Contrastive JS: A Novel Scheme for Enhancing the Accuracy and Robustness of Deep Models

Weiwei Xing, Jie Yao, Zixia Liu, Weibin Liu, Shunli Zhang, Liqiang Wang

***Abstract*—Deep learning technologies have been applied in various computer vision tasks in recent years. However, deep models suffer performance decay when some unforeseen data are contained in the testing dataset. Although data enhancement techniques can alleviate this dilemma, the diversity of real data is too tremendous to simulate. To tackle this challenge, we study a scheme for improving the robustness and efficiency of the deep network training process in visual tasks. Specifically, first, we build positive and negative sample pairs based on a class-sensitive strategy. Then, we construct a feature-consistent learning strategy based on contrastive learning to constrain the representations of interclass features while paying attention to the intraclass features. To extend the effect of the consistent strategy, we propose a novel contrastive Jensen–Shannon divergence consistency loss (JS loss) to restrict the probability distributions of different sample pairs. The proposed scheme successfully enhances the robustness and accuracy of the utilized model. We validated our approach by conducting extensive experiments in the domains of model robustness and few-shot object detection (FSOD). The results showed that the proposed method achieved remarkable gains over state-of-the-art (SOTA) methods. We obtained a 3.2% average improvement over the best-performing FSOD method.**

***Index Terms*—Contrastive Learning, Deep Model Robustness, Few-Shot Object Detection, Image Augmentation**

I. Introduction

In recent years, due to the rapid development of technology, various forms of data have emerged. The analysis these data can serve diverse applications and enhance our lives. Among them, image-based and video-based data analysis and processing are some of the most popular research topics [1]. Furthermore, deep learning techniques have been successfully used across a wide spectrum of vision tasks, such as object tracking [2] and person re-identification [3]. Such complex scenarios usually involve the design of specific network modules to extract features from multimedia data and process the extracted features through a designated algorithm to obtain the final results. Although these modules are diverse, most of them rely on some widely used convolutional neural networks. Hence, the stability levels and feature extraction abilities of deep networks become factors that affect the performance of these modules.

Obtaining a well-trained deep network usually requires abundant training data. Additionally, deep networks are fragile, as some tiny distortions can cause their performance to crash [4]. Therefore, a reliable and robust deep visual perception system becomes of vital importance. To solve the stability and efficiency problems of deep models, an increasing number of data sources are being used to accomplish the training process. It is generally believed that machine learning models can learn the distribution characteristics of the input data in this way. However, the learning process is quite complicated, and some mistaken attention paid to minute details is more likely to lead to high sensitivity and instability for the deep learning classifier. It is difficult to ensure that the distribution of the utilized training data is appropriately in line with reality. In practice, the distributions of the training set and the testing set may be biased [5]. At the same time, during the actual testing process, the quality of the testing images can be interfered with by a variety of factors, such as weather, blur, or other factors that are considered uncontrollable. In such circumstances, the data acquisition program may only capture a small part of the underlying data distribution. Due to this weakness, mismatches between the training and testing data are commonplace, and the obtained classifier may be easily fooled by some tiny imperceptible perturbations or corruptions within the query image. This requires a model that is able to generalize robustly across data distribution shifts.

Considering the fragility of deep models, many researchers have begun to challenge the robustness of deep learning classifiers and proposed a variety of training schemes to improve their robustness. The addition of some small corruptions to the data is a typical kind of distortion and is sufficient for subverting existing classifiers. Although techniques for improving corruption robustness remain few, an increasing number of researchers have begun to study this issue. Hendrycks et al. [6] established a new benchmark for evaluating the performance achieved on some common corruptions and found that some models that perform well on clean datasets cannot be used directly on corrupted datasets, as they cannot maintain their performance. Then, they [7] proposed a data processing technique to improve the robustness and uncertainty estimates of image classifiers. This method is effective because it extends the distribution of the training dataset and helps models resist unforeseen corruptions. However, this method pays too much attention to data enhancement and the prediction distributions

This work is supported by the National Natural Science Foundation of China [grant number 61876018].

Weiwei Xing and Shunli Zhang are with the School of Software Engineering, Beijing Jiaotong University, Beijing, 100044, China.

Jie Yao is with the School of Software Engineering, Beijing Jiaotong University, Beijing, 100044, China and the School of Information Management, Beijing Information Science and Technology University, Beijing, 100192, China.

Weibin Liu is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China.

Zixia Liu and Liqiang Wang are with the Department of Computer Science, University of Central Florida, 32816, USA.

Corresponding author: Jie Yao. Email: 17112098@bjtu.edu.cn, jieyao@bistu.edu.cn. Jie Yao and Weiwei Xing contribute equally to this paper.

among inputs from the same category and fails to learn the differences between objects from different categories. On the other hand, while increasing the scale of the training data benefits the performance of the utilized model, it incurs a large extra cost during the training process. Moreover, in the data collection process, there may be a situation where the data of some specific objects cannot be obtained in large quantities. This requires the employed model to have the ability to adequately learn object features from a limited dataset.

To address these issues, we investigate and propose a novel training scheme for deep visual recognition models to provide them with strong robustness and high accuracy in the face of unseen testing data. In addition to the regular optimization process in the network, we employ the contrastive learning technique to learn to distinguish features from different classes. Specifically, we build positive sample pairs using clean samples and their corrupted variants, and we build negative sample pairs using clean samples and samples from different categories. Then, we simultaneously input these sample pairs into the network and construct a feature-consistent learning strategy based on contrastive learning to guide the feature extraction process. Furthermore, we broaden the effect of the contrastive strategy and propose the contrastive Jensen–Shannon divergence consistency loss (JS loss) to restrict the prediction distributions of different sample pairs.

To the best of our knowledge, this is the first study to use contrastive learning for enhancing the robustness and accuracy of a model. We verify the effectiveness of the proposed method in two applications: model robustness and few-shot object detection (FSOD). For the deep model robustness task, we select Augmix [7] as our strong baseline, which achieves state-of-the-art (SOTA) performance on the robustness benchmark [6]. Moreover, we carry out comprehensive ablation studies to investigate the characteristics of our method. For FSOD, we choose few-shot object detection via contrastive proposal encoding (FSCE) [8] as our strong baseline. The comparison results demonstrate that the proposed scheme achieves robustness improvements and has superior object detection performance. In summary, this paper proposes a novel scheme for enhancing the stability and efficiency of deep models, and the major contributions of this paper are summarized as follows.

- A novel and class-sensitive sample pair building strategy is proposed to guarantee that the pairwise information is object-sensitive for positive sample pairs and border-sensitive for negative sample pairs.
- A feature-consistent learning strategy is constructed to constrain the representations of interclass features while paying attention to the intraclass features.
- A novel probability distribution consistency constraint, the contrastive JS loss, is proposed to broaden the effect of the contrastive strategy and restrict the prediction distributions of different sample pairs.
- Extensive experiments are conducted to evaluate the applicability and performance of the proposed scheme in terms of improving the robustness and data effectiveness of the given model.

The rest of this paper is organized as follows. We review the relevant literature in Section II, and then the proposed scheme is elaborated in Section III. Section IV demonstrates the experimental results. Section V discusses the limitations of this study and future work. Section VI draws conclusions about the proposed scheme.

II. Related Work

A. Deep Model Robustness

Convolutional neural networks can better simulate the human visual system than other methods. However, these mechanisms are complicated and different from the traditional human visual system, which can lead to high vulnerability problems [9]. In terms of this issue, a series of studies have focused on using various methods to challenge the robustness of deep neural networks. Some of the research on visual robustness has focused on the main challenge of adversarial samples [4]. Adversarial samples are machine learning model inputs that are deliberately designed by the attacker [10]; these samples lead to model errors by adding some invisible disturbances. Most of these approaches generate adversarial samples by using gradient-based methods. The way that adversarial samples are generated determines that they can only represent limited testing data distributions.

