Towards Learning Spatially Discriminative Feature Representations

Chaofei Wang*, Jiayu Xiao*, Yizeng Han, Qisen Yang, Shiji Song, Gao Huang†
Department of Automation, Tsinghua University

Abstract

The backbone of traditional CNN classifier is generally considered as a feature extractor, followed by a linear layer which performs the classification. We propose a novel loss function, termed as CAM-loss, to constrain the embedded feature maps with the class activation maps (CAMs) which indicate the spatially discriminative regions of an image for particular categories. CAM-loss drives the backbone to express the features of target category and suppress the features of non-target categories or background, so as to obtain more discriminative feature representations. It can be simply applied in any CNN architecture with neglectable additional parameters and calculations. Experimental results show that CAM-loss is applicable to a variety of network structures and can be combined with mainstream regularization methods to improve the performance of image classification. The strong generalization ability of CAM-loss is validated in the transfer learning and few shot learning tasks. Based on CAM-loss, we also propose a novel CAAM-CAM matching knowledge distillation method. This method directly uses the CAM generated by the teacher network to supervise the CAAM generated by the student network, which effectively improves the accuracy and convergence rate of the student network.

1. Introduction

In the past few years, convolutional neural networks (CNNs) have achieved excellent performance in many visual classification tasks. To handle the increasingly complex data, CNNs have continuously been improved with deeper structures (AlexNet [22], VGGNet [32], ResNet [14], ResNext [46], DenseNet [18]). However, deep networks are prone to overfitting while they get stronger learning ability. Many researchers have proposed effective regularization solutions, such as Dropout [33], Weight Decay [10], Stochastic Depth [19], Mixup [54], Shakedrop [47], Cutmix [51]. An alternative solution is to design different loss functions to obtain more distinguishing feature representations, which increase intra-class compactness and inter-class separability. Inspired by such idea, contrastive loss [12], triplet loss [30], center loss [45] are proposed to bring in additional constraints on the basis of cross entropy loss. Unfortunately, they usually dramatically increase the computational cost. L-Softmax [26] and SM-Softmax [25] are proposed to modify the original softmax function mathematically, leading to potentially larger angular separability between feature vectors. Implicit semantic data augmentation (ISDA) loss [41] is proposed to optimize the upper bound of expected cross-entropy loss. However, when adopting the above loss functions, an input image is represented by a one-dimensional feature vector, which collapses the spatial information.

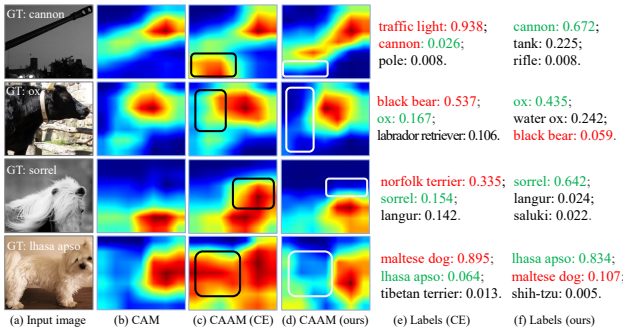

Figure 1. Some examples to illustrate our motivation. A ResNet-50 model trained on ImageNet is adopted. "GT" represents the ground truth label. "CE" represents cross entropy loss. "ours" represents CAM-loss. Black bounding boxes show main differences between (b) and (c), while white ones show main differences between (c) and (d).

In this paper, we propose to construct a novel loss function by leveraging the class activation maps (CAMs [56]) with rich spatial information. CAM indicates the spatial discriminative regions to identify a particular category. It is easily obtained by computing a weighted sum of the feature maps of the last convolutional layer. In fact, we can also obtain a class-agnostic activation map (CAAM [2]) by computing the sum of the feature maps directly, which indicates the spatial distribution of the embedded features. To describe our motivation visually, in Figure 1 we show some validation images misclassified by a pre-trained ResNet-

*Equal contribution.
†Corresponding author.

