IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 11, 2020, accepted July 25, 2020, date of publication July 29, 2020, date of current version August 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012720

Robust Compare Network for Few-Shot Learning

YIXIN YANG, YANG LI, RUI ZHANG, JIABAO WANG, AND ZHUANG MIAO
Command and Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China

Corresponding author: Rui Zhang (3959966@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61806220, and in part by the Frontier Science and Technology Innovation Project of Army Engineering University of PLA.

ABSTRACT Making machines learn like humans is the ultimate goal of artificial intelligence. Few-shot learning attempts to simulate the learning mechanism of humans, which is a task that can learn novel concepts from very few labeled samples. Due to the lack of sufficient labeled training data, existing methods often increase the risk of over-fitting or cause a considerable gap between clean and augmented data. To solve these problems, we present a novel compare network to perform robust few-shot learning in a meta-learned end-to-end manner. Specifically, we argue that it is desirable to learn a robust encoder that can draw inferences about other cases from one example. To this end, we improve the accuracy of few-shot learning by mining the internal mechanism of deep networks, which can leverage label information more effectively. By introducing shift-invariant blocks and a self-attention block in our architecture, all these components are seamlessly integrated into our framework, which can give feedback to each other without data augmentation. Furthermore, we provide ablative analyses of different blocks to help understand how each term contributes to performance. Extensive experiments demonstrate that our method can provide more robust results and outperform state-of-the-art few-shot learning methods.

INDEX TERMS Few-shot learning, shift-invariant, attention, compare network.

I. INTRODUCTION

Making machines learn like humans is the ultimate goal of artificial intelligence [1]–[3]. Although deep learning models have made an unprecedented breakthrough in the field of image recognition [4]–[9], these methods often require enormous supervision information and many iterations to achieve such performance. Different from machine learning algorithms, people can learn a new concept from just few examples [10]. For example, one picture with a new object can be very impressive for children to generate richer representations. On the other hand, without enough supervised information, deep learning models often suffer the over-fitting problem, which will lead to poor generalization results.

Few-shot learning (FSL) [11]–[13] attempts to simulate the learning mechanism of the human brain, which can learn novel concepts from very few labeled samples. Due to the lack of sufficient labeled training samples, FSL is often formulated as transfer learning problems [14], [15] or meta-learning problems [16], [17]. From the perspective of transfer learning, FSL methods can learn a deep model on common classes with sufficient samples first, and then transform the model into novel classes based on only a few labelled samples. On the other hand, some FSL approaches based on meta-learning follow the key idea of learning-to-learn mechanism. Specifically, it samples FSL tasks from the base training set and optimizes the network to perform well on all these subtasks [18]. Although meta-learning and transfer learning methods have achieved promising results for FSL, most of these methods still can not generalize better on unseen tasks with few examples [19]–[21].

Since the lack of labeled samples is the core problem for FSL, data augmentation methods [22], [23] are often be used as an effective assistant to solve FSL problems. Classical data augmentation methods often perform one or some operations on the input images from different transformations, such as scale, crop, flip, cut out, and elastic distortions [24], [25]. Intuitively, data augmentation methods can teach a deep model about robustness, which is used to overcome the weakness of rotational invariance. However, increasing the training data can also increase the risk of over-fitting and cause a considerable gap between clean and augmented data [26]. In addition, the fundamental issue still remains: Data augmentation will generate a lot of data, which is inefficient for machines to learn.

The associate editor coordinating the review of this manuscript and approving it for publication was Shangce Gao.

To deal with the above challenge, we propose an alternative data augmentation method for FSL. In fact, our method does not augment data at all. Instead, we improve the accuracy of FSL by mining the internal mechanism of deep networks, which can make FSL to leverage label information more effectively. We argue that it is desirable to learn a robust encoder, which can draw inferences about other cases from one example. To this end, we introduce shift-invariant blocks and a self-attention block in our architecture (as shown in Fig. 1 (b)).

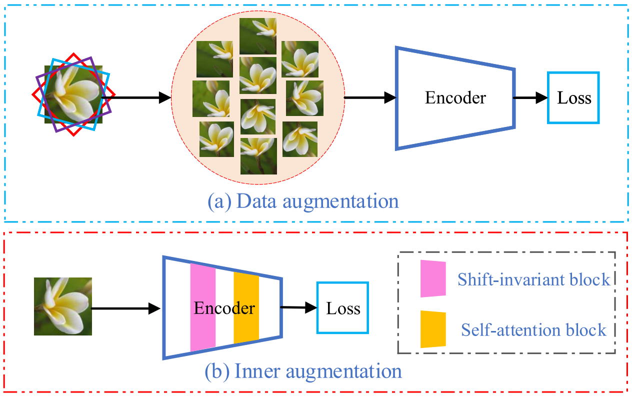

FIGURE 1. Conceptual comparison of two data augmentation mechanisms. These two kinds of mechanisms are different in how the encoder is trained. (a): The encoder is trained by the augmented data. (b): The encoder is trained by the original data but the encoder can leverage label information more effectively.