In addition to imperceptible perturbations, small corruptions in the data distribution are sufficient to subvert existing classifiers as well [11]. Vasiljevic et al. [12] found that a model trained from high-quality images would suffer significant performance degradation when applied to images degraded by some blur augmentations due to the mismatches between the training and testing datasets. Lakshminarayanan et al. [13] found that training probabilistic networks by ensembling classifier predictions can improve the resulting prediction performance and achieve higher performance on test examples from known and unknown distributions. Recently, Hendrycks et al. [6] investigated the differences between problems of robustness against adversarial perturbations and corruptions, integrated multiple corruption methods, and proposed a novel benchmark to measure the robustness of models to unseen corruptions. Our work mainly revolves around this benchmark, as it comprehensively includes a variety of corruption styles, such as Gaussian noise, frost, and glass blur.

Data augmentation techniques can yield improved generalization performance because the augmented data they generate can be used as complementary data for model training. Several data augmentation techniques, such as random flipping and cropping, have already been applied to many commonly used deep learning frameworks. Numerous data augmentation methods have been proposed in recent years [14]–[16]. Recently, Ekin et al. [17] proposed a simplified search space to reduce the computational expense of automated augmentation and permitted the removal of a separate proxy task. Hendrycks et al. [7] focused on the situation where the distributions of the training and testing sets are mismatched and proposed a data augmentation technique called Augmix, which utilizes the combinations of different augmentation operations in concert with a consistency loss. Yulin et al. [18] proposed implicit

semantic data augmentation (ISDA) to complement traditional augmentation schemes.

However, most of these methods pay exorbitant attention to data augmentation and fail to distinguish between the features of different categories. If the differences among categories are tiny, the differences among their augmented variants may be even smaller. In this case, focusing solely on the features within classes may not satisfy the model training requirements. Therefore, we need to constrain the interclass features while paying attention to the intraclass features.

B. FSOD

Few-shot learning is the problem of making predictions based on a limited number of samples. Few-shot learning is different from standard supervised learning. It is difficult to obtain comprehensive object features from limited data. Therefore, one of the most important ideas of few-shot learning is "learning to learn" [19]. FSOD helps detectors adapt to unseen classes with few training instances and is useful when manual annotation is time-consuming or when data acquisition is limited [20]. An increasing number of research works have been published to solve this problem.

Chen et al. [21] proposed a low-shot transfer detector (LSTD) that integrates the advantages of both the single-shot detector (SSD) [22] and the Faster region-based convolutional neural network (RCNN) [23] into a unified deep framework. Karlinsky et al. [24] proposed a distance metric learning method that simultaneously learns the backbone network parameters, the embedding space, and the multimodal distribution of each training category. Yang et al. [25] proposed a flexible module called the context transformer, which leverages source-domain object knowledge as guidance and exploits the contexts derived from the target training images. Then, it integrates this relational information to enhance the discriminative ability of the detector. Fan et al. [26] proposed an attention-based region proposal network (RPN) and a multi-relation detector to exploit the similarity between the few-shot support set and query set. Wu et al. [20] proposed a multiscale positive sample refinement (MPSR) approach to generate multiscale positive samples and refine the prediction process at various scales. Karlinsky et al. [27] proposed a differentiable nonparametric star model detection and classification head, named StarNet. Zhu et al. [28] utilized the semantic relational consistency between novel classes and the base classes and introduced explicit relation reasoning to the learning process for novel object detection. These methods take full advantage of the input training data from different perspectives and achieve competitive performance. However, some of them are complicated and difficult to apply to other problems.

Recently, Sun et al. [8] indicated that object proposals with different intersection-of-union (IoU) scores are augmented variants of ground-truth (GT) objects and proposed FSCE to learn contrastive-aware object proposal encodings. This work presented a novel approach for solving the FSOD problem, but their method introduces additional parameters and puts extra pressure on the model training procedure.

C. Contrastive Learning

To avoid the extensive cost of collecting and annotating large-scale datasets, as a subset of unsupervised learning methods, self-supervised learning (SSL) methods have been proposed to learn general image features from large-scale unlabeled data without using any human-annotated labels [29]. SSL includes two main strategies: generative and contrastive methods [30]. Generative methods, including autoencoding (AE) models [31], and hybrid generative models [32], mainly focus on the reconstruction error in the pixel space to learn representations.

Contrastive techniques build representations by learning to encode what makes two things similar or different. For example, Hjelm et al. [33] proposed Deep InfoMax, which maximizes the mutual information between local features and global features. Tian et al. [34] constructed positive and negative sample pairs through multimodal information. He et al. [35] built a dynamic dictionary with a queue and a moving average-based encoder. Chen et al. [36] simplified recently proposed contrastive self-supervised learning algorithms and proposed a simple framework called SimCLR for conducting contrastive learning on visual representations. To address the abovementioned stability and efficiency issues, we need to construct distance metrics on the feature space and appropriately form positive and negative sample pairs.

III. Methodology

In this section, we present the details of our approach. As shown in Fig. 1, we introduce the derivation of our method in the following ways: 1) introducing how to construct the positive and negative sample pairs; 2) constraining the feature representations during the training process using the contrastive loss; 3) developing the contrastive JS loss based on the idea of the contrastive loss to further control the prediction distributions.

A. Positive and Negative Sample Pairs

Data augmentation techniques can improve the generalization performance of the utilized model. For images that have little differences among themselves, the differences between their augmented variants can be even smaller, bringing an additional burden to the model in terms of accurately identifying objects from different classes. To eliminate this issue, the basic idea of our approach is to constrain the interclass features while paying attention to the intraclass features.

To do so, we construct positive and negative sample pairs. Suppose that f_s is the feature extraction subnetwork and f_c is the classifier subnetwork, so the whole model can be defined as $f = f_s \cup f_c$. Analogously, θ_s is the parameter of f_s, and θ_c is the parameter of f_c. Therefore, the whole model parameter set can be defined as $\theta_f = \theta_s \cup \theta_c$. θ_s and θ_c are updated by the backpropagation algorithm and the chain rule. We input x and its label y sampled from dataset D, and $aug(input = x, augmented = True)$ is the augmentation operation function ,where the first variable is the input image, and the second variable is the control signal for indicating whether the input image will be augmented (the

