24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

A Two-Stream Network with Image-to-Class Deep Metric for Few-Shot Classification

Qinghua Gu[1] and **Zhengding Luo**[1] and **Yuesheng Zhu**[*1]

Abstract. Few-shot learning in image classification aims to learn classifiers for new classes when few examples are available for each class. Though recent work has greatly advanced promising classification performance, they mainly focus on the feature maps extracted from RGB images and the task-invariant image-to-image metrics. In this paper, we argue that richer features need to be learned and the general metrics are not effective enough due to the scarcity of examples in few-shot learning. Specifically, we propose a Two-Stream Neural Network (TSNN) with a learnable Image-to-Class Deep Metric (ICDM) for few-shot learning, which is trained end-to-end from scratch upon the recent episodic training mechanism. We not only extract features from RGB images to find contrast differences in semantic information, but also leverage the steganalysis features extracted from a steganalysis rich model filter layer to discover the local inconsistencies between different categories. Meanwhile, we extend our model to fine-grained few-shot classification, which is benefit from the proposed novel ICDM. The experimental results on three benchmark datasets show that our approach attains superior performance, with the largest improvement of 6.01% in classification accuracy over related competitive baselines.

1 Introduction

Supervised deep learning [18] has shown great success in many applications such as computer vision [17, 25], speech processing [11, 12] and natural language processing [4]. However, these achievements have relied on a large amount of labeled data for the training process of the model, which is resource-intensive. In contrast, humans can learn a specific task based on a small number of known samples. For example, a child can learn the main features of the dogs according to several pictures rather than massive labeled data. Moreover, these supervised models are often limited in certain scenes and difficult to promote to general tasks. For image classification tasks, the classifiers trained with a certain dataset can only identify the classes in this dataset and cannot identify other new classes.

Few-shot learning [5] was proposed to learn a model with good generalization capability, such that it can adapt to recognize new classes based on few labeled support examples. As it is shown in Figure 1, few-shot learning in image classification aims to classify the query images by accessing the few images in support set, and these classes are not seen during training. There have been many contributions to the study of few-shot learning. The early work mainly focuses on the transfer learning methods [1, 21], where the pre-trained data is usually employed to fine-tune the few-shot learning models in order to ensure generalization performance.

[1] Communication and Information Security Lab, Shenzhen Graduate School, Peking University, China, email: {guqh, luozd, zhuys}@pku.edu.cn, *corresponding author

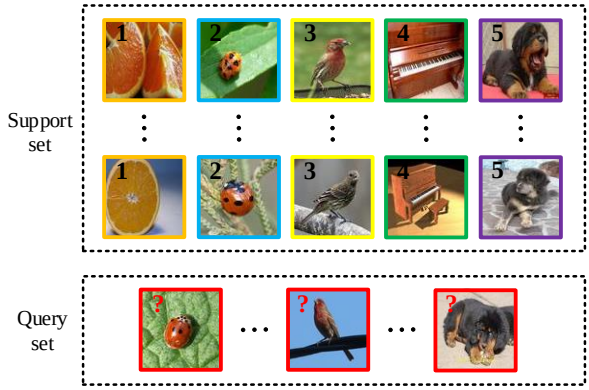


Figure 1. Illustration of few-shot image classification. The task aims to classify the query images in query set based on few samples in support set (e.g., there will be 10 support samples in the 5-way 2-shot task).

In recent years, a large number of studies on few-shot learning have been proposed, most of them can be divided into meta-learning based and metric-learning based. The meta-learning based approaches [24, 6, 26, 13] usually adopt recurrent neural networks or long short term memory networks [32] to store knowledge, and encourage models to adjust the appropriate parameters according to the previous experience. The metric-learning based approaches [16, 29, 27, 8, 28] are proposed to learn the distance distributions among samples relying on different metrics and episodic training mechanisms [29]. Most of the existing methods mainly focus on embedding RGB images, and the transfer of task-invariant distance metrics [16, 29, 27, 8] (e.g., Cosine Similarity and Euclidean Distance) or relation measure [28], but they don't pay attention that the richer features of images and a more efficient distance metric are critical for various few-shot tasks.

Therefore, we propose a novel Two-stream Neural Network (TSNN) for few-shot classification, which performs end-to-end training upon episodic training mechanism. Our approach aims to make full use of the few support samples and classify query images more directly and effectively. Specifically, in addition to detecting intrinsic contrast differences through embedding RGB images, we consider capturing inconsistencies in local steganalysis features. Then, feature maps of two streams are fused based on multimodal spatial fusion methods [23], which will find a joint representation that contains complementary information from different modalities.

The intuition behind the second stream is that the local steganalysis features between the different categories are unlikely to match. At the same time, these local steganalysis features can ignore some continuous background information, which may hinder the identifi-

cation of images. To utilize these features, we transform the RGB images into the steganalysis domain and use the local steganalysis features as the input to the second stream. Based on recent work on steganalysis rich model (SRM) [7] of digital images, we select SRM filter kernels to produce the steganalysis features.

Meanwhile, we consider that the general task-invariant measure of image-to-image distance is not effective enough for different few-shot classification tasks. For instance, Euclidean Distance shows poor performance in fine-grained image classification. In this case, we propose to learn the feature representations of support classes based on CNNs, which is conducive to reduce intra-class gaps and increase inter-class gaps. Thereafter, the adaptive Image-to-Class Deep Metric (ICDM) is utilized to measure the similarities between queries and each class flexibly in various few-shot classification tasks.

