE ADDRESS OF THEIR SOURCE CODES BY WHICH WE REIMPLEMENT THEM FOR RFSL PROBLEMS.

Approaches	**Address of souce code**
MATCHING NETS [7]	https://github.com/gitabcworld/MatchingNetworks
MAML [9]	https://github.com/katerakelly/pytorch-maml
RELATION NETS [13]	https://github.com/floodsung/LearningToCompare_FSL
GNN [25]	https://github.com/vgsatorras/few-shot-gnn
ACTS2PARAMS(C64E) [27]	https://github.com/joe-siyuan-qiao/FewShot-CVPR
PROTOTYPICAL NETS [10]	https://github.com/jakesnell/prototypical-networks

layer is equipped with a K-way softmax operation, which generates the K-dimensional attention scores.

2) Comparison Methods: Our experiments involve two settings: typical FSL and RFSL. In typical FSL setting, we compare our RapNets with other advanced FSL methods on benchmark datasets, and the experimental performance of other comparison methods are cited from their original papers. For RFSL problems, we put emphasis on several cutting-edge FSL methods and reimplement them according to their corresponding original papers and the publicly available codes, as shown in Table VI. We unify the dataset and data preprocessing, and ensure that the reimplemented methods can reach their reported performance w.r.t typical FSL settings. As for the unreported task settings, we adjust their related hyperparameters to make the $\mathrm{Acc}(N, K, 0)$ as high as possible.

3) Generation of LOs and ROs: In the evaluation phase of RFSL experiments, one LO is produced by randomly replacing one support data of S' with another irrelevant data from the remaining test classes. Comparably, one RO is generated by randomly adding one type of noises (GN, SN, PN and SPN) into one support data. Specifically, we utilize the image processing library *scikit-image*[4] in python to achieve noise interference. The noise of four types is randomly selected to be added to the clean data and several various severities of noises are considered. For GN and SN, we set five different severities about the variance of random distribution, *i.e.* 0.1, 0.2, 0.3, 0.4 and 0.5. For SPN, we also set five different severities about the proportion of image pixels to be replaced with salt or pepper noises, *i.e.* 1%, 2%, 3%, 4% and 5%. For PN, we use the default settings of *scikit-image* to generate noise. The noise severity in one RO is also randomly decided.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

[1] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Eleventh Annu. Conf Int. Speech Commu. Association (INTERSPEECH)*, 2010.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Sig. Process. Mag.*, vol. 29, 2012.

[4] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proc. Annu. Meet. Cogni. Sci. Society (CogSci)*, vol. 33, no. 33, 2011, pp. 2568–2573.

[5] B. M. Lake, R. R. Salakhutdinov, and J. Tenenbaum, "One-shot learning by inverting a compositional causal process," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2526–2534.

[6] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. (ICML) Deep Learning Workshop*, vol. 2, 2015.

[7] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3630–3638.

[8] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, 2017.

[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1126–1135.

[10] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4077–4087.

[11] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, 2018.

[12] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, 2018.

[13] F. S. Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1199–1208.

[14] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4367–4375.

[15] E. Shi, Y. Niu, M. Jakobsson, and R. Chow, "Implicit authentication through learning user behavior," in *Proc. Int. Conf. on Inf. Security*. Springer, 2010, pp. 99–113.

[16] J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu, "Additive adversarial learning for unbiased authentication," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 11 428–11 437.

[17] L. Fe-Fei *et al.*, "A bayesian approach to unsupervised one-shot learning of object categories," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2003, pp. 1134–1141.

[18] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Analy. Mach. Learn. (TPAMI)*, vol. 28, no. 4, pp. 594–611, 2006.

[19] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 1473–1480.

[20] E. Triantafillou, R. Zemel, and R. Urtasun, "Few-shot learning through an information retrieval lens," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 2255–2265.

[21] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial Intelligence Review*, vol. 18, no. 2, pp. 77–95, 2002.

[22] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1842–1850.