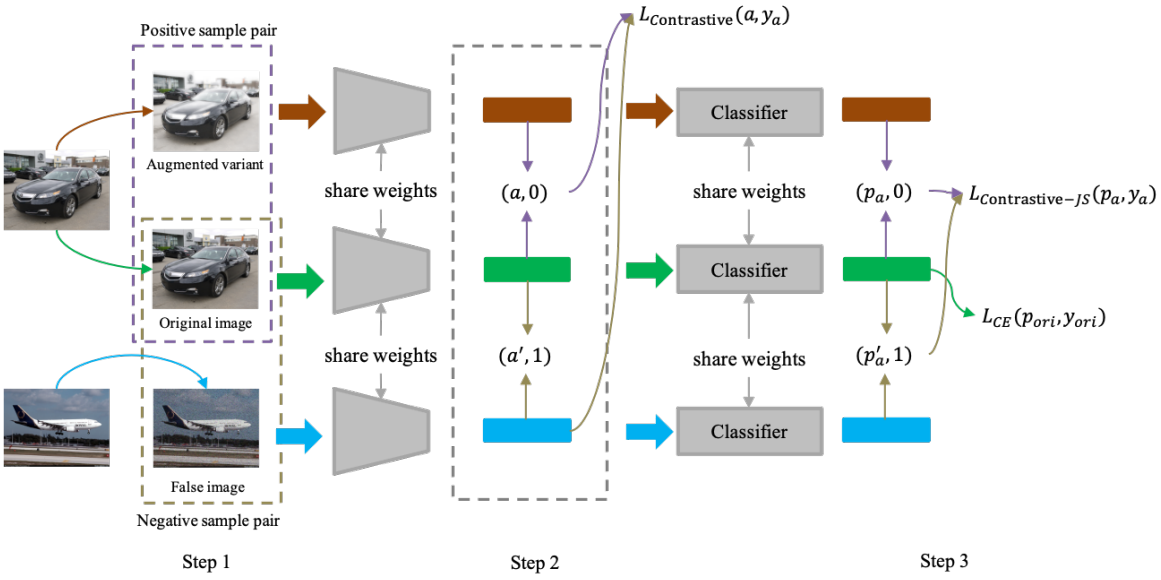


Fig. 1. The process flow of our method. Our method consists of three steps. For step 1, we build positive sample pairs based on a class-sensitive strategy. For step 2, we input these sample pairs into the network simultaneously and construct a feature-consistent learning strategy to guide the feature extraction process. In step 3, we broaden the effect of the contrastive strategy and propose the contrastive JS loss to restrict the prediction distributions of different sample pairs.

default value is True). Therefore, a positive pair can be defined as Equation 1.

$$a_p = (x_{ori}, x_{aug}) \tag{1}$$

where $x_{aug} = aug(x_{ori}, augmented = True)$. Correspondingly, the negative pair is defined as Equation 2.

$$a_n = (x_{ori}, x_{false}) \tag{2}$$

where $x_{false} = aug(x_{false}, augmented = aug\_flag)$. x_{false} is a input image sampled from other categories of D and aug_flag is the random choice from $\{True, Flase\}$.

Different from construction methods [37] that build pairs based on the instance level and potentially define different images from the same category as negative sample pairs, our method focuses on the category level and ensures that x_{false} is sampled from a different category. After building positive and negative pairs, the Siamese network [38] is used to carry out the training process. Apart from the constructed pairs, we keep the original images as clean features and use them to generate prediction distributions as the inputs of the cross-entropy (CE) loss function.

B. Constraint Imposed on Feature Representations

As shown in step 2 in Fig. 1, after constructing the positive and negative pairs, we input these pairs into the Siamese network to fulfill the training process. As we aim to minimize the differences between positive samples and enlarge the differences between negative samples, the contrastive loss [39] is used to control the feature extraction process and is defined as Equation 3.

$$\begin{aligned}\mathcal{L}_{Contrastive}(F_a, y_a) =& \frac{1}{2}\ (1 - y_a)\ D_F \\ &+ \frac{1}{2}\ y_a\ max\{0, m - D_F\}\end{aligned} \tag{3}$$

$$\begin{aligned}D_F(F_a[0], F_a[1]) &= ||F_a[0] - F_a[1]||_2 \\ &= (\sum_{i=0}^{n-1}((F_a[0])_i - (F_a[1])_i)^2)^{\frac{1}{2}}\end{aligned} \tag{4}$$

where F_a is the feature generated by the feature extraction subnetwork f_s. Therefore, $F_a = f_s(a; \theta_s)$. As features are the main objects that we constrain in this section and features can exhibit patterns similar to that of the input image, we use the Euclidean distance measure to quantify the representations of features in sample pairs a (a_p *or* a_n) and D_F is defined as Equation 4. At the same time, during the experiment, it can be found that using the Euclidean distance to calculate the feature consistency does not lead to the gradient explosion phenomenon caused by some overly large calculation results. The model parameters can be stably optimized. $F_a[0]$ and $F_a[1]$ are the features of the inputs in the sample pairs. $y_a \in \{0, 1\}$ is the annotation of the corresponding pair. When a represents the positive pair a_p, y_a equals to 0, and vice versa. m is the margin that decides the upper bound of the pair distances.

C. Contrastive JS Loss

As we construct several sample pairs in a single iteration, to achieve a stronger restraining effect, in addition to constraining the feature learning process using a contrastive loss function, we should also set up a loss function that can provide guidance to constrain the prediction distributions. Inspired by [7], we

propose the contrastive JS consistency loss to constrain the prediction distributions generated by positive and negative pairs, as illustrated in step 3 in Fig. 1. The formula is shown as follows.

$$\begin{aligned}&\mathcal{L}_{Contrastive-JS}(p_a, y_a) = \\ &\frac{1}{2}\ (1-y_a)\ \left(\sum_{i=0}^{b} KL\left[p_a\left[i\right]||\mathcal{M}\right]\right) + \\ &\frac{1}{2}\ y_a\ max\left\{0, m' - \left(\sum_{j=0}^{b} KL\left[p_a\left[j\right]||\mathcal{M}\right]\right)\right\}\end{aligned} \tag{5}$$

where p_a is the prediction distribution of sample group a (a_p *or* a_n), and $y_a \in \{0, 1\}$ is the annotation of this pair. b is the number of samples in the sample group. In this paper, b is set to 1 as there are two samples in a single pair and b counts from 0. p_a is the prediction distribution generated by the classifier subnetwork f_c and $p_a = f_c(F_a; \theta_c)$. $p_a\left[0\right]$ and $p_a\left[1\right]$ are the first and second prediction distribution of input in the pair a respectively. $\mathcal{M} = \frac{1}{2}\left(p_a\left[0\right] + p_a\left[1\right]\right)$ and KL is the Kullback–Leibler divergence. m' is the margin that shares the same meaning as m in Equation 3.

According to this formula, we can clearly determine that when $y_a = 0$, the model updates to minimize the distance between the clean image x_{ori} and its augmented variant x_{aug}. In contrast, when $y_a = 1$ and the distance between x_{ori} the loss value equals 0, which means that the model is not updated; otherwise, the model is updated until the distance between the negative pairs reaches m'. This characteristic satisfies our requirement. First, the proposed scheme saves computing resources and time, as the model is not updated when the distance between negative samples is larger than m'. Second, the proposed method has no additional model parameters, as we use features to diametrically compute the contrastive value without any fully connected layers. Then, based on these characteristics, this loss function can theoretically be placed behind any convolution layer. Various combinations of the locations of the contrastive loss function and the other loss function may achieve different performances, and we study this in the ablation part.

IV. Experiments

A model trained from limited data can generalize well to a wider range of testing data, which is an important measure of the model's stability and efficiency. In this section, to demonstrate the universality of the proposed scheme, we choose model robustness against corruption and FSOD to evaluate the performance of our method. The former application uses clean data as the training set and evaluates the model performance on datasets that are distorted by various corruptions. The latter application utilizes several images, e.g., 1, 2, 3, 5, or 10, as the training set and then tests the performance of the model on the whole testing set.

A. Deep Model Robustness

In this section, we design various experiments to justify our approach, including both comparison experiments with SOTA methods and ablation studies concerning the characteristics of the proposed method.

1) Experimental Settings:

a) Datasets: Our experiments are conducted on CIFAR-10 and CIFAR-100. Both are widely used benchmarks for classification performance evaluations and contain 60,000 (50,000 for training and 10,000 for testing) 32×32×3 color images in 10 and 100 different classes, respectively.