Our contribution is three-fold. First, a two-stream framework is employed in few-shot learning tasks for the first time, and we show that the two streams are complementary for few-shot classification. Second, a novel Image-to-Class Deep Metric (ICDM) is proposed to adapt to various few-shot classification tasks. Third, the experimental results on *mini*ImageNet show that our approach achieves the best results among related competitive approaches. The 5way-1shot and 5way-5shot accuracies on *mini*ImageNet have increased by 4.77% and 2.20% respectively. Besides, we extend our model to fine-grained few-shot classification, and conduct experiments on Stanford Dogs and CUB-200. The experimental results show that we gain the largest improvement over the second-best result by 6.01% on CUB-200 in 5-way 5-shot setting.

2 Related Work

Recent years have witnessed a vast amount of work on few-shot learning tasks. They can be roughly categorized into meta-learning based and metric-learning based.

Meta-learning based approaches: In these approaches, a meta-learner that learns to optimize model parameters extracts some transferable knowledge between tasks to leverage in the context of few-shot learning. Meta-LSTM [24] uses LSTM as a model updater and treats the model parameters as its hidden states. This allows to learn the initial values of parameters and update the parameters by reading few-shot examples. The MAML approach [6] optimizes the process of gradient descent through specific tasks (e.g., few-shot learning), then the parameters of a learner model are optimized so that they can be quickly adapted to a particular task. Another generic meta learner, SNAIL [19], is with a novel combination of temporal convolutions and soft attention to learn an optimal learning strategy. However, most meta-learning based approaches need complicated network architectures. Instead, we adopt a simple and effective CNN framework for few-shot learning, which can be trained end-to-end from scratch based on recent episodic training mechanism.

Metric-learning based approaches: These approaches [16, 29, 27, 8, 28] mainly focus on learning transferable embeddings and the distance distribution between images. Vinyals et al. [29] proposed the Matching Nets, a neural network that utilizes the Cosine Similarity to compute the distance of images and employs attention with memory that enables rapid learning. They also introduced the episodic training mechanism which is very effective for few-shot learning tasks. Snell et al. [27] proposed the Prototypical Networks to learn a metric space in which classification can be performed by computing Euclidean Distance to prototype representations of each class. The Relation Network [28] utilizes the Sigmoid function to convert the embeddings among RGB images into relation scores, and Mean Square Error (MSE) is selected as the loss function to train the network.

Our framework TSNN is a model based on measuring distance essentially. However, the most critical difference from existing approaches is that our TSNN is the first to employ the two-stream feature input in few-shot learning tasks. More specifically, in addition to the RGB image features used in previous models, we also extract steganalysis features based on the steganalysis rich model (SRM). In terms of distance metric selection, the previous approaches almost measure the distance between queries and support samples. Meanwhile, the fixed metrics [16, 29, 27, 8] or relation measure [28] are usually utilized in their work, which are not flexible enough to deal with various tasks. In contrast, we propose a novel learnable metric: ICDM, which can adaptively measure the similarities between queries and support classes for few-shot classification.

3 Proposed Approach

3.1 Task Formulation

When given a (small) support set $\mathcal{S}$ which consists of C different classes and K image samples each class, few-shot learning aims to classify the query images in the query set $\mathcal{Q}$ which consists of the images selected in the remaining images in the above C classes (i.e., $\mathcal{S}$ and $\mathcal{Q}$ share the same class label space). This setting is called the C-way K-shot in the few-shot learning.

The support set $\mathcal{S}$ and query set $\mathcal{Q}$ can be described as:

$$\mathcal{S} = \{(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_n, y_n)\} \quad (1)$$

$$\mathcal{Q} = \{(x_{n+1}, y_{n+1}), \cdots, (x_j, y_j), \cdots, (x_{n+m}, y_{n+m})\} \quad (2)$$

Where n is the number of the support images ($n = C \cdot K$), m is the number of the query images, x_i, x_j are support and query instances respectively, and y_i, y_j are their corresponding class labels.

For few-shot learning tasks, the training set $\mathcal{D}_{train}$, evaluation set $\mathcal{D}_{val}$ and test set $\mathcal{D}_{test}$ have a disjoint class label space with each other for classifiers. The $\mathcal{D}_{val}$ is used to evaluate the generalization performance of the model during the training process. In this case, the classifier trained by the training set $\mathcal{D}_{train}$ needs to be able to identify the new category on the test set $\mathcal{D}_{test}$ correctly. We need to consider how to train such a classifier with good generalization performance. In addition to the ideas of the model framework, training strategy is also crucial for few-shot learning.

The training strategy in Matching Networks [29] called episodic training mechanism has achieved good performance in few-shot learning, and it will be adopted in our work. More specifically, the training procedure is the same with few-shot classification procedure. In each training episode, we perform the same C-way K-shot setting as we do during the testing. After hundreds of thousands of training episodes, the model can be directly used for few-shot classification tasks with the support set $\mathcal{S}$ and $\mathcal{Q}$ which are selected from images with new classes. More details will be shown in the network training section and the experimental part.