50 [14] model with cross entropy loss on ImageNet [29]. The CAMs of target categories, CAAMs and output labels are shown in columns (b), (c), and (e), respectively. A visual conclusion is that CAAMs generally show larger activated areas and richer features than CAMs of target categories. Unfortunately, the redundant feature representations (black bounding box areas in column (c)) result in the confidence scores of non-target categories (red labels in column (e)) exceeding those of target categories (green labels in column (e)), which lead to misclassification. For example, the expression of body makes the model misidentify an ox as a black bear, and the expression of ears and mane makes the model misidentify a horse as a dog. Intuitively, if we constrain CAAMs closer to CAMs of target categories, features of the target categories will be expressed well and those of non-target categories will be suppressed simultaneously. This effectively enforces intra-class compactness and inter-class separability.

Based on the above inspiration, we construct a new loss function, termed as CAM-loss, by minimizing the distance between the CAAM and the CAM of target category for each training image. CAM-loss drives the backbone to learn more discriminative feature representations from a spatial perspective. We train another ResNet-50 model with CAM-loss, and show CAAMs and output labels of the same images in Figure 1 (d) and (f). It shows that CAAMs produced by CAM-loss are usually cleaner than those produced by cross entropy loss (comparing column (c) with column (d)). Some features of non-target categories are suppressed (white bounding box areas in column (d)), which greatly improves the accuracy of labels (comparing column (e) with column (f)). In fact, extensive experiments show that CAM-loss effectively improves the performances of various classification models.

As an independent loss module, CAM-loss can also be combined with the mainstream regularization methods to improve their performances. Furthermore, we verify the strong generalization ability of CAM-loss in transfer learning and few shot learning tasks. CAM-loss particularly boosts the baseline method by 7.04% (1-shot) and 4.75% (5-shot) on CUB [38], 2.78% (1-shot) and 1.68% (5-shot) on Mini-ImageNet [37] in the setting of few shot learning. This is attributed to the key role of CAM-loss in reducing the negative effect of image background.

In the traditional teacher-student knowledge distillation framework, the existing methods all use a certain type of teacher knowledge to supervise the same type of student knowledge, such as soft target [15], hints [28], attention map [52], relationship between samples [27] or layers [49]. Inspired by CAM-loss, we propose a different idea to match different types of knowledge between teacher and student, that is, to directly supervise the CAAMs generated by the student network with the CAMs generated by the teacher network. We call it CAAM-CAM matching (CCM) knowledge distillation. The experimental results show that CCM can effectively improve the accuracy and convergence rate of the student network.

The main contributions of our work are:

- We propose a novel loss function CAM-loss from the perspective of spatial information. It can effectively improve the classification performance of various CNN models with neglectable additional parameters and calculations, and easily be combined with the mainstream regularization methods to achieve the state-of-the-art on CIFAR-100 and ImageNet.
- CAM-loss shows strong generalization capability in transfer learning and few shot learning tasks. In particular, CAM-loss significantly improves the performance of few shot image classification.
- We propose a novel knowledge distillation method named CAAM-CAM matching, which matches different types of knowledge between teacher (CAMs) and student (CAAMs), and improves the accuracy and convergence rate of the student network simultaneously.

2. Related work

2.1. Class Activation Map

Generating class activation maps (CAMs [56]) with CNN classification models plays an important role in computer vision. Grad-CAM [31] and Grad-CAM++ [3] generalize CAM [56], so that CAMs can be obtained in any CNN-based classification models. The CAM technique derived from the classification network has been widely used for other weakly supervised visual tasks, such as localization [2, 48], detection [39, 43, 55], segmentation [1, 23, 44].

In image classification tasks, CAM is usually used as a visualization technique, but few researchers treat it as something that can be fed back into training [8, 11, 34]. [8] introduced a complicated multi-branch structure consisting of an attention mechanism, an attention branch (based on CAM), and a perception branch. [11] used dual image input and two-branch structure to do the attention consistency (based on CAM). These two methods rely on complex network structures and result in additional computation cost. In contrast, CAM-loss directly introduces the distance constraint between CAAM and CAM, both of which are generated in the normal training process of the classification model. Our method is very clean, direct, low-cost but effective. [34] proposed to suppress the features of negative categories by minimizing the CAMs of top-k negative categories or constraining them with a uniform spatial distribution. Compared with [34], CAM-loss suppresses more extensive non-target regions such as the background. Furthermore, because the CAM of the target category and those of the non-target categories may overlap in some regions,

CAM-loss can better avoid the risk of simultaneously suppressing the features of the target category.