In addition to verifying the standard classification accuracy of the model on the clean testing set, we measure the robustness of the model to data corruptions based on two other datasets, CIFAR-10-C and CIFAR-100-C [6]. These datasets are constructed by corrupting the original CIFAR testing sets, and there are a total of 15 kinds of noise in each dataset.

b) Metrics: For the robustness evaluation, given a corruption type, the corruption classification error is calculated in the same way as the standard classification error. We compute the mean corruption error (mCE) of all corruptions based on Equation 6.

$$\mathbb{E}_{c\sim C}\left[P_{(x,y)\sim D}(f(c(x)) = y)\right] \tag{6}$$

where C is the corruption operation set and f is the classifier. This equation measures the classifier's average-case performance on corruptions C. When C is set to the empty set, this formula represents the standard classification performance on the clean testing dataset.

c) Training Setting: We select several commonly used networks to carry out comparison experiments with some SOTA methods, namely, the All Convolutional Network [41], DenseNet [42], Wide ResNet [43] and ResNeXt [44]. For the ablation studies, we use another highly used network, ResNet [45]. The learning rate is initialized to 0.1 and adjusted according to the cosine annealing schedule. Stochastic gradient descent (SGD) is used as the optimizer, and the total number of training epochs is 100 for all experiments. The batch size is set to 128 for all experiments. All the results are averaged from three independent experiments with the same settings. Furthermore, our experiments are based on PyTorch-1.1.0 and CUDA-10.0.

2) Quantitative Results: We select Augmix [7], which achieves SOTA model robustness [6], as our main comparison method. In addition to Augmix, we also select some other SOTA data augmentation methods as our comparison methods, such as ISDA [18], RandAugment [17], CutMix [15], etc. As introduced in Section III, the proposed method uses clean samples and their augmented variants to form positive pairs. We adopt the augmentation operations in Augmix to generate augmented inputs. We generate two augmented variants and select two false inputs for each experiment, and all the results are averaged over three independent experiments conducted with the same settings.

In the robustness performance evaluation, as shown in Table I, our method outperforms other methods. Specifically, our method achieves 1.3%, 0.5%, 0.3%, and 0.6% improvements over Augmix on AllConvNet, DenseNet, WideResNet, and ResNeXt on the CIFAR-10-C dataset, respectively. Apart from this, on the CIFAR-100-C dataset, our method achieves

TABLE I
MEAN CORRUPTION ERROR (MCE) ON CIFAR-10-C AND CIFAR-100-C DATASETS.

Method	AllConv	DenseNet	WRN	ResNeXt	AllConv	DenseNet	WRN	ResNeXt
Task	CIFAR-10-C				CIFAR-100-C			
Standard	31.6	30.7	27.5	29.1	56.2	58.1	53.1	53.4
Cutout [40]	30.7	32.1	26.8	28.9	56.8	59.6	53.5	54.6
Mixup [14]	24.6	24.6	22.3	22.6	53.4	55.4	50.4	51.4
CutMix [15]	31.3	33.5	27.1	29.5	56.0	59.2	52.9	54.1
AdvTraining [10]	28.1	27.6	26.2	27.0	56.0	55.2	55.1	54.4
AutoAugment [16]	29.2	26.6	23.9	24.2	55.1	53.9	49.6	51.3
RandAugment [17]	22.8	20.1	17.8	18.4	51.2	51.6	46.3	46.0
ISDA [18]	24.2	19.3	14.3	14.5	46.6	50.5	40.8	37.1
Augmix [7]	15.4	12.7	11.0	11.8	42.7	39.3	35.8	34.8
Ours	**14.1**	**12.2**	**10.7**	**11.2**	**39.7**	**38.4**	**35.4**	**34.6**

TABLE II
STANDARD CLASSIFICATION PERFORMANCE ON CIFAR-10 DATASET.

Method	AllConv	DenseNet	WRN	ResNeXt	Mean
Standard	6.3	5.9	5.4	4.8	5.6
Cutout [40]	6.1	4.8	4.4	4.4	4.9
Mixup [14]	6.3	5.5	4.9	4.2	5.2
CutMix [15]	6.4	5.3	4.6	3.9	5.0
AdvTraining [10]	18.9	17.9	17.1	15.4	17.3
AutoAugment [16]	6.6	4.8	4.8	3.8	5.0
RandAugment [17]	6.4	4.6	4.7	4.3	5.0
ISDA [18]	9.8	6.8	5.2	4.2	6.5
Augmix [7]	6.4	5.1	4.8	4.7	5.3
Ours	**5.4**	**5.1**	**4.3**	**4.3**	**4.8**

more significant increases over Augmix, which are 3%, 0.9%, 0.4%, and 0.2% on AllConvNet, DenseNet, WideResNet, and ResNeXt, respectively.

In addition to robustness, we also carry out comparison experiments on the clean dataset. As shown in Table II, we conduct experiments involving the backbones used in the robustness evaluation on the CIFAR-10 dataset. According to the results, we accomplish a 0.5% improvement on average over Augmix. We attain the best classification performance on average compared to other data augmentation methods.

TABLE III
PERFORMANCE EVALUATION BASED ON DIFFERENT SCHEMATA. THE MCE (MEAN CORRUPTION ERROR) VALUE IS COMPUTED BY AVERAGING ACROSS ALL 15 CORRUPTION ERROR VALUES.

Schema	CIFAR-10-C		CIFAR-100-C	
	Clean Error	mCE	Clean Error	mCE
Augmix	4.62	11.40	23.44	35.11
Contrastive-Only	4.51	10.79	22.67	34.09
Contrastive-JS-Only	4.56	10.84	22.69	34.86
Both	**4.43**	**10.34**	**22.43**	**33.56**

3) Ablation Study: In this section, we carry out ablation studies based on the contrastive consistency constraint and the contrastive JS consistency constraint. We select Augmix as our comparison method in the following comparison experiments. We use ResNet [45] as the backbone network. Our ablation studies are conducted according to the following aspects.

a Comparing the effect of each constraint.
b Investigating the best position to apply this contrastive constraint.
c Exploring the relationship between the performance and the depth of the network.
d Analyzing the proportional relationship between positive and negative samples.

- Contrastive VS Contrastive JS

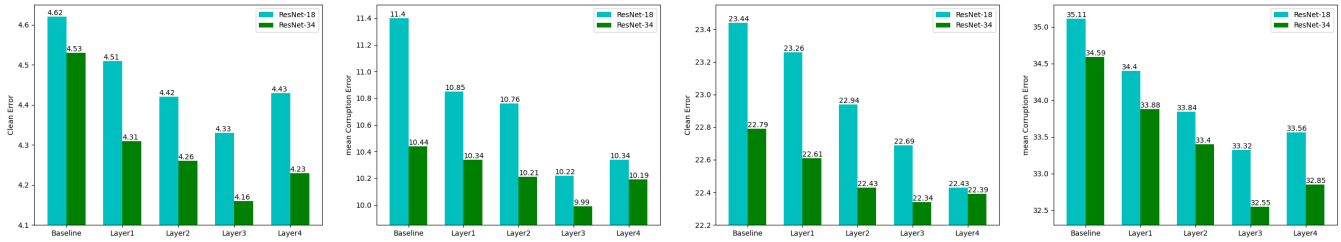
In this part, we evaluate the performances achieved based on different schemes, as shown in Table III. First, we run Augmix using ResNet-18 as the baseline performance. Then, we adopt the augmentation operation of Augmix and use either the contrastive loss or the contrastive JS loss to carry out the constraint separation experiments. Finally, we execute the experiment using the integral method. It is worth noting that the contrastive JS loss handles the probability distributions that are processed by the softmax function and contain the meanings corresponding to the category level. The contrastive loss constrains the feature representations whose meanings correspond to the image level. Therefore, this is different from the position study. Furthermore, to conduct a better comparison with Augmix, for the second scheme, we use not only the contrastive loss to constrain the feature representations but also the JS loss and CE loss to provide guidance for the training process.