3.2 Model Description

Figure 2 shows our Two-stream Neural Network (TSNN) architecture of the 5-way K-shot setting, which is based on a metric called Image-to-Class Deep Metric (ICDM). In the 5-way K-shot setting, the support sets of two streams consist of 5 different classes with K images per class. The RGB stream aims to detect visual differences and contrast effect between query images and support samples. Meanwhile, the steganalysis features can provide additional

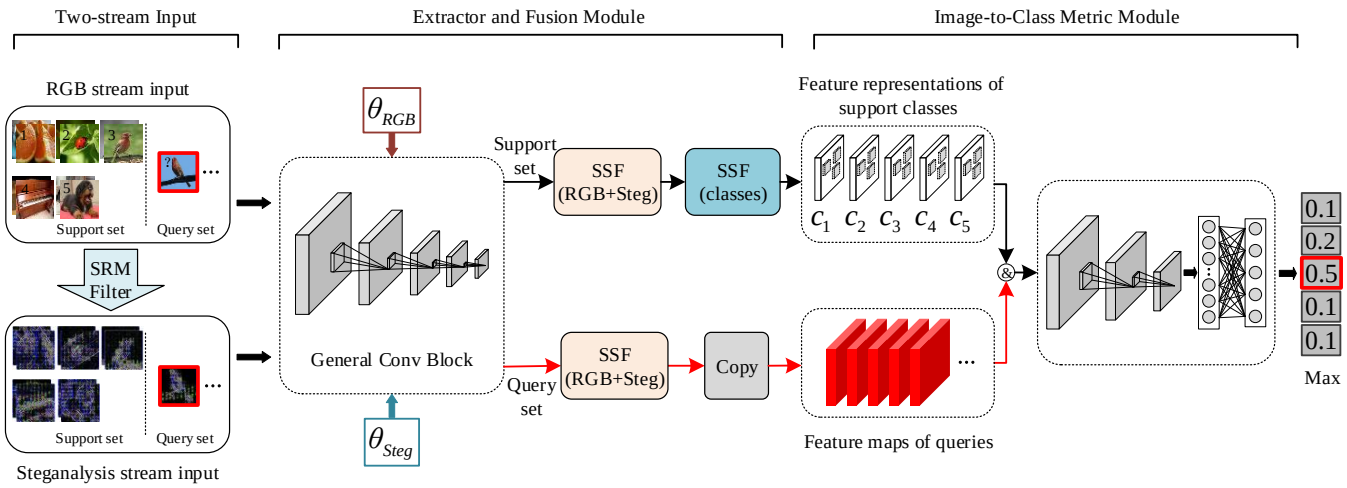


Figure 2. Architecture of our TSNN for a few-shot learning task in 5-way K-shot setting. Both two streams contain support set and query set (a query image is marked with a red box). The General Convolutional Block is used to embed two streams through corresponding parameters θ_{RGB} and θ_{Steg}. SSF: Sum Spatial Fusion. The ICDM learns to discover the distance between queries and class c_i, and we convert it into the probability distribution. &: Concatenate.

critical information (e.g., the local inconsistencies between different classes) to assist few-shot classification. Then, we fuse the two streams through spatial fuse methods, which will find a joint representation that contains complementary information from different modalities. The support samples belonging to the same class are uniformly fused as the feature representations corresponding to c_i $(i = 1, 2, \cdots, 5)$. Finally, the probability distributions of the query images can be calculated through the image-to-class metric module.

3.2.1 *Two-stream Input*

The RGB stream input directly consists of the original RGB images, which can find contrast visual differences in semantic information. However, the single RGB stream is not sufficient to discover all the key information of images. On the one hand, we consider designing an efficient and generic network structure instead of a complex deep network. On the other hand, it is challenging for single RGB images to extract general feature maps to improve generalization performance due to lacking support samples.

So, we utilize the local steganalysis distributions of the images to provide additional evidence. In contrast to the RGB stream, the steganalysis stream is designed to pay more attention to local intrinsic features rather than semantic image content. This is novel for few-shot learning while current approaches focus on extracting features from RGB image content, no prior work in few-shot learning has investigated learning from steganalysis distributions. Inspired by rich models for steganalysis of digital images, we use SRM filters [7] to extract the local steganalysis features from RGB images as the input to our steganalysis stream. Examples of two stream inputs are shown in Figure 3. We can see that local steganalysis can filter out many smooth backgrounds, and the steganalysis features of main objects are mainly retained, which is very helpful for few-shot classification.

In our setting, steganalysis is modeled by the residual between a pixel's value and the estimate of that pixel's value produced by interpolating only the values of neighboring pixels. Starting from 30 basic filters, SRM quantifies and truncates the output of these filters and extracts the nearby co-occurrence information as the final features. We find that only using 3 kernels can achieve decent performance in our work, and applying all 30 kernels does not give significant performance gain. Therefore, we choose 3 kernels, whose weights are shown in Figure 4. We define the kernel size of the SRM filter layer in the steganalysis stream to be $5 \times 5 \times 3$.

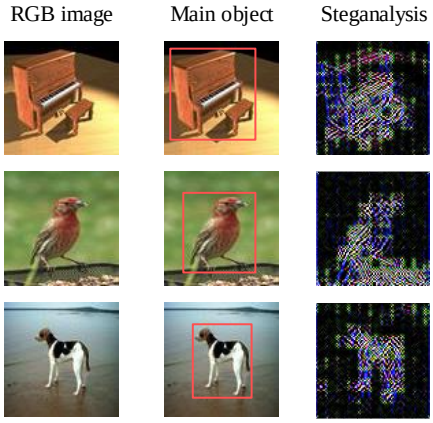


Figure 3. Examples of two-stream inputs. The first column shows the original RGB images. Meanwhile, we mark the main objects to be identified in the second column with red bounding boxes. The third column shows the local steganalysis stream input obtained by the SRM filter layer.

$$\frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix}$$

Figure 4. Three SRM filter kernels used to extract steganalysis features.

3.2.2 *Extractor and Fusion Module*

In the feature extraction module, we design a General Convolutional Block to extract both feature maps of RGB stream input and steganalysis stream input through corresponding network parameters θ_{RGB}

and θ_{Steg}. More specifically, the General Convolutional Block consists of four convolutional blocks, each of them contains a convolutional layer which including 64 filters of size 3×3, a batch normalization layer and a ReLU nonlinear activation layer. The first two blocks also contain a 2×2 max-pooling layer to reduce feature dimensions and prevent overfitting.