2.2. Loss function

The cross entropy loss is widely used in CNNs due to its simplicity and probabilistic interpretation. Despite its popularity, it does not explicitly encourage the intra-class compactness and inter-class separability. One solution route is to add an additional loss term to assist the cross entropy loss. The contrastive loss [12] was proposed to simultaneously minimize the distances between positive image pairs and enlarge the distances between negative image pairs. Similarly, the triplet loss [30] was proposed to apply a similar strategy to image triplets rather than image pairs. The center loss [45] was proposed to minimize the euclidean distance between the feature vector and the corresponding class centroid. A major drawback of these losses is the expensive calculation on image pairs or triplets explosion, class centroids update. Another solution route is to modify the softmax cross entropy loss. L-Softmax [26], SM-Softmax [25] and AM-Softmax [40] were proposed to introduce some margin parameters into the softmax function. ISDA [41] was proposed to optimize the upper bound of expected cross-entropy loss. However, in these methods, images are represented by one-dimensional feature vectors, which do not include any spatial information.

Different from the previous methods, CAM-loss utilizes the spatial information of self-generated CAMs to constrain the feature maps, which drives the backbone of CNN to learn more spatially discriminative feature representations. It has notable visual interpretability, and requires little additional computation. These advantages make it suitable for a wide range of application scenarios.

2.3. Knowledge distillation

A vanilla knowledge distillation (KD [15]) proposed to transfer some knowledge of a strong capacity teacher model to a compact student model by minimizing the Kullback-Leibler divergence between the soft targets of two models. Since then, there have been works exploring variants of knowledge distillation. Fitnets [28] proposed to transfer the knowledge using not only final outputs but also intermediate ones. AT [52] proposed an attention-based method to match the activation-based and gradient-based spatial attention maps. FSP [49] proposed to compute the Gram matrix of features across layers for knowledge transfer. CCKD [27] proposed to transfer the correlation between instances. Existing methods all use a certain type of teacher knowledge to supervise the same type of student knowledge. Different from them, we first propose a new idea to match different types of knowledge between teacher and student, which can effectively improve the accuracy and convergence rate of the student network simultaneously.

3. Method

In this section, we first formally define and describe the proposed CAM-loss, then analyze the choice of the hyper parameters, finally introduce and explain CAAM-CAM matching knowledge distillation.

3.1. Definition of CAM-loss

Based on the procedure of generating CAMs in [56], we present how to get the CAM, CAAM, and CAM-loss in a CNN-based architecture as shown in Figure 2. Note that we can also use the generalized methods Grad-CAM [31] or Grad-CAM++ [3] to replace the method of CAM [56]. The only difference is that, due to the calculation of gradients, Grad-CAM [31] or Grad-CAM++ [3] will increase the computational cost. In the paper, the method of CAM [56] is chosen for the convenience of description and reduction of experiment cost. The formal description is shown below.

For a given image, the last convolutional layer outputs some units of feature map. Let $f_k(x, y)$ represents the activation of unit k at spatial location (x, y). Then, for unit k of height H and width W, the result of performing global average pooling, $F_k = \frac{1}{H \times W} \sum_{x,y} f_k(x, y)$. Thus, for a given class i, the input to the softmax, $z_i = \sum_k w_k^i F_k$, where w_k^i is the weight corresponding to class i for unit k. Essentially, w_k^i indicates the importance of F_k for class i. Finally the output of the softmax for class i, p_i is given by $\frac{e^{(z_i)}}{\sum_j e^{(z_j)}}$. By plugging $F_k = \frac{1}{H \times W} \sum_{x,y} f_k(x, y)$ into the class score z_i, we obtain

$$\begin{aligned} z_i &= \frac{1}{H \times W} \sum_k w_k^i \sum_{x,y} f_k(x, y) \\ &= \frac{1}{H \times W} \sum_{x,y} \sum_k w_k^i f_k(x, y). \end{aligned} \tag{1}$$

We define CAM_i as the class activation map for class i, where each spatial element is given by

$$\text{CAM}_i(x, y) = \sum_k w_k^i f_k(x, y). \tag{2}$$

Thus, $z_i = \frac{1}{H \times W} \sum_{x,y} \text{CAM}_i(x, y)$, where $\text{CAM}_i(x, y)$ directly indicates the importance of the activation at spatial location (x, y) leading to the image belonging to class i.