The results are shown in Table III. The complete scheme presented in this paper obtains the highest performance among these four cases. Compared with the baseline, our scheme achieves 0.19% and 1.01% clean error improvement and additional 1.06% and 1.55% mean corruption error improvements. Additionally, both the contrastive-only scheme and the contrastive JS-only scheme outperform Augmix, which illustrates that the contrastive strategy is effective in boosting the robustness of the deep model. More specifically, when we put the consistency constraint on the feature level, the difference between the clean object and its variant decreases, and the difference between the clean object and the false object increases. Distinct feature expressions generate distinct prediction distributions, so it is easier for the classifier to distinguish the object. On the other hand, when we put the consistency constraint on the prediction level, the distance behavior of the prediction distribution is similar to the feature level. However, unlike directly constraining the feature consistency, to obtain a more distinguishable prediction distribution, the model generates more inconsistent feature representations during the training process.

TABLE IV
CORRUPTION ERRORS OVER 15 CORRUPTION TYPES OF CIFAR-10-C DATASET BASED ON DIFFERENT SCHEMATA.

Schema	Weather				Digital			
	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Augmix	10.52	10.88	8.37	5.66	9.47	9.14	11.83	11.86
Contrastive-Only	10.20	10.42	**8.15**	**5.54**	9.18	9.25	11.80	11.81
Contrastive-JS-Only	10.10	10.39	8.40	5.59	9.35	9.13	11.28	11.67
Both	**9.90**	**9.83**	8.28	5.57	**8.96**	**8.95**	**11.24**	**11.24**

Schema	Blur				Noise			Mean
	Defocus	Glass	Motion	Zoom	Gauss.	Shot	Impulse	
Augmix	5.69	20.53	7.37	6.61	21.79	15.64	15.70	11.40
Contrastive Only	**5.64**	18.91	7.28	6.60	18.94	14.02	14.06	10.79
Contrastive-JS Only	5.80	19.44	**7.09**	6.71	19.27	14.37	13.97	10.84
Both	5.71	**18.01**	7.15	**6.58**	**18.86**	**13.82**	**11.05**	**10.34**



(a) Clean Error on CIFAR-10 (b) mCE on CIFAR-10-C (c) Clean Error on CIFAR-100 (d) mCE on CIFAR-100-C

Fig. 2. Position study on ResNet-18 and ResNet-34. We employ the contrastive-JS loss behind the classification layer group of ResNets and employ the contrastive loss behind each major feature extraction layer group separately.

TABLE V
PROPORTION STUDY OF POSITIVE AND NEGATIVE PAIRS BASED ON RESNET18.

Dataset	Porprotion	Clean Error	mean Corruption Error
CIFAR-10-C	Augmix	4.62	11.40
	2:1	4.43	10.33
	2:2	**4.33**	**10.22**
	2:4	4.58	10.40
	2:6	4.79	11.02
CIFAR-100-C	Augmix	23.44	35.11
	2:1	22.75	33.82
	2:2	**22.69**	**33.32**
	2:4	22.92	34.02
	2:6	23.07	34.43

Furthermore, the contrastive-only scheme obtains better results than the contrastive JS-only scheme. We envision that the cause of this phenomenon may be the following: the contrastive strategy is effective in improving the robustness of the model. Apart from learning the information within the input image, the model obtains the interclass information by using the constructed image pairs. In this way, the model gains a stronger ability to identify the category information to which the input image belongs. Additionally, as mentioned before, the contrastive JS loss handles the probability information, which contains the predicted probabilities across all classes. For the feature level, the valid information includes the activated features, and the contrastive information is either true or false regardless of the GT value of the input image. In this sense, the contrastive-only scheme is more straightforward than the contrastive JS-only scheme when addressing this issue. Moreover, this performance difference suggests that adding an extra constraint on the feature level is slightly more effective than doing so on the probability level. The last scheme, which is the proposed method, shows that the effects of both constraints can be superimposed so that these constraints focus on different contents.

To obtain an overall comprehension of the robustness performance of the model, we enumerate the corruption errors induced over 15 corruptions of the CIFAR-10-C dataset, as shown in Table IV. According to these results, we find that our method achieves stable improvements on digital and noise corruptions and fluctuates under the other two corruptions. We speculate that the reason for this phenomenon may be the following. The images processed by the digital and noise corruptions become complex, and the differences between adjacent pixels are increased. The differences between these images and their original counterparts increase accordingly. Hence, during the training stage, the contrastive strategy can have a more obvious effect. While the images are affected by the other two corruptions, the boundary between the core object and the background is smoother, which may cause the model to pay too much attention to the wrong details, and the recognition results fluctuate. Moreover, among all kinds of corruption, the proposed method achieves the best improvement on the noise type. This indicates that the proposed method has the potential to be applied to the area of denoising.

- Position Study

As discussed before, the effects of extra constraints can be superimposed. The position where the contrastive loss should be placed is undetermined. In this part, we address this issue

by employing the contrastive loss in different locations of the network. Most of ResNets, such as ResNet-18 and ResNet-34, consist of two main components: a feature extraction layer group and a classification layer group. The feature extraction layer group consists of four major layers, each of which is composed of a different number of basic blocks or bottlenecks. We employ the contrastive JS loss behind the classification layer group, which is the same location as that of the CE loss, and employ the contrastive loss behind each major feature extraction layer group separately and compare the performance gain from each case.

As shown in Fig. 2, every piece of equipment attains a performance gain over Augmix. Specifically, for ResNet-18, the proposed method yields the best standard classification and robustness performance improvements on the CIFAR-10-C dataset when we place the contrastive loss behind the third major layer group. For the CIFAR-100-C dataset, the best performance gain occurs when we place the contrastive loss behind the third and last major layer groups. We speculate that the reason for this phenomenon may be the following: the contrastive loss handles features directly, and the sizes of the feature maps processed after each major layer are dissimilar. As the number of convolutional layers for processing features increases, the size of the feature map gradually decreases, and the information within each image becomes more abstract, which means that the proportion of valid information increases. Therefore, the loss function is more efficient when performing comparative analysis, as it does not focus too much on other invalid information. On the other hand, the best improvement appears in the third layer, indicating that the size of the feature map to be processed by the contrastive loss should not be as small as possible. The information contained in a feature map that is too small may be too abstract to affect the comparison results. In summary, it is suggested to place the contrastive loss behind the high-level convolutional layers and place the contrastive JS loss behind the classification layer.

- Depth Study

The above analysis notes that the degree of feature abstraction and the feature size may affect the performance of the proposed method. Furthermore, networks with different depths contain different numbers of convolutional layers. The deeper the network is, the higher the feature abstraction degree. To explore whether the proposed method has a similar characteristic under a deeper backbone network, we carry out experiments based on ResNet-34.

Similarly, we employ the contrastive JS loss at the end of the classification module and employ the contrastive loss at the end of each major feature extraction layer separately. As shown in Fig. 2, on the whole, the experimental results of ResNet-34 are superior to those of ResNet-18, which demonstrates that ResNet-34 has a stronger feature extraction capability. The more refined feature extraction ability brought about by increasing the network depth helps to stabilize the experimental results. Moreover, in addition to achieving performance increases over Augmix across all constraint combinations, the best standard classification and robustness improvements on CIFAR-10-C and CIFAR-100-C datasets are achieved concomitantly at the third combination. This phenomenon supports our previous analysis. According to this observation, the features from the fourth layer group are slightly over abstracted compared to those from the third layer group for the contrastive loss. Combining the results of previous experiments, it is recommended to put the contrastive loss behind some high-level convolutional layers while leaving some higher layers for the subsequent training process.