As for spatial fusion module, assume for a particular data point, there are M group feature maps corresponding to the representation of different modalities. A fusion function $f : \{z^1, z^2, \cdots, z^m, \cdots, z^M\} \rightarrow Z$ fuses the M group feature maps and produces an output Z, where z^m denotes the feature maps of the mth modality. For simplicity we assume that all the input feature maps have the same dimension of $\mathbb{R}^{H\times W\times d^{in}}$, and the output has the dimension of $\mathbb{R}^{H^{'}\times W^{'}\times d^{out}}$. In our work, given an image with a resolution of 84×84, we can get $H = W = 19$ and $d^{in} = 64$ for both RGB and steganalysis streams because they have the same convolutional network architecture. At the same time, since we are using spatial fusion methods, the size of the feature maps do not change (i.e., $H^{'} = H = W = W^{'}, d^{in} = d^{out}$). Here, Sum Spatial Fusion (SSF) is adopted in our experiments, which can be expressed as:

$$Z_{i,j,k} = \sum_{m=1}^{M} z^m_{i,j,k} \tag{3}$$

where $Z_{i,j,k}$ denotes the value in the spatial position (i, j, k) in the output, and z^m denotes the input feature maps of mth modality. In our work, $M = 2$ (i.e., RGB stream and steganalysis stream).

Specifically, we have conducted SSF twice for the support images. The first is the fusion of the two-stream feature maps of support images, which is the same as the queries. For the support images belonging to the same class, we conduct the second SSF orderly to obtain the feature representations of support classes.What's more, average-pooling is utilized to reduce the variability between different modalities after sum spatial fusion. Since we add some zero paddings to the periphery of the feature maps, the output $\bar{Z} = \text{Avg-Pooling}(Z)$ have the same dimension with Z (i.e., $\bar{Z}, Z \in \mathbb{R}^{19\times19\times64}$). Of course, other fusion methods are also possible.

3.2.3 *Image-to-Class Metric Module*

In most previous approaches, image-to-image distance metrics are usually used to measure the similarities between images. However, in the few-shot classification, these metrics cannot effectively classify the queries due to lacking support samples. In our work, we adopt a novel learnable Image-to-Class Deep Metric (ICDM). The comparison of the two kinds of metrics is shown in Figure 5. In the image-to-class metric, the feature maps belonging to the same class are merged into the joint feature representations, which contains high-level semantic information of the corresponding class. In this way, the query image can measure its similarities to each class conveniently and flexibly in few-shot classification. Meanwhile, the adaptive deep metric can be applied to a variety of classification tasks, while task-invariant metrics show poor performance in certain tasks. More details will be shown in the experimental section.

Specifically, we concatenate the class representations of support images and feature maps of queries in depth. Then, two convolutional blocks and two fully-connected layers are used to learn the ICDM. The activation function of last output fully-connected layer is Softmax function in order to compute the probability values that the query x_j belongs to each class c_i, and $i = 1, 2, \cdots, C$. For a new

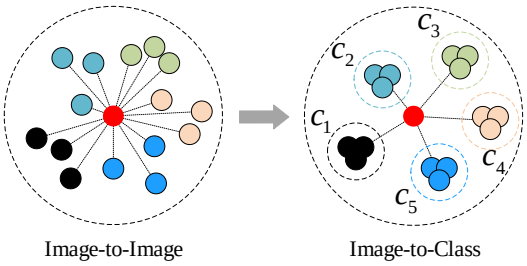


Figure 5. The comparison of two kinds of metrics in the 5-way 3-shot setting. Red indicates the feature maps of the queries, and the remaining colors indicate the feature maps of support samples of different classes.

query x_j, its predicting probability distribution $\hat{y}_j$ over all C support classes can be formulated by a Softmax function as:

$$p(\hat{y}_j = c_i | X_{c_i}, x_j) = \frac{\exp(\mathcal{M}(\mathcal{F}(X_{c_i}), \mathcal{F}(x_j)))}{\sum_{c_i{'}=1}^{C} \exp(\mathcal{M}(\mathcal{F}(X_{c_i{'}}), \mathcal{F}(x_j)))} \tag{4}$$

where X_{c_i} denotes the support samples of class c_i, $\mathcal{F}(\cdot)$ denotes the extractor and fusion process of two-stream features, $\mathcal{M}(\cdot, \cdot)$ represents the learned Image-to-Class Deep Metric in our work.

3.3 Network Training

Following episodic training mechanism [29], in each training episode, we randomly sample C classes from the training set $\mathcal{D}_{train}$. Then, we randomly select K labeled samples each class to compose a training support set $\mathcal{S}$ (e.g., in a 5-way 5-shot task, the $\mathcal{S}$ will have 25 samples). At the same time, we randomly select some samples from the remaining parts of above C classes to compose a training query set $\mathcal{Q}$ (i.e., $\mathcal{S}$ and $\mathcal{Q}$ have the same class label space).

Next, the TSNN is trained to minimize the error predicting the class labels $\hat{y}_j$ in the query set $\mathcal{Q}$ conditioned on the support set $\mathcal{S}$. More precisely, the TSNN training objective is denoted as:

$$\theta = \underset{\theta}{\text{argmax}}\, E_{\mathcal{S},\mathcal{Q}\in\mathcal{D}} \left[\sum_{(x_j,y_j)\in\mathcal{Q}} \log p_\theta(\hat{y}_j = y_j | x_j, \mathcal{S}) \right] \tag{5}$$

Training θ with Eq. 5 yields a model which works well when sampling $\mathcal{S}$ and $\mathcal{Q}$ from $\mathcal{D}_{test}$. Crucially, our approach does not need any fine-tuning when recognizing new categories, because it learns rich features from RGB and steganalysis images, and the intrinsic correlations between images and categories are learned based on the ICDM.