Furthermore, we define CAAM as the class-agnostic activation map for the input image. Each spatial element of CAAM is given by

$$\text{CAAM}(x, y) = \sum_k f_k(x, y). \tag{3}$$

To drive CAAM close to CAM_i, we define L_{cam} to measure the distance between CAAM and CAM_i. After the same min-max normalization of CAAM and CAM_i, we get

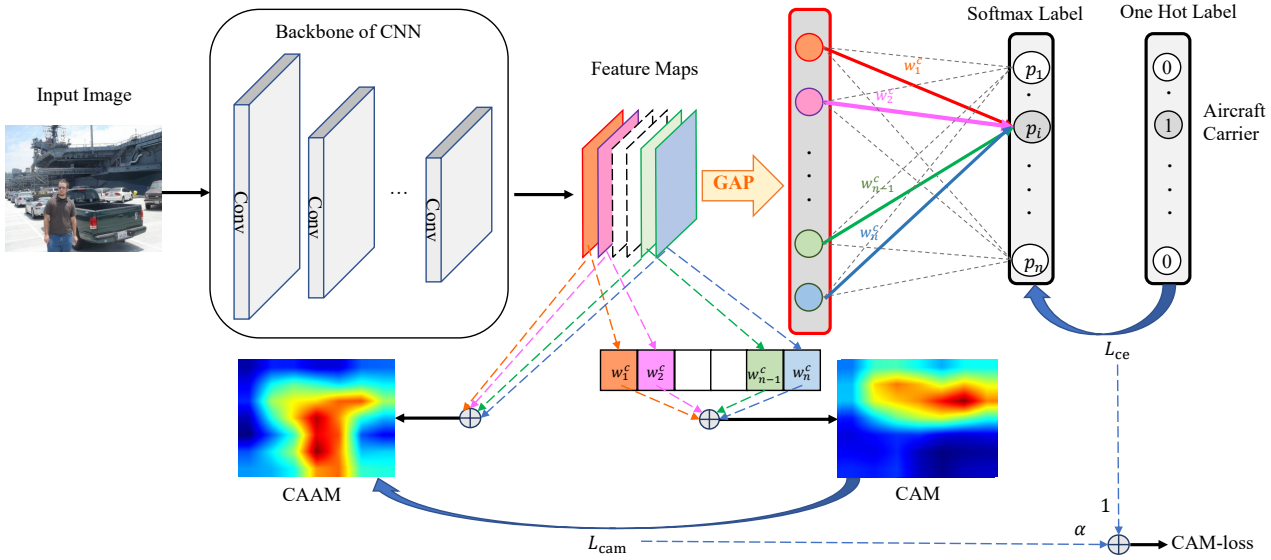

Figure 2. How to get the CAM, CAAM and CAM-loss in a CNN framework. CAM is a weighted sum of the feature maps of the last convolutional layer. CAAM is the sum of the feature maps directly. CAM-loss is the combination of L_{cam} and L_{ce}

$\text{CAAM}^{'}$ and $\text{CAM}_i^{'}$, and then use any pixel space distance to measure L_{cam}. In this paper, we simply choose l_1 distance (l_2 could be used as well). So, the formal expression of L_{cam} is given by

$$L_{\text{cam}} = \frac{1}{H \times W} \sum_{x,y} \left\| \text{CAAM}^{'}(x,y) - \text{CAM}_i^{'}(x,y) \right\|_{l_1}. \quad (4)$$

Of course, the cross entropy (CE) loss L_{ce} is still necessary and defined as follow:

$$L_{\text{ce}} = -\log \frac{e^{(z_i)}}{\sum_j e^{(z_j)}}. \quad (5)$$

When updating parameters of backbone, the two loss terms should be well combined as follow

$$\text{CAM-loss} = \alpha L_{\text{cam}} + L_{\text{ce}}, \quad (6)$$

where α represents the combine ratio.