- Proportion Study

In this work, we utilize positive and negative sample pairs. In theory, the sampling space of negative sample pairs is much larger than that of positive sample pairs. Models trained with unbalanced data suffer from serious deviations. Therefore, many works [37] have focused on the issue of unbalanced sampling. To determine if our method has such a burden, in this part, we construct experiments concerning the proportions of positive and negative sample pairs. To better compare our approach with the baseline method, for each clean image, we generate two augmented variants to construct positive pairs. We sample one false image for each anchor image from the dataset to build negative pairs as our basic situation. Then, we gradually increase the ratio of negative pairs while maintaining the ratio of positive pairs.

The experimental results are shown in Table V. When we increase the proportion of negative pairs, both the standard classification and robustness performances are improved. As we continue to increase the proportion of negative pairs, these indicators begin to decline. The best performance occurs when the numbers of positive and negative sample pairs are the same. Based on this inspection, the most suitable proportion for our method is to maintain the same number of positive and negative sample pairs. We speculate that the reason for such a phenomenon may be the following: the main concern regarding sampling imbalance is that it makes the model have a very serious bias. However, as we use the data enhancement method to create augmented data, our positive sampling space is enhanced as well. In this sense, the positive sample pairs have sufficient capacity to fulfill the learning requirements of the model.

B. FSOD

In this section, we demonstrate that the proposed scheme is not only suitable for improving model robustness but also available for enhancing data efficiency. More specifically, FSOD helps detectors adapt to unseen classes with few training instances and is useful when manual annotation is time-consuming or when data acquisition is limited. This task is carried out to evaluate the performance of the proposed scheme. We show that the proposed scheme can constantly yield better results than those of some other SOTA methods with no extra costs.

1) Detection Architecture: Inspired by the few-shot image classification task [52], earlier FSOD works mostly utilized meta-learning strategies [53]. Recently, it was revealed that two-stage fine-tuning-based approaches (TFAs) [49] have more potential to achieve improved FSOD performance. The basic idea of a TFA is freezing all base class model parameters

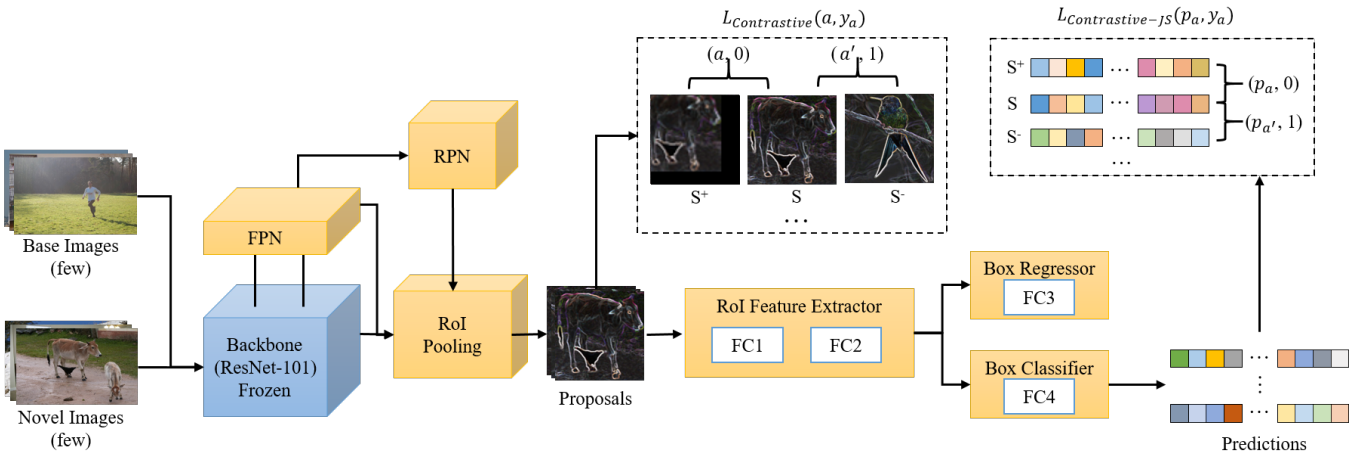


Fig. 3. The process flow of FSOD. The features of the GT object and the same object with the highest IoU score are constructed as positive sample pairs. The features of the GT object and the object from a different category are constructed as negative sample pairs.

TABLE VI
PERFORMANCE EVALUATION (NAP 50) OF EXISTING FEW-SHOT DETECTION METHODS ON THREE PASCAL VOC NOVEL SPLIT SETS. ALL RESULTS ARE AVERAGED BY OVER 5 RANDOM SEEDS.

Method / Shot	Backbone	Novel Split 1					Novel Split 2					Novel Split 3				
		1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
LSTD [46]	VGG-16	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
YOLOv2-ft [47]	YOLO V2	6.6	10.7	12.5	24.8	38.6	12.5	4.2	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
FSRW [48]		14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet [48]		17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
RepMet [24]	InceptionV3	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2
FRCN-ft [48]	FRCN-R101	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
FRCN+FPN-ft [49]		8.2	20.3	29.0	40.1	45.5	13.4	20.6	28.6	32.4	38.8	19.6	20.8	28.7	42.2	42.1
MetaDet [48]		18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [50]		19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA [49]		25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
FSIW [51]		24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
FSCE [8]		32.9	44.0	46.8	52.9	59.7	23.7	30.6	38.4	43.0	48.5	22.6	33.4	39.5	47.3	54.0
Ours		**36.0**	**47.0**	**47.8**	**55.6**	**60.8**	**24.1**	**33.1**	**41.5**	**45.0**	**50.3**	**29.6**	**41.4**	**46.0**	**51.1**	**55.9**

trained using abundant data and only fine-tuning the box classifier and box regressor with novel data. FSCE [8] presents a novel perspective in which proposals with various IoU scores can be regarded as augmented variants of the GT object. Following this, the proposed scheme in this paper is employed to solve this dilemma, and a novel FSOD architecture is presented.

The overall process consists of two stages: a data-abundant base training stage and a novel fine-tuning stage. The first stage is similar to most of TFAs [8], [49]. The process flow of the fine-tuning stage is shown in Fig. 3. The parameters of the backbone network are frozen first. Then, we employ the feature consistency constraint at the end of the region of interest (RoI) pooling operation, where S is the feature corresponding to the GT object, S^+ is the feature of the same object with the highest IoU score and is regarded as the positive sample and S^- is the feature of the object from a different category and is regarded as the negative sample. We utilize such samples to construct the sample pairs and maintain these pairwise relationships after they pass through the subsequent network module. At the end of the classifier, we use the predictions of the sample pairs as the inputs of the contrastive JS constraint to further control the prediction representations of proposals.

2) Experimental Settings:

a) Dataset: Extensive experiments are performed on the PASCAL VOC [54] benchmark. There are 20 categories in this dataset, which are divided into 15 base categories and 5 novel categories. All base category data from the PASCAL VOC 2007+2012 trainval sets are considered available, and K shots of novel instances are randomly sampled from the previously unseen novel classes for K = 1, 2, 3, 5 and 10. Following the existing work [8], we adopt the same three random partitions of the base and novel categories and the samplings introduced in [48], referred to as novel splits 1, 2, and 3. Furthermore, we report the average precision at 50 (AP50) for the novel predictions (nAP50) on the PASCAL VOC 2007 test set as the model performance evaluation indicator. It is worth noting that there might be very large variances between the different training set selections. Our results are consequently averaged over 5 random seeds.