4 Experiments

In this section, we execute experiments with C-way K-shot settings on three datasets to evaluate the performance of our proposed TSNN. In addition to the commonly used datasets *mini*ImageNet, we also adjust two fine-grained image classification datasets Stanford Dogs [14] and CUB-200 [31] to make them suitable for few-shot classification tasks.

The main goal of this section is to investigate three questions: (1) Will our TSNN, a model based on a novel deep metric and the episodic training mechanism, work well in the actual experiments? (2) Can our TSNN adapt to various tasks on different datasets? (3) What roles do the two streams play respectively in few-shot classification?

4.1 Datasets

Our experiments are executed on three datasets: *mini*ImageNet, Stanford Dogs and CUB-200. For two datasets used for fine-grained image classification, we have pre-processed them for few-shot learning.

***mini*ImageNet.** The *mini*ImageNet dataset, originally proposed by Vinyals et al. [29], consists of 60000 colour images of size 84×84 with 100 different classes and 600 samples each class. For our experiments, we use the splits introduced by Ravi and Larochelle [24] in order to directly compare with state-of-the-art algorithms for few-shot learning. The 100 classes are divided into 64, 16 and 20 classes for training, validation, and testing respectively.

Stanford Dogs. This dataset [14] is proposed for fine-grained image classification which contains 120 classes of dogs from around the world and 20580 images in total. In order to make it suitable for our experiments, we split the dataset with 75, 20, 25 classes for training, validation, and testing. It is worth mentioning that some scenes in this dataset are complex or have multiple objects, which is also a challenge for our model.

CUB-200. This dataset [31] is another dataset for fine-grained image classification which contains 6033 images from 200 classes of birds. We split the dataset with 135, 25, 40 classes for training, validation, and testing. For two fine-grained image classification datasets, all the images are resized to 84×84 as *mini*ImageNet. What's more, we augment two fine-grained datasets with random rotations by multiples of 90 degrees in order to prevent overfitting, because there are fewer images in these datasets than *mini*ImageNet.

4.2 Parameter Settings

For all the experiments on three datasets, we execute tasks of 5-way 1-shot with 15 query samples each class, and 5-way 5-shot with 10 query samples each class. More specifically, for the task of 5-way 5-shot setting, in one training episode, there will be $5 \times 5 = 25$ support samples and $5 \times 10 = 50$ query samples to form the support set $\mathcal{S}$ and query set $\mathcal{Q}$ respectively. The total 400000 training episodes are constructed for sufficient generalization performance. We adopt the Adam algorithm [15] with an initial learning rate 10^{-3} and halve the learning rate each 100,000 episodes.

During test, we cycle a total of 10 times of test procedures. For each test procedure, we perform 600 test episodes for all the datasets. For each test episode, we choose 15 query images per class to test the accuracy of the classification in all the settings. More specifically, for a test task on *mini*ImageNet of 5-way 5-shot setting, we randomly sample 25 support samples from $\mathcal{D}_{test}$ to form the support set $\mathcal{S}$, and test the accuracies of the classification of $5 \times 15 = 75$ query images in one test episode. The classification accuracies of all test procedure are the average of the accuracy of 600 test episodes. Further, we also record the 95% confidence interval of accuracy.

4.3 *mini*ImageNet Few-shot Classification

Baselines. For the experiments on *mini*ImageNet, we choose five various state-of-the-art baselines whose type is metric-learning based few-shot learning approaches, because our TSNN is a metric-learning based methods essentially, including Matching Nets [29], Prototypical Nets [27], Graph Neural Network (GNN) [8], Relation Net [28] and Maximum-entropy Reinforcement Learning [3]. It should be noted that we re-evaluate the performance of Prototypical Nets and GNN for a fair comparison, because original Prototypical Nets needs more support samples when training models, and original GNN uses more filters than ours in the same settings. The Max-entropy RL uses the same Cosine Similarity as the Matching Nets, but it adopts a Maximum Entropy Sampler for few-shot learning.

Results on *mini*ImageNet. Results comparing the five metric-learning based baselines to our model on *mini*ImageNet are shown in Table 1. The second column in the table refers to the type of this model. The third column refers to different metrics in these approaches. Our TSNN achieves the best results among all related metric-learning based approaches. More specifically, we gain the largest improvement over the second-best result by 4.77% and 2.20% in 5-way 1-shot and 5-shot settings. Here we answer the first question, our approach can perform well in few-shot classification.

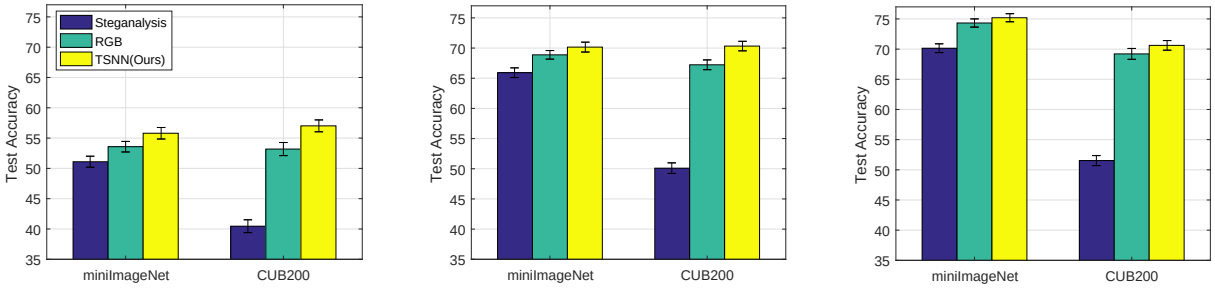
In order to show all types of few-shot learning approaches, we also list some competitive meta-learning based approaches: Meta-Learner LSTM [24], MAML [6], MM-Net [2] and Dynamic-Net [10]. Compared with these models, our TSNN also achieves certain improvements over them. As for the Dynamic-Net, it pre-trains its model with all classes together before performing the few-shot training. It should be emphasized that our TSNN does not require storing past information, which is much simpler than the meta-learning based approaches.