The training process is summarized in algorithm 1. Note that L_{ce} is used to update W while CAM-loss is used to update θ. The purpose is to eliminate the correlation between W and L_{cam}, which may cause W to approach a all one vector, resulting in an illusory decline of CAM-loss.

3.2. Choice of α

How to choose α is an open question. Intuitively, CAMs obtained in the headmost epochs are too discrete to guide the CAAMs, while CAMs obtained in the latter epochs are more effective for guiding. So, we consider α as a simple step function formally described as follows

$$\alpha = \begin{cases} 0, x < t \\ c, x \geqslant t \end{cases}, \quad (7)$$

Algorithm 1 Training process with CAM-loss

Initialization: the parameters of backbone θ; the parameters of the following full connection layer W;

Optimization:

1: **for** number of training iterations **do**
2: Calculate CAAM and CAM;
3: Update W with $\nabla_W L_{\text{ce}}$;
4: Update θ with ∇_θCAM-loss
5: **return** optimal parameters θ^* and W^*

where t is the jump point (or start epoch). It means that L_{cam} will be added to L_{ce} from the t^{th} epoch. We simply set $c = 1$ to analyze the relationship between the value of t and the error rate as shown in the left part of Figure 3. It shows that the best t is 30. Further analysis, we find that the train error and test error are approximately less than 50% at the 30^{th} epoch. At the moment, CAMs already have obvious target category features. In this sense, adaptively setting the value of t as the epoch when training accuracy exceeds 50% is a simple but smart choice.

With the position of the start epoch t fixed, we analyze the effect of the size of c on the error rate as shown in the right part of Figure 3. It shows that the improvement of error rate is maximal when $c = 3$. In fact, the choice of the optimal α is related to the dataset and the number of training epoch. α can also be any other function or a kind of probability distribution. It is difficult to traverse all possibilities, but we can always beat the baseline with a simple selection strategy as shown in Figure 3.

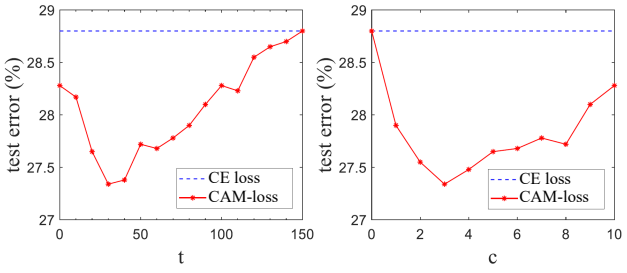

Figure 3. Ablation of t (left) and c (right). We conduct an image classification experiment on CIFAR-100 with ResNet-56. The baseline adopts the cross entropy loss while our method adopts CAM-loss. The experimental setting can be seen in Sec. 4.2

3.3. CAAM-CAM matching

The classical knowledge distillation (KD [15]) lets a weak student mimic a strong teacher's behavior by minimizing the Kullback-Leibler divergence of their soft targets. The formal description is shown below.

Given a vector of logits z as the output of a deep model (or the input to the softmax), such that z_i is the logit for the class i, and then the probability p_i that the input image belongs to the class i can be estimated by a softmax function, $p_i = \frac{e^{(z_i)}}{\sum_j e^{(z_j)}}$. A temperature factor τ is introduced to control the importance of each soft target as $p_i^\tau = \frac{e^{(z_i/\tau)}}{\sum_j e^{(z_j/\tau)}}$, where a higher temperature produces a softer probability distribution over classes. The distillation loss term of KD [15] is

$$L_{\text{kd}} = \frac{1}{n}\sum_{i=1}^{n}\tau^2(p_{ti}^\tau \log p_{ti}^\tau - p_{ti}^\tau \log p_{si}^\tau), \tag{8}$$

where p_{si}^τ and p_{ti}^τ represent the soft targets p_i^τ of the student and teacher respectively. n is the number of classes.