Fig. 4. Visualization of the detection results. The first column represents the original image and its GT bounding boxes, and the remaining columns are the detection results obtained under different shots. The odd rows are the results of FSCE, and the even rows are those of the proposed scheme.

b) Training Setting: For the detection model, we use Faster RCNN [23] with ResNet-101 and a feature pyramid network [55]. All experiments are run on 4 TITAN RTX GPUs with a standard batch size of 16. The solver is standard SGD with a momentum of 0.9 and a weight decay of 1e-4. For the data-abundant base training stage, we use all base category data from PASCAL VOC 2007+2012 as the training dataset. For the novel fine-tuning stage, we use the model parameters trained from the previous stage to initialize the network and fix the backbone. We basically follow the protocol in FSCE [8] and use various training steps and learning rates for different numbers of training shots. We set the margin in Equation 3 to 1 and set the margin in Equation 5 to 0.1 in most experiments. Moreover, every detail will be open-sourced in a self-contained codebase to facilitate future research. Similar to the previous section, our experiments are based on PyTorch-1.1.0 and CUDA-10.0.

3) Quantitative Results: The comparison results obtained for all three random novel splits from PASCAL VOC are shown in Table VI. Our method significantly outperforms all existing works with any number of shots and all splits. Thus, the effectiveness of our method is fully demonstrated. In fact, we are the first to achieve nAP50 results exceeding 60% on novel split 1 and 50% on novel split 2. Moreover, the proposed method obtains 2.18%, 1.96% and 5.24% improvements on average over FSCE [8], which is the current SOTA approach. More specifically, on novel split 1, our method obtains 3.1%, 3.0%, 1.0%, 2.7% and 1.1% improvements under the 1-, 2-, 3-, 5-, and 10-shot cases, respectively. On novel split 2, our method obtains 0.4%, 2.5%, 3.1%, 2.0% and 1.8% improvements and achieves values of 7.0%, 8.0%, 5.5%, 3.8% and 1.9% on novel split 3, respectively.

In addition to comparing the nAP50 performance with that of existing works, we visualize the detection results of our method and FSCE. As shown in Fig.4, the first column represents the original image and its GT bounding boxes, and the remaining columns are the detection results obtained under different shots. The odd rows are the results of FSCE, and the even rows are those of the proposed method. According to the visualization results, as the number of shots increases, the recognition scores of both methods gradually increase. However, when the number of shots is low, FSCE causes duplication detection problems or missed detection problems for some images with complex contexts. Moreover, when the

given image contains multiple objects, the detection stability of our method is significantly better than that of FSCE. For example, when the number of shots equals 5, FSCE successfully detects the person in the second testing sample, but it misses this person as the number of shots increases to 10. Not only does our method precisely detect this person, but as the number of shots increases, so does his confidence score provided by the model. During the testing process, a large number of detection areas are generated, these areas are filtered by an IoU threshold, and the remaining regions are given a classification score to produce the final detection result. According to this, the appearance of the above detection problems means that the detector does not rectify the deviated regions effectively during the training process. Furthermore, when the number of images used for training is low, the feature identification capability of the model is limited. A model trained under such conditions has difficulty producing precise activation responses for all objects in an image. While our method can fully improve the data utilization efficiency of the model during the training process, our method outperforms FSCE. According to this, it can be discerned that the model trained by our scheme is more effective in making full use of the scant training data. Moreover, our method does not introduce extra parameters. In fact, from observation, the model size of FSCE is 235.66 MB, while our model is even smaller at only 231.07 MB. This also illustrates that our method enhances the capacity and effectiveness of the model.

TABLE VII
PERFORMANCE EVALUATION (BAP 50) ON PASCAL VOC BASE SPLIT 1.

Method / Shot	1	2	3	5	10
Baseline-FPN [20]	56.9	-	66.2	67.9	-
FRCN-ft [48]	68.9	69.4	66.1	66.7	66.0
FSCE [8]	**78.9**	74.3	74.1	76.6	77.5
Ours	77.1	**74.7**	**75.5**	**76.8**	**78.5**

Moreover, as we adopt the two-stage training strategy to carry out the FSOD experiments, in addition to the novel split data, we evaluate the performance of our model on the base split data to verify the ability of our method to maintain the properties of base datasets. As shown in Table VII, our method achieves improvements over FSCE, which shows that our scheme enables the detector to retain the lower base forgetting property, as in FSCE.

V. LIMITATIONS AND FUTURE WORK

With the advancement of technology, various forms of data have endlessly emerged, among which vision-based data are representative. Deep learning techniques have been successfully used to solve various practical computer vision problems, such as video segmentation. However, networks tend to memorize the properties of certain data distributions [4]; hence, these models are vulnerable when the testing set is more complicated than the training set. To address this issue, we investigate a training scheme for deep visual recognition models with stronger robustness and effectiveness by utilizing contrastive learning to enhance the consistency of the responses for the same object in different scenarios.

Although most of our experimental results indicate that the proposed method can continuously improve its performance across multiple backbone networks and vision tasks, we also observe a limitation that might be associated with the margin in the consistency constraints. Margins are related to datasets and tasks. Currently, the setting of the margin mainly relies on experimental experience. Sometimes, finding a proper margin is time-consuming. Thus, it becomes necessary to design an end-to-end trainable framework that allows the network to determine the value of the margin during the training process.

To the best of our knowledge, this is the first study that integrates contrastive learning to focus on the robustness and effectiveness of deep neural networks. Our method can boost the robustness of the model and can be applied to other multimedia-related works. We believe that further research in this direction is valuable. For example, our method can be applied to some more specific applications, such as multivehicle detection [56] and cross-dataset action recognition [57]. These complex decision systems typically perform multistage analyses of multimedia data collected by sensors using different modules, and their testing scenarios are usually much more intricate than the training scenarios. Our method has the potential to be applied to these applications, as it can facilitate the activation responses of features in the network.

VI. CONCLUSIONS

[58]

In this paper, we propose a training scheme to boost the robustness and effectiveness of deep models by adopting the contrastive learning technique. Specifically, we construct positive and negative sample pairs using clean images with either their augmented variants or false images sampled from other classes. Then, we simultaneously input these pairs into the network and use the contrastive loss to guide the feature extraction process. Following this, we propose a contrastive JS loss to restrict the probability distributions. According to the experimental results, our method outperforms some typical SOTA methods in the domains of both model robustness and FSOD. Moreover, based on our ablation studies, to achieve better performance, it is suggested to place the contrastive loss behind the high-level convolutional layers and to place the contrastive JS loss behind the classification layer.

REFERENCES

[1] S. Wan, Y. Xia, L. Qi, Y.-H. Yang, and M. Atiquzzaman, "Automated colorization of a grayscale image with seed points propagation," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1756–1768, 2020.

[2] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning deep multi-level similarity for thermal infrared object tracking," *IEEE Transactions on Multimedia*, vol. 23, pp. 2114–2126, 2020.

[3] Y. Zhang, F. Zhang, Y. Jin, Y. Cen, V. Voronin, and S. Wan, "Local correlation ensemble with gcn based on attention features for cross-domain person re-id," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 19, no. 2, pp. 1–22, 2022.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014, pp. 1–10.

[5] Z. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," in *International conference on machine learning*. PMLR, 2018, pp. 3122–3130.