Table 1. The mean accuracies on *mini*ImageNet and with 95% confidence intervals. Our baselines are five metric-based learning approaches. $*$: Results re-implemented in the same setting for a fair comparison. The best-performing approach is highlighted.

Model	**Type**	**Metric**	**5-way Accuracy(%)**	
			1-shot	5-shot
Meta-Learner LSTM [24]	Meta	-	43.44 ± 0.77	60.60 ± 0.71
MAML [6]	Meta	-	48.70 ± 1.84	63.11 ± 0.92
MM-Net [2]	Meta	-	53.37 ± 0.48	66.97 ± 0.35
Dynamic-Net [10]	Meta	-	55.45 ± 0.89	70.13 ± 0.68
Matching Nets [29]	Metric	Cosine Similarity	43.56 ± 0.84	55.31 ± 0.73
Prototypical Nets$*$ [27]	Metric	Euclidean Distance	47.45 ± 0.93	66.24 ± 0.52
Prototypical Nets [27]	Metric	Euclidean Distance	49.42 ± 0.78	68.20 ± 0.66
GNN$*$ [8]	Metric	Absolute Difference	49.34 ± 0.85	63.25 ± 0.81
GNN [8]	Metric	Absolute Difference	50.33 ± 0.36	66.41 ± 0.63
Relation Net [28]	Metric	Relation Measure	50.44 ± 0.82	65.32 ± 0.70
Max-entropy RL [3]	Metric	Cosine Similarity	51.03 ± 0.78	67.96 ± 0.71
TSNN (Ours)	Metric	Image-to-Class Metric	$\mathbf{55.80 \pm 0.95}$	$\mathbf{70.16 \pm 0.82}$

Table 2. The mean accuracies on Stanford Dogs and CUB-200 and with 95% confidence intervals. $*$: Results re-implemented in the same setting for a fair comparison. The best-performing approach is highlighted.

Model	Type	*Stanford Dogs*		*CUB-200*	
		5-way Accuracy(%)			
		1-shot	5-shot	1-shot	5-shot
Matching Nets$*$ [29]	Metric	35.68 ± 0.98	48.64 ± 1.05	45.26 ± 1.03	58.47 ± 1.02
Prototypical Nets$*$ [27]	Metric	36.59 ± 1.03	49.02 ± 1.00	38.04 ± 1.01	51.87 ± 0.99
Relation Net$*$ [28]	Metric	43.50 ± 0.84	55.84 ± 0.74	52.22 ± 0.98	64.03 ± 0.76
Max-entropy RL$*$ [3]	Metric	45.31 ± 0.82	57.49 ± 0.78	52.47 ± 0.87	64.32 ± 0.83
GNN$*$ [8]	Metric	46.06 ± 0.90	61.89 ± 0.95	51.64 ± 0.89	63.17 ± 0.94
TSNN (Ours)	Metric	$\mathbf{48.62 \pm 0.99}$	$\mathbf{63.45 \pm 0.84}$	$\mathbf{57.02 \pm 0.98}$	$\mathbf{70.33 \pm 0.79}$



(a) 5-way 1-shot (b) 5-way 5-shot (c) 5-way 10-shot

Figure 6. Comparison with single-stream test accuracies of three settings on *mini*ImageNet and CUB200.

4.4 Fine-grained Few-shot Classification

Fine-grained image classification [22, 9, 20, 30] tasks are more challenging than basic image classification tasks, because they aim to recognize hundreds of subclasses under the same basic-level class. In the fine-grained few-shot classification, the model is required to have a good ability to distinguish local fine feature differences and good generalization performance due to lacking support samples.

Baselines. For the experiments both on Stanford Dogs and CUB-200, we choose the same five metric-learning based baselines as *mini*ImageNet, including Matching Nets [29], Prototypical Nets [27], Relation Net [28], GNN [8] and Maximum-entropy Reinforcement Learning [3]. We re-evaluate the performance of these models on Stanford Dogs and CUB-200 for a fair comparison.

Results on Stanford Dogs. Experimental results on Stanford Dogs are given in Table 2. We find that the accuracies on Stanford Dogs are lower than that on *mini*ImageNet for all the models. The reason is that the Stanford Dogs dataset contains 120 classes of dogs, many of which have many similar or identical features and confusing multiple objects. However, our approach consistently performs better than other baselines in this case.