AT [52] proposes to match the attention maps between two different models. Using only the CAMs generated by the last convolutional layer, the distillation loss term of AT [52] can be simplified as follows

$$L_{\text{at}} = \left\| \text{CAM}_{si}^{'} - \text{CAM}_{ti}^{'} \right\|_{l_1}, \tag{9}$$

where $\text{CAM}_{si}^{'}$ and $\text{CAM}_{ti}^{'}$ represent the normalized CAM of the student and teacher to the target class i respectively. l_1 distance is used instead of l_2 to keep consistent.

Different from AT [52], CAAM-CAM matching (CCM) adopts the normalized CAM of the target category generated by the teacher to constrain the normalized CAAM generated by the student. The CCM distillation loss term is

$$L_{\text{ccm}} = \left\| \text{CAAM}_{s}^{'} - \text{CAM}_{ti}^{'} \right\|_{l_1}, \tag{10}$$

where $\text{CAAM}_{s}^{'}$ represents the normalized CAAM of the student. It's easy to find the difference between Eq. (9) and Eq. (10) is that $\text{CAAM}_{s}^{'}$ replaces $\text{CAM}_{si}^{'}$. Further, Eq. (10) can be transformed as follows

$$\begin{aligned} L_{\text{ccm}} &= \left\| \text{CAAM}_{s}^{'} - \text{CAM}_{ti}^{'} \right\|_{l_1} \\ &\leqslant \left\| \text{CAAM}_{s}^{'} - \text{CAM}_{si}^{'} \right\|_{l_1} + \left\| \text{CAM}_{si}^{'} - \text{CAM}_{ti}^{'} \right\|_{l_1} \\ &= L_{\text{cam}} + L_{\text{at}}. \end{aligned} \tag{11}$$

Eq. (11) states that $L_{\text{cam}} + L_{\text{at}}$ is the upper bound of L_{ccm}. In other words, CCM can be approximated as adding CAM-loss to the student network on the basis of AT [52]. Because CAM-loss can independently improve the performance of the student network, it is reasonable to infer that CCM can obtain better performance than AT [52].

In practical application, like AT [52], CCM also needs to combine the cross entropy loss term and soft target loss term to achieve the best performance. The loss function is

$$L = \beta L_{\text{ce}} + (1-\beta)L_{\text{kd}} + \gamma L_{\text{ccm}}, \tag{12}$$

where β and γ represent the combine ratio. Because the teacher is a well-trained network that generates good CAMs, the optimization of β and γ can be done directly by linear search, without considering the influence of epoch like α in Sec.3.2.

4. Experiments

In this section, we first introduce the datasets in our experiments (Sec. 4.1). Then we evaluate the performance of CAM-loss in image classification tasks (Sec. 4.2), including application to various networks, combination with mainstream regularization methods, and comparison with different loss functions. We also verify the generalization ability of CAM-loss in transfer learning and few shot learning tasks (Sec. 4.3, 4.4). Finally, we apply CAAM-CAM matching method to knowledge distillation tasks (Sec. 4.5).

4.1. Datasets

CIFAR-10 and CIFAR-100 [21]. The CIFAR-10 and CIFAR-100 comprise 32×32 pixel RGB images with 10 and 100 classes, containing 50,000 training and 10,000 test images. We follow the standard augmentation in [17]. That is, the training images are padded 4 pixels, and then randomly cropped to 32×32 combined with random horizontal flipping. The original 32×32 images are used for testing.

ImageNet-1K [5] and Mini-ImageNet [37]. The ImageNet-1K contains 1.2 million training and 50,000 validation images of 1000 classes. The Mini-ImageNet consists of a subset of 100 classes from the ImageNet and contains 600 images for each class. We adopt the same augmentation strategy as [51] and apply a center cropping in testing. In the few shot learning task, following [24], we randomly split Mini-ImageNet [37] dataset into 64 base, 16 validation, and 20 novel classes.