Specifically, we propose a novel Robust Compare Network (RCN) as a way of "inner data augmentation" mechanism for FSL. As illustrated in Fig. 2, our method is composed of three key modules: (1) Embedding module: a deep neural network with shift-invariant blocks and a self-attention block to learn invariance visual features. (2) Compare module: one convolutional network with shift-invariant blocks to learn feature relation between images. (3) Loss function module: a loss function to measure whether the compare features are from the same categories or not. All the three components are seamlessly integrated into our RCN method, which can give feedback to each other without data augmentation in an end-to-end way.

The main contributions of our RCN method are summarized as follows:

- We present a novel compare model to perform robust FSL in a meta-learned end-to-end manner. This can be seen as an alternative data augmentation method for FSL.
- We introduce shift-invariant blocks and a self-attention block in our framework, which can leverage label information more effectively for FSL. We provide ablative analyses of different blocks to help understand how each term contributes to the performance.
- Extensive experiments show that our RCN method can outperform state-of-the-art methods in few-shot recognition applications.

The rest of this paper is organized as follows. In Section II, we discuss the related work of FSL. Section III presents the details of our RCN architecture. Section IV shows the experimental results, and Section V concludes this paper and points out some possible directions for future pursuit.

II. RELATED WORK

The goal of FSL [11]–[13] is to simulate the learning mechanism of the human brain, which can learn novel concepts from very few labeled samples. Earlier works on FSL tended to explore a variational Bayesian method on the basis of prior knowledge [11] or introduce a generative model to simulate human learning [27]. Recent FSL approaches most follow the meta-learning (learning-to-learn) framework, which samples FSL tasks from the base training set and optimizes the network to perform well on all these subtasks [16], [17]. Under this perspective, various meta-learning based FSL methods are proposed, which can be generally divided into three main categories: metric-based methods, optimization-based methods, and data augmentation based methods.

A. METRIC-BASED METHODS

Metric learning approaches attempt to learn a deep representation with a metric, which can preserve the semantic relationship of the original data. As the name suggests, metric-based methods always compare positive and negative examples to learn feature representation. For example, matching networks [28] propose an attention mechanism to classify query samples with cosine distance, which can be viewed as a weighted nearest-neighbor classifier applied within an embedding space. Prototypical networks [29] propose to represent classes by the mean of the training samples in a representation space, which can learn a metric space by computing Euclidean distances to prototype representations of each class. Relation networks [30] further generalize this framework by learning a deep non-linear distance metric for computing the similarity of different samples. And Cao and Zhang [31] introduces a novel semantic alignment loss to compare relations, which is robust to content misalignment. However, the comparison ability of these methods is still limited due to the problem of unable to understand the spatial relationship between features [32]. Therefore, we further introduce shift-invariant blocks and a self-attention block in our framework to learn a more robust metric ability, which can avoid this problem.

B. OPTIMIZATION-BASED METHODS

Besides explicitly considering the semantic relationship of data, optimization-based methods focus on the optimization process over support-set during FSL. The MAML [33] approach attempts to continuously update the original parameter using the fine-tuning manner. The success of this strategy depends heavily on an appropriate based model, and the result of the fine-tuning operation can not lead to a qualitative leap with a few samples. The latent embedding optimization method [34] overcomes the over-fitting