[6] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*. OpenReview.net, 2019, pp. 1–16.

[7] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *International Conference on Learning Representations*. OpenReview.net, 2020, pp. 1–15.

[8] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7352–7362.

[9] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *Journal of Machine Learning Research*, vol. 20, no. 184, pp. 1–25, 2019.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–11.

[11] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *International Conference on Learning Representations*. OpenReview.net, 2018, pp. 1–22.

[12] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," *arXiv preprint arXiv:1611.05760*, 2016.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, 2017, pp. 6402–6413.

[14] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*. OpenReview.net, 2018, pp. 1–13.

[15] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6023–6032.

[16] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1–14.

[17] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.

[18] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[19] Z. Chen, Y. Fu, Y.-X. Wang, L. Ma, W. Liu, and M. Hebert, "Image deformation meta-networks for one-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8680–8689.

[20] J. Wu, S. Liu, D. Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *European conference on computer vision*. Springer, 2020, pp. 456–472.

[21] H. Chen, Y. Wang, G. Wang, and Y. Qiao, "Lstd: A low-shot transfer detector for object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[24] L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, and A. M. Bronstein, "Repmet: Representative-based metric learning for classification and few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5197–5206.

[25] Z. Yang, Y. Wang, X. Chen, J. Liu, and Y. Qiao, "Context-transformer: tackling object confusion for few-shot detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 653–12 660.

[26] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4013–4022.

[27] L. Karlinsky, J. Shtok, A. Alfassy, M. Lichtenstein, S. Harary, E. Schwartz, S. Doveh, P. Sattigeri, R. Feris, A. M. Bronstein, and R. Giryes, "Starnet: towards weakly supervised few-shot object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2021, pp. 1743–1753.

[28] C. Zhu, F. Chen, U. Ahmed, Z. Shen, and M. Savvides, "Semantic relation reasoning for shot-stable few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8782–8791.

[29] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[30] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[31] D. Nguyen, D. T. Nguyen, R. Zeng, T. T. Nguyen, S. N. Tran, T. Nguyen, S. Sridharan, and C. Fookes, "Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1313–1324, 2021.

[32] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *NeurIPS*, 2019, pp. 5754–5764.

[33] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*. OpenReview.net, 2019, pp. 1–24.

[34] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[37] M. Patacchiola and A. J. Storkey, "Self-supervised relational reasoning for representation learning," in *NeurIPS*, 2020.

[38] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 773–782.

[39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[40] C. Termritthikun, Y. Jamtsho, and P. Muneesawang, "An improved residual network model for image recognition using a combination of snapshot ensembles and the cutout technique," *Multimedia Tools and Applications*, vol. 79, pp. 1475–1495, 2020.

[41] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–15.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[43] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference 2016*, R. C. Wilson, E. R. Hancock, and W. A. P. Smith, Eds. BMVA Press, 2016, pp. 1–15.

[44] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] H. Chen, Y. Wang, G. Wang, and Y. Qiao, "Lstd: A low-shot transfer detector for object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[47] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9925–9934.

[48] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8420–8429.

[49] X. Wang, T. E. Huang, J. Gonzalez, T. Darrell, and F. Yu, "Frustratingly simple few-shot object detection," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, vol. 119. PMLR, 2020, pp. 9919–9928.

[50] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9577–9586.

[51] Y. Xiao and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," in *European conference on computer vision*. Springer, 2020, pp. 192–210.

[52] M. Goldblum, S. Reich, L. Fowl, R. Ni, V. Cherepanova, and T. Goldstein, "Unraveling meta-learning: Understanding feature representations for few-shot tasks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3607–3616.

[53] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9577–9586.

[54] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[55] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

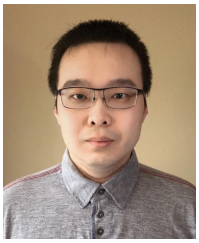
[56] D. Roy, Y. Li, T. Jian, P. Tian, K. R. Chowdhury, and S. Ioannidis, "Multi-modality sensing and data fusion for multi-vehicle detection," *IEEE Transactions on Multimedia*, 2022.

[57] T. Perrett and D. Damen, "Ddlstm: dual-domain lstm for cross-dataset action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7852–7861.

[58] K. Yang, J. H. Yau, L. Fei-Fei, J. Deng, and O. Russakovsky, "A study of face obfuscation in imagenet," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 313–25 330.

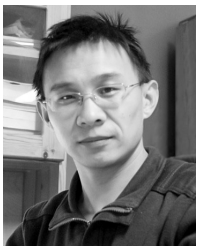
Weiwei Xing received her B.S. degree in Computer Science and Technology and Ph.D. degree in Signal and Information Processing from Beijing Jiaotong University, in 2001 and 2006 respectively. During 2011 to 2012, she was a visiting scholar at University of Pennsylvania. Currently, she is a professor at School of Software Engineering, Beijing Jiaotong University. Her research interests mainly include intelligent information processing and artificial intelligence.

Jie Yao received the B.S. and Ph.D. degree in Software Engineering from Beijing Jiaotong University, in 2016 and 2022 respectively. During 2019 to 2020, he was a visiting student at University of Central Florida. Currently, he is a faculty member in School of Information Management, Beijing Information Science and Technology University. His research interests include image processing, computer vision and deep model robustness.

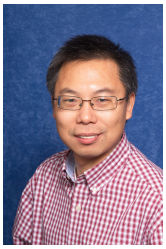


Zixia Liu received the Ph.D. degree in computer science from the University of Central Florida, USA in 2022. He was formerly a student of mathematics and accordingly received his B.S., M.S. degrees from Jilin University, China in 2007, 2009 and M.A. degree from University of Kansas, USA in 2012. His research interests include machine learning and distributed big data computing. He is interested in analyzing machine learning techniques as well as their potentials for improving distributed big data computing performance.

Weibin Liu received the Ph.D. degree in Signal and Information Processing from Institute of Information Science at Beijing Jiaotong University, China, in 2001. During 2001-2005, he was a researcher in Information Technology Division at Fujitsu Research and Development Center Co., LTD. Since 2005, he has been with the Institute of Information Science at Beijing Jiaotong University, where currently he is a professor in Digital Media Research Group. He was also a visiting researcher in Center for Human Modeling and Simulation at University of Pennsylvania, PA, USA during 2009-2010. His research interests include computer vision, computer graphics, image processing, deep learning, and pattern recognition.



Shunli Zhang received the B.S. and M.S. degrees in electronics and information engineering from Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree in signal and information processing from Tsinghua University in 2016. He was a visiting scholar in Carnegie Mellon University, Pittsburgh, from 2018 to 2019. He is currently a faculty member in School of Software Engineering, Beijing Jiaotong University. His research interests include pattern recognition, computer vision, and image processing.



Liqiang Wang is a professor in the Department of Computer Science at the University of Central Florida. He is the director of Big Data Lab. He was a faculty member (2006-2015) in the Department of Computer Science at the University of Wyoming. He received Ph.D. in Computer Science from Stony Brook University in 2006. He was a visiting Research Scientist in IBM T.J. Watson Research Center during 2012 to 2013. His research focuses on big data techniques, which include the following aspects: (1) improving the accuracy, security, privacy, and fairness of big data analytics; (2) optimizing performance, scalability and resilience of big data processing, especially on Cloud and GPU platforms; (3) using program analysis and deep learning techniques to detect and avoid programming errors, execution anomaly, as well as performance defects in big data programs. He received NSF CAREER Award in 2011 and Castagne Faculty Fellowship (2013-2015).