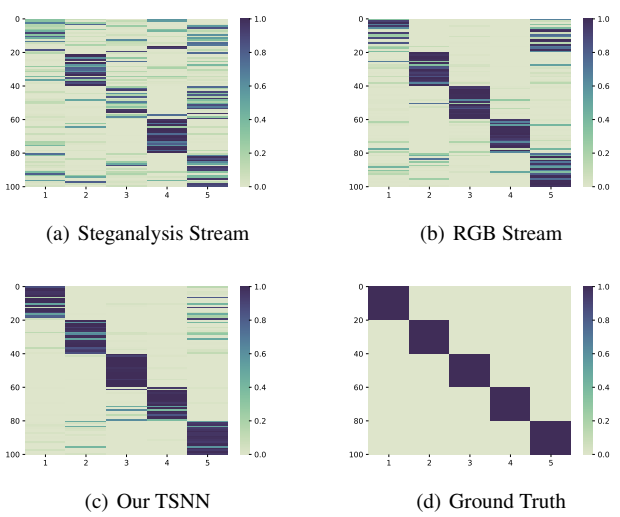
Results on CUB-200. For another fine-grained dataset CUB-200, we also perform experiments with the same settings on Stanford Dogs and the experimental results are recorded in the Table 2. Compared with five competitive metric-learning based approaches, the proposed TSNN leads to some persistent improvements in all the settings. We find that Relation Net and Max-entropy RL also achieve not bad results on CUB-200, the former converts RGB feature maps into relationship scores, and the latter learns a Maximum Entropy Sampler for few-shot learning based on reinforcement learning. Instead, a two-stream framework and a more effective distance metric are adopted in our work. We gain the largest improvement over the second-best approach Max-entropy RL* by 6.01% in 5-way 5-shot setting. In this section, we answer the second question, our TSNN can continuously achieve superior performance beyond the related baselines even on fine-grained datasets.

4.5 Analysis and Discussion

4.5.1 Single-stream Experiments and Analysis

The above experimental results show that our two-stream network architecture can achieve good performance in few-shot classification, but the performance of single-stream input is not clear. Does the steganalysis stream play an equal role in different datasets? Here, we conduct single-stream experiments and analyse the results on *mini*ImageNet and CUB200. For the specific training parameters, in one training episode, we execute 5-way 1-shot, 5-way 5-shot and 5-way 10-shot with 15, 10 and 5 query samples each class respectively. Figure 6 shows the comparison of single-stream experimental results of three settings on *mini*ImageNet and CUB200.

Firstly, in all experiments on two datasets, the test accuracy of a single steganalysis stream has always been lower than that of a single RGB stream. This is reasonable because RGB images contain deep semantic information, which is much more important than local steganalysis features in the classification tasks. What's more, the performance of the single steganalysis stream on the CUB200 is far less than that on the *mini*ImageNet in all the settings. For instance, in Figure 6(a), the single steganalysis stream attains only 40.46% test



(a) Steganalysis Stream (b) RGB Stream

(c) Our TSNN (d) Ground Truth

Figure 7. Visualization of experimental results on CUB200 of the 5-way 5-shot setting. The horizontal axis represents labels of 5 classes, and the vertical axis represents 100 query images selected from 5 classes.The darker the color, the greater the probability of belonging to corresponding class.

accuracy on CUB200, which is much less than the 51.10% test accuracy on *mini*ImageNet. This is because CUB200 is a fine-grained image classification dataset, and the steganalysis features of delicate differences tend to interfere with each other. In this case, the performance of few-shot classification is definitely poor if we only rely on local steganalysis features.

Secondly, it can be clearly seen that although a single steganalysis stream performs very poor on CUB200, our TSNN performs well. In other words, the gain effect of the steganalysis stream is more pronounced on CUB200 than *mini*ImageNet. For the example in Figure 6(b), in the 5-way 5-shot setting, the test accuracy of a single RGB stream on CUB200 is about 67.23%, and the accuracy is improved by 3% after adding steganalysis stream. However, the test accuracy only increases from 68.88% to 70.16% on *mini*ImageNet. Why can the two-stream network achieve such a large gain on CUB200? On the one hand, the RGB features learn deep semantic information, and a large number of futile steganalysis features are eliminated under the coordination of the features of RGB images. On the other hand, the key steganalysis features from SRM help analyze the delicate differences, and learn rich features for few-shot classification. So far, we have answered the third question at the beginning of this section.

Thirdly, we can see that the advantage of transduction narrows with the shots increase. In other words, the gain from steganalysis stream gradually decreases since more labeled data are used. Taking the results on *mini*ImageNet as an example, in the 5-way 1-shot setting, the accuracy of the two-stream increases from 53.58% of RGB stream to 55.80%. However, in the 5-way 5-shot and 5-way 10-shot settings, the improvements become less and less. This is because when the number of support samples is large enough, the RGB images can contain enough information to improve the generalization performance of the model. It also shows that steganalysis features can play an important role when lacking support samples.

Result Visualization. In order to demonstrate the training results of our classifier more intuitively, we visualize the experimental test results on CUB200 of the 5-way 5-shot setting. As shown in Figure 7,

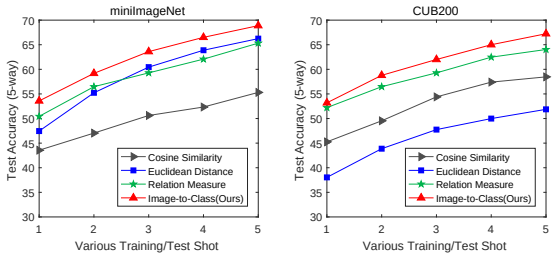


Figure 8. Performance with different metrics on *mini*ImageNet and CUB200 of the 5-way K-shot setting ($K = 1, 2, 3, 4, 5$).

we randomly select 5 samples each class to form a support set, and 20 queries each class to form a query set (i.e., 100 queries in total). We show the classification performance of a single steganalysis stream, a single RGB stream, our TSNN, and the ground truth respectively. The darker the color in the graph, the greater the probability of belonging to that class. It can be clearly seen from the figure that our TSNN performs better than a single stream.

4.5.2 *Effect of Metrics*

It is obvious that the choice of different metrics will have effects on the performance of few-shot classification. To further investigate this, we conduct 5-way K-shot ($K = 1, 2, 3, 4, 5$) experiments with different metrics on *mini*ImageNet and CUB200. It should be noted that we use the same CNNs to extract features in order to ensure the fairness of the experiments.