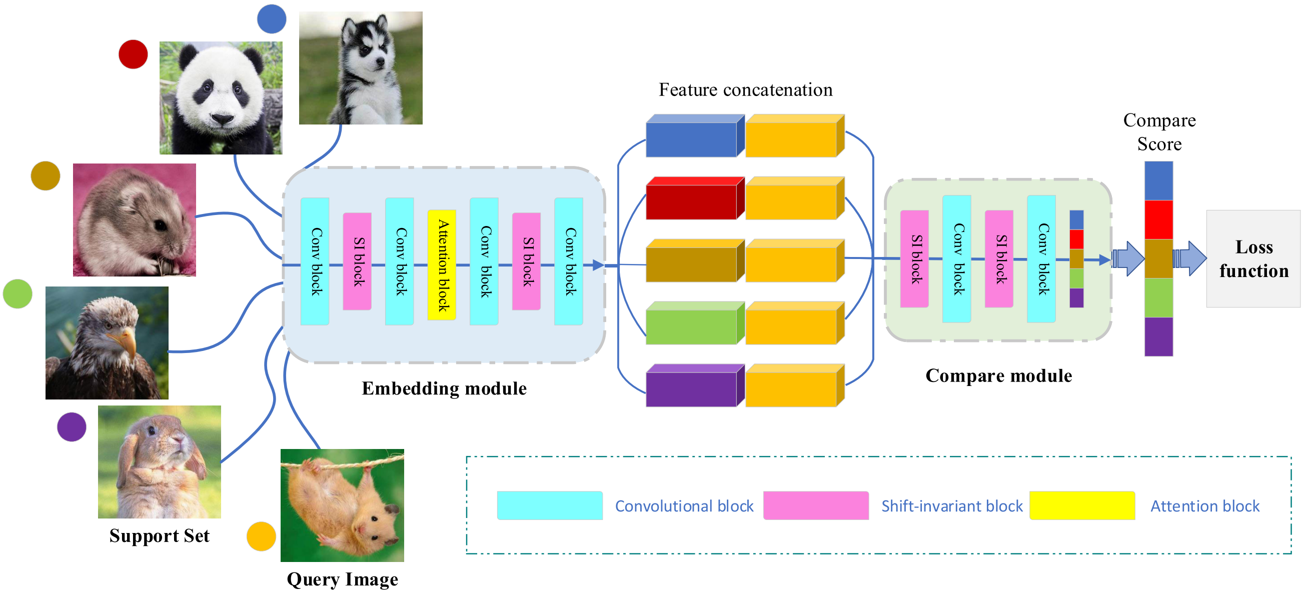

FIGURE 2. An overview of our method under the ***C-way K-shot*** **($C = 5, K = 1$) object recognition task with one query sample. There are three parts of our method: embedding module, compare module, and loss function module.**

problem by learning a initial representation of model parameters and implementing gradient descent as the adaptation process in this low-dimensional latent space. MetaOptNet [21] exploits two properties of linear classifiers, which use high-dimensional embeddings with improved generalization at a modest increase in computational overhead. In this paper, we introduce a simple smooth L_1 loss to enhance the generalization ability of our method.

C. DATA AUGMENTATION BASED METHODS

Another line of FSL try to utilize data augmentation mechanism. The classical data augmentation methods often perform different transformations on the input images, such as scale, crop, flip, cut out, and elastic distortions [24], [25]. More sophisticated augmentation methods have been explored in FSL as well. For example, Wang *et al.* [35] proposes to combine a meta-learner with a ''hallucinator'' that produces additional training examples. Zhong *et al.* [36] randomly selects a rectangle region in an image and erases its pixels with random values, which reduces the risk of over-fitting and makes the model robust to occlusion. Hariharan and Girshick [37] proposes a regularization technique to hallucinate additional training examples for data-starved classes. Chen *et al.* [38] proposes to directly synthesize instance features by leveraging semantics using a novel auto-encoder network called dual TriNet. Wu *et al.* [39] proposes a position-aware relation network, which uses a deformable feature extractor to discover the diversity between data.

The most relevant works of existing methods are [30] and [39]. However, they ignored the robustness of the network. By contrast, we apply the shift-invariant and self-attention mechanisms, which can draw inferences about other cases from one example. This new method requires no additional learning cost and is more generalized.

III. METHODOLOGY

In this section, firstly, we give a formal definition to FSL. Then, we introduce the overall framework of our network. Finally, we describe the shift-invariant block and self-attention block in our architecture in detail.

A. PROBLEM DEFINITION

FSL is actually a weakly supervised learning task [13], which aims to achieve the image classification with a data set $\mathcal{D} = \{\mathcal{D}^{\text{train}}, \mathcal{D}^{\text{support}}, \mathcal{D}^{\text{test}}\}$. Formally, given a labeled dataset $\mathcal{D}^{\text{train}}$ with a large amount of images in each class, the goal of FSL is to learn concepts in novel classes $\mathcal{D}^{\text{new}} = \{\mathcal{D}^{\text{support}}, \mathcal{D}^{\text{test}}\}$ with a few samples in each class. In a *N-way K-shot* few-shot task, the support set $\mathcal{D}^{\text{support}}$ contains K labelled samples in N different classes, the test set $\mathcal{D}^{\text{test}}$ contains Q samples in N different classes. And the goal is to classify the $N \times Q$ unlabelled samples into different N classes. More details about FSL definition can be found in [9], [30].

Episodic training mechanism is an effective way to train the network [28]. In each training iteration, we randomly select C classes from the training set $\mathcal{D}^{\text{train}}$ with K labelled samples to act as the *sample* set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{m}$ $(m = K \times C)$. At the same time, we choose the remainder of those C classes' sample to serve as the *query* set $\mathcal{Q} = \left\{\left(x_j, y_j\right)\right\}_{j=1}^{n}$. This *sample* set $\mathcal{S}$ and *query* set $\mathcal{Q}$ are designed to mimic the support set $\mathcal{D}^{\text{support}}$ and the test set $\mathcal{D}^{\text{test}}$. The model trained from *sample* set $\mathcal{S}$ and *query* set $\mathcal{Q}$ can be further fine-tuned using the support set $\mathcal{D}^{\text{support}}$.

However, due to the different label space between the training set $\mathcal{D}^{\text{train}}$ and the novel set $\mathcal{D}^{\text{new}}$, the performance of FSL is usually not satisfactory. To address this problem, meta-learning is used on the training set, which is to extract the prior knowledge for better FSL. Nevertheless, few-shot samples inevitably contain noise through acquisition, which makes meta-learning unstable. Therefore, in this paper, we attempted to perform a novel contrastive learning model, whose robustness is directly built in the architecture. Our method extends traditional contrastive learning in three key ways: it provides the shift-invariant feature for relation network, it focuses on the global contextual information of each input, and it uses a more powerful loss function. Each of these modifications provides improvements over other methods.