[23] T. Munkhdalai and H. Yu, "Meta networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 2554–2563.

[24] L. Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, 2017.

[25] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, 2018.

[26] L. Bertinetto, J. F. Henriques, J. Valmadre, P. H. S. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 523–531.

[27] S. Qiao, C. Liu, W. Shen, and A. L. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7229–7238.

[28] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[29] Q. Ke and T. Kanade, "Robust l/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 739–746.

[30] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[4] https://scikit-image.org/

[31] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Trans. Know. and Data Eng. (TKDE)*, vol. 26, no. 9, pp. 2250–2267, 2014.

[32] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[33] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 582–588.

[34] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[35] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *IEEE Int. Conf. on Data Min. (ICDM)*, 2008, pp. 413–422.

[36] J. Rissanen, "Mdl denoising," *IEEE Trans. on Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[37] M. C. Motwani, M. C. Gadiya, R. C. Motwani, and F. C. Harris, "Survey of image denoising techniques," in *Proc. of GSPX*, 2004, pp. 27–30.

[38] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[39] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process. (TIP)*, vol. 15, no. 12, pp. 3736–3745, 2006.

[40] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 60–65.

[41] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 341–349.

[42] F. Agostinelli, M. R. Anderson, and H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 1493–1501.

[43] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Trans. Image Process. (TIP)*, vol. 26, no. 7, pp. 3142–3155, 2017.

[44] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, 2013.

[45] G. Rätsch, T. Onoda, and K. R. Müller, "An improvement of adaboost to avoid overfitting," in *Proc. of the Int. Conf. on Neural Inf. Process.*, 1998.

[46] G. Rätsch, B. Schölkopf, A. J. Smola, S. Mika, T. Onoda, and K.-R. Müller, "Robust ensemble learning for data mining," in *Proc. Pacific-Asia Conf. Knowledge Disc. Data Mining*. Springer, 2000, pp. 341–344.

[47] Y. Freund, "An adaptive version of the boost by majority algorithm," *Mach. Learn.*, vol. 43, no. 3, pp. 293–318, 2001.

[48] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.

[49] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[50] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3156–3164.

[51] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4438–4446.

[52] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015.

[53] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4651–4659.

[54] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *IEEE Trans. Pattern Analy. Mach. Learn. (TPAMI)*, vol. 39, no. 12, pp. 2321–2334, 2017.

[55] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[56] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015.

[57] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proc. IEEE Int. Conf. on Comput. Vis. (ICCV)*, 2015, pp. 1511–1519.

[58] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, 2014.

[59] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[60] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2574–2582.

[61] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

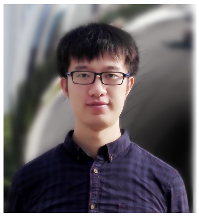
[63] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[64] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *Int. Jour. Comput. Vis. (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[65] H. Edwards and A. Storkey, "Towards a neural statistician," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, 2017.

Jiang Lu received the B.S. degree from Tsinghua University, Beijing, China, in 2013, where he is currently working toward the Ph.D. degree in the Department of Automation. His research interests include machine learning, deep learning and computer vision.

Sheng Jin received the B.S. degree from Tsinghua University, Beijing, China, in 2017, where he is currently working toward the M.S. degree in the Department of Automation. His research interests include machine learning and deep learning.



Jian Liang received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2012 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2018.

He is currently a senior researcher in the Wireless Security Products Department of the Cloud and Smart Industries Group at Tencent, Beijing. His research interests include authentication, transfer learning, semi-supervised learning. He won the Best Short Paper Award at the 2016 IEEE International Conference on Healthcare Informatics (ICHI).

Changshui Zhang (M'02–SM'15–F'18) received the B.S. degree in mathematics from Peking University, Beijing, China, in 1986, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, in 1992.

He is currently a Professor with the Department of Automation, Tsinghua University. His current research interests include artificial intelligence, image processing, pattern recognition, machine learning, and evolutionary computation.