Figure 8 shows that our learnable ICDM performs consistently better than other metrics with varying shots, because it can measure the similarities between queries and categories adaptively. The performance of Euclidean Distance is significantly reduced on the fine-grained dataset CUB200 due to fine differences between categories. Another observation is that the Relation Measure also performs well on both two datasets. It is also a nonlinear metric which converts the similarities between images into relationship scores. However, it doesn't measure the probabilities that queries belong to each category directly. It turns out that our ICDM is more effective for few-shot classification than other metrics.

5 CONCLUSION

In this paper, we propose a novel two-stream network by using both an RGB stream and a steganalysis stream to learn richer features for few-shot classification. Experimental results confirm that the two streams are complementary when lacking support samples. Meanwhile, the proposed nonlinear Image-to-Class Deep Metric consistently performs well on different datasets compared with the related competitive approaches. In the future, we would like to move forward to apply the current framework in other applications such as person re-identification and relation classification.

ACKNOWLEDGEMENTS

This work was supported in part by NSFC-Shenzhen Robot Jointed Founding under Grant U1613215, in part by the Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing), and in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010137001.

REFERENCES

[1] Yoshua Bengio, 'Deep learning of representations for unsupervised and transfer learning', in *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 17–36, (2012).
[2] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei, 'Memory matching networks for one-shot image recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4080–4088, (2018).
[3] Wen-Hsuan Chu, Yu-Jhe Li, Jing-Cheng Chang, and Yu-Chiang Frank Wang, 'Spot and learn: A maximum-entropy patch sampler for few-shot image classification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6251–6260, (2019).
[4] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, et al., 'Natural language processing (almost) from scratch', *Journal of Machine Learning Research (JMLR)*, **12**(Aug), 2493–2537, (2011).
[5] Li Fei-Fei, Rob Fergus, and Pietro Perona, 'One-shot learning of object categories', *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **28**(4), 594–611, (2006).
[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine, 'Model-agnostic meta-learning for fast adaptation of deep networks', in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 1126–1135. JMLR. org, (2017).
[7] Jessica Fridrich and Jan Kodovsky, 'Rich models for steganalysis of digital images', *IEEE Transactions on Information Forensics and Security (TIFS)*, **7**(3), 868–882, (2012).
[8] Victor Garcia and Joan Bruna, 'Few-shot learning with graph neural networks', *arXiv preprint arXiv:1711.04043*, (2017).
[9] Efstratios Gavves, Basura Fernando, Cees GM Snoek, Arnold WM Smeulders, and Tinne Tuytelaars, 'Local alignments for fine-grained categorization', *International Journal of Computer Vision (IJCV)*, **111**(2), 191–212, (2015).
[10] Spyros Gidaris and Nikos Komodakis, 'Dynamic few-shot visual learning without forgetting', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4367–4375, (2018).
[11] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, 'Speech recognition with deep recurrent neural networks', in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649. IEEE, (2013).
[12] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, et al., 'Deep neural networks for acoustic modeling in speech recognition', *IEEE Signal Processing Magazine*, **29**(6), 82–97, (2012).
[13] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio, 'Learning to remember rare events', *arXiv preprint arXiv:1703.03129*, (2017).
[14] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, 'Novel dataset for fine-grained image categorization: Stanford dogs', in *Proceedings of CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, (2011).
[15] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).
[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, 'Siamese neural networks for one-shot image recognition', in *Proceedings of ICML workshop on deep learning*, volume 2, (2015).
[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks', in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, (2012).
[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep learning', *Nature*, **521**(7553), 436, (2015).
[19] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel, 'A simple neural attentive meta-learner', *arXiv preprint arXiv:1707.03141*, (2017).
[20] Maria-Elena Nilsback and Andrew Zisserman, 'Automated flower classification over a large number of classes', in *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729. IEEE, (2008).
[21] Sinno Jialin Pan and Qiang Yang, 'A survey on transfer learning', *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, **22**(10), 1345–1359, (2009).
[22] Yuxin Peng, Xiangteng He, and Junjie Zhao, 'Object-part attention model for fine-grained image classification', *IEEE Transactions on Image Processing*, **27**(3), 1487–1500, (2017).
[23] Dhanesh Ramachandram and Graham W Taylor, 'Deep multimodal learning: A survey on recent advances and trends', *IEEE Signal Processing Magazine*, **34**(6), 96–108, (2017).
[24] Sachin Ravi and Hugo Larochelle, 'Optimization as a model for few-shot learning', (2016).
[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, et al., 'Imagenet large scale visual recognition challenge', *International Journal of Computer Vision (IJCV)*, **115**(3), 211–252, (2015).
[26] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap, 'Meta-learning with memory-augmented neural networks', in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1842–1850, (2016).
[27] Jake Snell, Kevin Swersky, and Richard Zemel, 'Prototypical networks for few-shot learning', in *Advances in Neural Information Processing Systems (NIPS)*, pp. 4077–4087, (2017).
[28] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, et al., 'Learning to compare: Relation network for few-shot learning', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1199–1208, (2018).
[29] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., 'Matching networks for one shot learning', in *Advances in Neural Information Processing Systems (NIPS)*, pp. 3630–3638, (2016).
[30] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, et al., 'Learning fine-grained image similarity with deep ranking', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1386–1393, (2014).
[31] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, et al., 'Caltech-ucsd birds 200', (2010).
[32] Jason Weston, Sumit Chopra, and Antoine Bordes, 'Memory networks', *arXiv preprint arXiv:1410.3916*, (2014).