B. NETWORK ARCHITECTURE

As illustrated in Fig. 2, our Robust Compare Network (RCN) consists of three modules: an embedding module f_φ, a compare module g_ϕ, and a loss function module. In the embedding module, we use four-layer convolutional network with shift-invariant blocks and an attention block to extract features. Let $f_\varphi(x_i)$ and $f_\varphi\left(x_j\right)$ denote the features maps of one support image x_i and a query image x_j, respectively. Then, we concatenate the feature maps $f_\varphi(x_i)$ and $f_\varphi\left(x_j\right)$ in depth.

In order to determine whether x_i and x_j are from the same class, the combined feature maps $\mathcal{C}\left(f_\varphi(x_i), f_\varphi\left(x_j\right)\right)$ are feed into the compare module g_ϕ. Concretely, the architecture of compare module g_ϕ is composed of two shift-invariant blocks and two convolution blocks. The compare module g_ϕ will produce a compare score, which is in the range of 0 to 1 for representing the similarity of x_i and x_j.

Therefore, in the 5-*way* 1-*shot* setting, we can generate 5 compare scores $r_{i,j}$ for x_i and x_j,

$$r_{i,j} = g_\phi\left(\mathcal{C}\left(f_\varphi(x_i), f_\varphi\left(x_j\right)\right)\right), \quad i = 1, 2, \ldots, 5 \tag{1}$$

In K-*shot* setting, where $K > 1$, we element-wise sum the feature maps of all samples from each training class as proposed in [30]. Thus, for one-shot or few-shot setting, the quantity of compare scores for every query is always the same.

Smooth L_1 loss (a kind of Huber loss) was used to train our model, which constrain the compare score $r_{i,j}$ to the ground truth. If z is the difference between $r_{i,j}$ and the ground truth, the smooth L_1 loss can be calculated as:

$$\text{smooth}\, L_1(z) = \begin{cases} 0.5z^2 & \text{if } |z| < 1 \\ |z| - 0.5 & \text{otherwise} \end{cases} \tag{2}$$

It is worth to notice that the smooth L_1 loss is different from the standard choice mean square error (MSE) loss in FSL as in [30]. MSE loss is convenient to make a derivation and has a stable result because each point is continuously smooth. However, once there is an abnormal and far from the center point, MSE will appear as a gradient explosion. In other words, MSE is stable but not robust. Therefore, in this paper, we introduce the smooth L_1 loss to avoid gradient explosion. Smooth L_1 loss is actually a piecewise function. Based on the MSE, the smooth L_1 loss uses the mean absolute mechanism that can adaptively employ the abnormal points and have satisfying robustness and stability.

C. SHIFT-INVARIANT FOR ENHANCING ROBUSTNESS

Feature extractor is one of the most significant steps for few-shot object recognition. However, convolutional networks are not shift-invariant [40]. A small input translation will cause unstable output for FSL. In practice, most existing FSL methods adopt data argumentation to achieve shift-invariance [38], [39]. In this paper, we adopt shift-invariant blocks in our embedding module f_φ and compare module g_ϕ without data augmentation.

We assume the output of one middle layer l of our network to be $F_l \in \mathbb{R}^{H\times W\times C}$ with spatial resolution $H \times W$ and C channels. Shift-invariant means the shifting operation can obtain an identical representation compared to the input.

$$F(x) = F\left(\text{Shift}_{\Delta h,\Delta w}(x)\right) \quad \forall(\Delta h, \Delta w) \tag{3}$$

where $F(.)$ represents a feature extractor, Δh and Δw represent the shift variables.

In convolutional networks, the max-pooling operation is not shift-invariant. Formally, we can decompose the max-pooling operation into two functions:

$$\text{MaxPool}_{k,s} = \text{Subsample}_s \circ \text{Max}_k \tag{4}$$

where k, s are the kernel and stride, respectively.

By combining the blur and subsample operations, the defect of subsampling may be compensated by adding an anti-aliasing filter with kernel $m \times m$, which is denoted as Blur_m, as illustrated in Fig. 3. Specifically, the max operation (padding=1, kernel=2) does not change the feature dimensionality, and the anti-aliasing filter (a low-pass filter) is used to compute a convincing feature map circularly without dimension reduction. Finally, a common pooling operation do downsampling in a shift-invariant max-pooling block. Therefore, the shift-invariant max-pooling can be defined as:

$$\begin{aligned} \text{MaxPool}_{k,s} &\rightarrow \text{Subsample}_s \circ \text{Blur}_m \circ \text{Max}_k \\ &= \text{BlurPool}_{m,s} \circ \text{Max}_k \end{aligned} \tag{5}$$

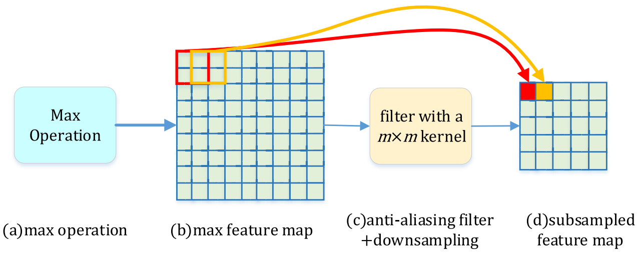

FIGURE 3. Shift-invariant computation combined of max-pooling operation.

As illustrated in Fig. 2, we applied the shift-invariant max-pooling four times in our framework, which is used twice in the embedding module and twice in the relation module. By applying the shift-invariant model, we can further improve the consistency in FSL when the data distribution is changing. In addition, as far as we known, we are the first one to apply this method to improve FSL.

D. ATTENTION MECHANISM FOR BETTER FEATURE REPRESENTATIONS

Due to the inadequate data in few-shot setting, traditional convolution operations can not fully explore global contextual information. We attempt to mitigate this problem by introducing a self-attention mechanism into our embedding module, as illustrated in Fig. 4. Our attention mechanism contains two attention blocks, a position attention block and a channel attention block, with a common input $\mathbf{A} \in \mathbb{R}^{H\times W\times C}$. Two identical dimensions of output are fed into a sum fusion operation without changing dimension as well.

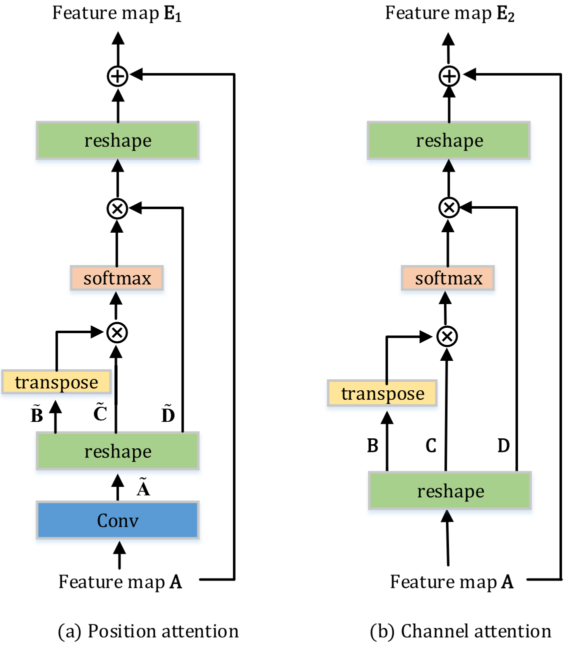

FIGURE 4. Position attention block and channel attention block are illustrated in (a) and (b).

1) POSITION ATTENTION BLOCK

All classification tasks are based on the clearly distinguishable feature map. However, recent works [41]–[43] find the features generated by FCNs [44] might lead to misclassification. The position attention module attempts to generate a spatial related feature with a long range of contextual characteristic.

As illustrated in Fig. 4 (a), we first feed the feature $\mathbf{A} \in R^{H\times W\times C}$ into a convolution layer and get new feature maps $\tilde{\mathbf{A}} \in R^{H\times W\times C}$. Then, we reshape the output into three new feature maps $\{\tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{D}}\} \in \mathbb{R}^{C\times N}$, where $N = H \times W$. Secondly, we multiply the transpose matrix $\tilde{\mathbf{B}}$ by matrix $\tilde{\mathbf{C}}$, and feed their result into a softmax layer to obtain a spatial attention matrix. We multiply the spatial attention matrix by $\tilde{\mathbf{D}}$ and reshape the result to $\mathbb{R}^{C\times H\times W}$. Finally, we perform an element-wise sum with features $\mathbf{A}$ and obtain the position attention output $\mathbf{E_1} \in \mathbb{R}^{C\times H\times W}$ as follows:

$$\mathbf{E_1} = \alpha * reshape(\tilde{\mathbf{D}} \times softmax(\tilde{\mathbf{B}}^T \times \tilde{\mathbf{C}})) + \mathbf{A} \qquad (6)$$

where α is set to 0 at the beginning and gradually changes during the training process.

2) CHANNEL ATTENTION BLOCK

The channel maps can reflect internal dependencies between different semantic classes [43]. We attempt to improve the feature representation of classification by obtaining channel attention maps (as illustrated in Fig.4 (b)). Different from the position attention module, we firstly reshape and copy the feature $\mathbf{A}$ into three new features $\{\mathbf{B}, \mathbf{C}, \mathbf{D}\} \in \mathbb{R}^{C\times N}$ without a convolutional operation. Secondly, we multiply the transpose of $\mathbf{B}$ by $\mathbf{C}$. After that, the result is sent into a softmax layer to obtain a channel attention matrix. We multiply the channel attention matrix by $\mathbf{D}$ and reshape the result to $\mathbb{R}^{C\times H\times W}$. Finally, we perform an element-wise sum with features $\mathbf{A}$ and obtain the channel attention output $\mathbf{E_2} \in \mathbb{R}^{C\times H\times W}$ as follows:

$$\mathbf{E_2} = \beta * reshape(softmax(\mathbf{C} \times \mathbf{B}^T) \times \mathbf{D}) + \mathbf{A} \qquad (7)$$

where β is set to 0 at the beginning and gradually changes during the training process.

The self-attention mechanism enhances the global correlation of local features, which allows different classes to clear the boundaries and prevent errors from being identified by ignoring some small features. For FSL, a clear and definite feature map can alleviate the recognition difficulty, and improve the accuracy of the model. We introduce it into the embedding module and use a small amount of memory and computational cost, but obtain a good result.

IV. EXPERIMENTS

To evaluate the effectiveness of our method, we perform few-shot classification experiments on the Omniglot dataset [1]. In the next subsections, we first introduce the dataset and implementation details, then we perform a series of ablation experiments on the Omniglot dataset. Finally, we report our results.

A. DATASETS AND SETTINGS

The Omniglot dataset [1] is designed to develop human-like learning algorithms. Human participants can achieve an error rate of 4.5% on this dataset. It has 1623 handwritten characters with 24×24 resolution from 50 different alphabets. Each of these characters is drawn online by 20 different people from Amazon's Mechanical Turk (as illustrated in Fig. 5). We select 1200 classes as training set and the rest 423 classes are used as testing set.

Following the training setting of [30], the 5-*way* 1-*shot* has 19 query images for each class. The 5-*way* 5-*shot* contains 15 query images. And the 20-*way* 1-*shot* and 20-*way* 5-*shot*

TABLE 1. The recognition accuracy of different methods on the Omniglot dataset. The best results are highlighted by **bold**, and the second best results are highlighted by underline.

Model	5-*way* 1-*shot*	5-*way* 5-*shot*	20-*way* 1-*shot*	20-*way* 5-*shot*
Mann [45]	82.8%	94.9%	-	-
Convolutional Siamese Nets [46]	96.7%	98.4%	88.0%	96.5%
Matching Nets [28]	98.1%	98.9%	93.8%	98.5%
Siamese Nets with Memory [47]	98.4%	99.6%	95.0%	98.6%
Neural Statistician [48]	98.1%	99.5%	93.2%	98.1%
Meta Nets [49]	99.0%	-	97.0%	-
Prototypical Nets [29]	98.8%	<u>99.7</u>%	96.0%	98.9%
Two-Stage [50]	99.2%	99.5%	97.2%	98.9%
Relation Net [30]	<u>99.6</u>%	**99.8**%	<u>97.6</u>%	<u>99.1</u>%
Sampler-FC [51]	97.4%	99.5%	-	-
Our method	**99.8**%	**99.8**%	**98.1**%	**99.3**%

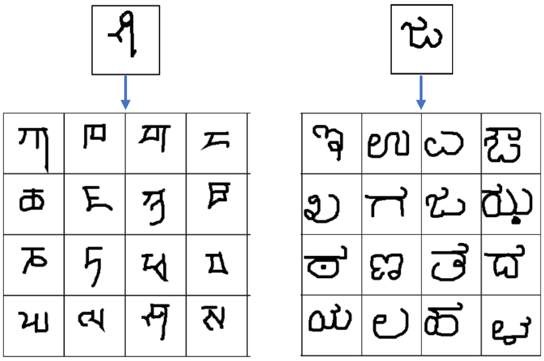

FIGURE 5. A new character is presented on the top. And the goal of FSL is to select another example from the same alphabet amongst other characters.

contain 10 and 5 query images respectively. And we also resized all input images into 28 × 28. Finally, our model is tested over 600 episodes and averaged as the last accuracy.

All experiments are trained from scratch by Pytorch [52] with an initialized learning rate of 0.001 and Adamax optimizer. And the parameter α and β are set to 0 at the beginning and gradually change during the training process. What's more, our models use end-to-end training and finish after 100,000 episodes.

B. MAIN RESULTS

We compare the proposed method with other 10 state-of-the-art FSL approaches, including Mann [45], Convolutional Siamese Nets [46], Matching Nets [28], Siamese Nets with Memory [47], Neural Statistician [48], Meta Nets [49], Prototypical Nets [29], Two-Stage [50], Relation Net [30], and Sampler-FC [51].

The comparative results of different methods are shown in Table 1. From Table 1, we can see that our method can outperform other state-of-the-art FSL methods. More specifically, in the 5-*way* 1-*shot* recognition experiment, our model outperforms the other methods and increase 0.2% than the second best result. For 20-*way* 1-*shot* and 20-*way* 5-*shot*, we improve the accuracy of 0.5% and 0.2% over the second best result. In 5-*way* 5-*shot* recognition task, our method and Relation Net [30] achieve the best result, which is 0.1% higher than the second best result.

There are three main reasons for the good results of our method. First, shift-invariant block increases the robustness of FSL. Second, the attention block enforces the embedding module to pay more attention to the relation information inside images, which encourage our network to learn more essential features. Finally, the smooth L_1 loss boosts the stability of the optimization objective.

C. ABLATION STUDY

In order to better demonstrate the contribution of each module of our method, we compare our method (with smooth L_1 loss, shift-invariant mechanism, and attention mechanism, named Baseline+M1+M2+M3) with three simplified versions: (1) the original compare network with MSE loss (Baseline); (2) the original compare network with smooth L_1 loss (Baseline+M1); (3) the original compare network with smooth L_1 loss and shift-invariant mechanism (Baseline+M1+M2); (4) our method.

The ablative results on the Omniglot dataset are shown in Table 2. And we can have three observations.

Baseline vs. Baseline+M1: The smooth L_1 loss can improve the stability of the model to some extent. The 20-*way* 1-*shot* recognition task useing this loss has a significant improvement, which improves 0.4% than the baseline model. However, the accuracy of 5-*way* 5-*shot* recognition task has a little improvement, which improves 0.04% than the baseline model. This phenomenon shows that the smooth L_1 loss can further improve the performance in more difficult tasks.

TABLE 2. Ablation study results for our method under different setting. Baseline is the original model. "M1" represents smooth L_1 loss function, "M2" represents shift-invariant mechanism, and "M3" represents attention mechanism.

strategies	5-*way*		20-*way*	
	1-*shot*	5-*shot*	1-*shot*	5-*shot*
Baseline	99.63%	99.70%	97.07%	99.13%
Baseline+M1	99.72%	99.74%	97.47%	99.21%
Baseline+M1+M2	99.76%	99.76%	97.99%	99.24%
Baseline+M1+M2+M3	99.84%	99.82%	98.05%	99.29%

Baseline+M1+M2: Shift-invariant mechanism have a significant impact on the 5-*way* 1-*shot* and the 20-*way* 1-*shot* recognition tasks. They improve 0.13% and 0.92% respectively compared to the baseline model.

Baseline+M1+M2+M3: The attention mechanism can further improve the accuracy of FSL. The 5-*way* 1-*shot* and 20-*way* 1-*shot* recognition task achieve 99.84% and 98.05%, which is 0.2% and approximately 1% higher than original model, respectively. In addition, the 5-*way* 5-*shot* recognition task has an obvious improvement than the prior strategies, which is 0.12% than the original model.

D. ROBUST PERFORMANCE ANALYSIS

In FSL, unpleasant shifts and geometric changes of input images will influence the robustness of outputs. We test our method and Relation Net [30] with manual pixel shift conditions to demonstrate the robust ability of our method. The results of 5-*way* 1-*shot* with pixel movement in different directions on the Omniglot dataset are illustrated in Table 3.

TABLE 3. The robust performance of different methods. Left, right, upper, and down denote one pixel shift in this direction.

	Relation Net [30]	Our model
Left	99.53%	99.71%
Right	99.59%	99.74%
Upper	99.60%	99.76%
Down	99.43%	99.79%

From Table 3, we can see that pixel movement greatly degrades the performance of Relation Net [30]. This phenomenon demonstrates that some pixel perturbations in the input image can cause lower stability and robustness for models and decrease the recognition accuracy. However, suffering from some perturbations in the input image, our method still can outperform the Relation Net [30].

E. FEATURE VISUALIZATION

In this subsection, we perform visualization experiments on the Omniglot dataset to demonstrate the credibility of our method. As illustrated in Fig. 6, the results from our model (second rows) obviously better than the results from Relation Net [30] (first rows). Specifically, we find that our method can extract the features from the regions that the target object corresponds to the characters. In addition, our visualization results further reveal the existence of a strong correlation between human attention and the explicit saliency field.

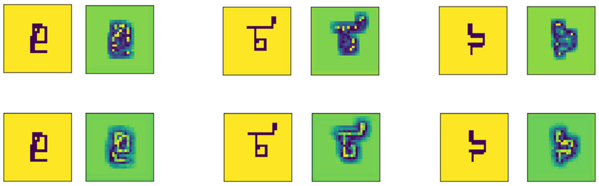

FIGURE 6. The visualization results for the same input images. The first rows are Relation Net [30], and the second rows are our method.

V. CONCLUSION

In this paper, we propose a novel compare network to perform robust FSL in a meta-learned end-to-end manner. Firstly, we introduce shift-invariant blocks and a self-attention block in our framework, which can leverage label information more effectively for FSL. Secondly, we introduce the smooth L_1 loss to avoid gradient explosion. Therefore, our method is more robust and flexible to learn better representations. Finally, we provide ablative analyses of different blocks to help understand how each term contributes to performance. Extensive experiments demonstrates the superiority of our RCN method in FSL recognition applications.

In the future, we will focus on using multi-modality information in designing FSL methods. In addition, we also want to extend the Auto-fpn methods [53] to FSL in an economic, efficient, and effective manner.

REFERENCES

[1] 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, Dec. 2015.

[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behav. Brain Sci.*, vol. 40, p. e253, Apr. 2017.

[3] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 1, no. 1, pp. 1–34, 2020.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. IEEE Conf. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1106–1114.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–4.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[9] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[10] S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum, "Computational rationality: A converging paradigm for intelligence in brains, minds, and machines," *Science*, vol. 349, no. 6245, pp. 273–278, Jul. 2015.

[11] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[12] M. Guo, E. Chou, D.-A. Huang, S. Song, S. Yeung, and L. Fei-Fei, "Neural graph matching networks for fewshot 3D action recognition," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 673–689.

[13] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–10.

[14] T. Tommasi, F. Orabona, and B. Caputo, "Learning categories from few examples with multi model knowledge transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 928–941, May 2014.

[15] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[16] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot Learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.

[17] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," in *Proc. IEEE Conf. Robot Learn.*, Sep. 2017, pp. 357–368.

[18] Y. Chen, X. Wang, Z. Liu, H. Xu, and T. Darrell, "A new meta-baseline for few-shot learning," 2020, *arXiv:2003.04390*. [Online]. Available: http://arxiv.org/abs/2003.04390

[19] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 403–412.

[20] X. Wang, F. Yu, R. Wang, T. Darrell, and J. E. Gonzalez, "TAFE-net: Task-aware feature embeddings for low shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1831–1840.

[21] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10649–10657.

[22] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "F-VAEGAN-D2: A feature generating framework for any-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10267–10276.

[23] A. Li, T. Luo, Z. Lu, T. Xiang, and L. Wang, "Large-scale few-shot learning: Knowledge transfer with class hierarchy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7205–7213.

[24] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[25] Z. Tang, X. Peng, T. Li, Y. Zhu, and D. Metaxas, "AdaTransform: Adaptive data transformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2998–3006.

[26] Z. He, L. Xie, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Data augmentation revisited: Rethinking the distribution gap between clean and augmented data," 2019, *arXiv:1909.09148*. [Online]. Available: http://arxiv.org/abs/1909.09148

[27] B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum, "One shot learning of simple visual concepts," in *Proc. 33th Annu. Meeting Cogn. Sci. Soc.*, 2011, pp. 2568–2573.

[28] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3630–3638.

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

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

[31] C. Cao and Y. Zhang, "Learning to compare relation: Semantic alignment for few-shot learning," 2020, *arXiv:2003.00210*. [Online]. Available: http://arxiv.org/abs/2003.00210

[32] J. Choi, H. Seo, S. Im, and M. Kang, "Attention routing between capsules," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1981–1989.

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

[34] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," 2018, *arXiv:1807.05960*. [Online]. Available: http://arxiv.org/abs/1807.05960

[35] Y.-X. Wang, R. B. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7278–7286.

[36] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

[37] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3037–3046.

[38] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, and L. Sigal, "Multi-level semantic feature augmentation for one-shot learning," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4594–4605, Sep. 2019.

[39] Z. Wu, Y. Li, L. Guo, and K. Jia, "PARN: Position-aware relation networks for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6658–6666.

[40] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7324–7334.

[41] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1743–1751.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[43] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[44] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

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

[46] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. Workshop*, vol. 2, 2015, pp. 1–8.

[47] L. Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–10.

[48] H. Edwards and A. J. Storkey, "Towards a neural statistician," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

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

[50] D. Das and C. S. G. Lee, "A two-stage approach to few-shot learning for image recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 3336–3350, 2020.

[51] W.-H. Chu, Y.-J. Li, J.-C. Chang, and Y.-C.-F. Wang, "Spot and learn: A maximum-entropy patch sampler for few-shot image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6251–6260.

[52] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[53] H. Xu, L. Yao, Z. Li, X. Liang, and W. Zhang, "Auto-FPN: Automatic network architecture adaptation for object detection beyond classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6649–6658.

YIXIN YANG received the B.S. degree from the Army Engineering University of PLA, Nanjing, China, in 2019, where he is currently pursuing the M.S. degree with the Command and Control Engineering College. His current research interests include image classification and the FSL.

YANG LI received the M.S. degree from the PLA University of Science and Technology, Nanjing, China, in 2010, and the Ph.D. degree from the Army Engineering University of PLA, Nanjing, in 2018. He is currently an Associate Professor with the Army Engineering University of PLA. His current research interests include computer vision, deep learning, and image processing.

RUI ZHANG received the Ph.D. degree from the PLA University of Science and Technology, Nanjing, China, in 2004. He is currently a Professor with the Army Engineering University of PLA, Nanjing. His current research interests include data engineering and information fusion.

JIABAO WANG received the Ph.D. degree in computational intelligence from the PLA University of Science and Technology, Nanjing, China, in 2013. He is currently an Assistant Professor with the Army Engineering University of PLA, Nanjing. His current research interests include computer vision and machine learning.

ZHUANG MIAO received the Ph.D. degree from the PLA University of Science and Technology, Nanjing, China, in 2007. He is currently an Associate Professor with the Army Engineering University of PLA, Nanjing. His current research interests include artificial intelligence, pattern recognition, and computer vision.