CIFAR-100		ImageNet	
Model	top 1	Model	top 1
ResNet-56 [14]	28.80	ResNet-50 [14]	23.68
ResNet-56 [14] + CAM-loss	**27.34**	ResNet-50 [14] + CAM-loss	**22.98**
Wide-ResNet-28-10 [53]	18.37	ResNet-101 [14]	22.30
Wide-ResNet-28-10 [53] + CAM-loss	**17.49**	ResNet-101 [14] + CAM-loss	**21.73**
ResNext-29, 8×64d [46]	18.01	ResNext-50, 4×32d [46]	22.42
ResNext-29, 8×64d [46] + CAM-loss	**17.24**	ResNext-50, 4×32d [46] + CAM-loss	**21.91**
DenseNet-bc-190-40 [18]	17.67	ResNext-101, 8×32d [46]	21.04
DenseNet-bc-190-40 [18] + CAM-loss	**16.98**	ResNext-101, 8×32d [46] + CAM-loss	**20.45**

Table 1. Applicability of CAM-loss to different network structures. Top 1 error rate (%) is adopted, and results of CAM-loss are **bold-faced**.

CIFAR-100			ImageNet		
Baseline Model PyramidNet-200 (alpha=240) [13]	top 1	top 5	Baseline Model ResNet-50 [14]	top 1	top 5
CE loss	16.45	3.69	CE loss	23.68	7.05
CAM-loss	**15.79**	**3.28**	CAM-loss	**22.98**	**6.52**
Cutout [6]	16.53	3.65	Cutout [6]	22.93	6.66
Manifold Mixup [36]	16.14	4.07	Manifold Mixup [36]	22.50	6.21
StochDepth [19]	15.86	3.33	StochDepth [19]	22.46	6.27
DropBlock [9]	15.73	3.26	DropBlock [9]	21.87	5.98
Mixup [54]	15.63	3.99	Mixup [54]	22.58	6.40
Shakedrop [47]	15.08	2.72	-	-	-
Shakedrop [47] + CAM-loss	**14.56**	**2.56**	-	-	-
Cutmix [51]	14.47	2.97	Cutmix [51]	21.54	5.92
Cutmix [51] + CAM-loss	**14.01**	**2.93**	Cutmix [51] + CAM-loss	**21.16**	**5.79**
Cutmix + Shakedrop	13.81	2.29	-	-	-
Cutmix + Shakedrop + CAM-loss	**13.49**	**2.18**	-	-	-

Table 2. Combination with mainstream regularization methods on CIFAR-100 and ImageNet. Top 1 and Top 5 error rate (%) are adopted, and results of CAM-loss are **bold-faced**. Baseline results are obtained from [51]

CUB-200-2011 [38] and Stanford Dogs [20]. The bird dataset contains 5,994 training and 5,794 testing images of 200 classes. The dog dataset contains 12,000 training and 8,580 testing images of 120 classes. For the data augmentation strategy, we rescale the input images to the resolution of 600×600, randomly crop a 448×448 region, and apply a center cropping in testing. In the few shot learning task, following [24], we randomly split the bird dataset into 120 base, 30 validation, and 50 novel classes.

4.2. Image Classification

We evaluate the image classification performances of CAM-loss on three benchmark datasets: CIFAR-10, CIFAR-100 and ImageNet-1K. On CIFAR datasets, we run 160 epochs with batch size 128, initial learning rate 0.1 and cosine learning rate decay. On ImageNet, we run 120 epochs with batch size 1024, initial learning rate 0.4 (batch size 512 and learning rate 0.2 for ResNext-101 due to computation limit) and cosine learning rate decay. Specially we set $c = 3$ and $t = 20$ (In fact, due to the robustness of CAM-loss to hyper parameters, we simply but confidently adopt the same setting in subsequent experiments).

Apply to various network structures. We perform ResNet [14], Wide-ResNet [53], ResNext [46] and DenseNet [18] keeping all hyper parameters the same as original papers. Table 1 shows that CAM-loss can be widely used in a variety of network structures to improve the performances of baselines. Specifically, CAM-loss has brought 0.51-0.70% improvement on ImageNet and 0.69-1.46% improvement on CIFAR-100, which are significant under these large network structures. For further analysis, we focus on the relationship between the number of epoch and the error rate as shown in Figure 4. It is observed that the model trained with CAM-loss achieves higher train error but lower test error, which proves that CAM-loss has a positive effect on avoiding overfitting.

Combine with regularization methods. Table 2 shows the evaluation of different regularization methods on CIFAR-100 and ImageNet following the setup of [51]. We observe that CAM-loss can be widely combined with the mainstream regularization methods to improve their performances further. Specifically, CAM-loss reduces the top-1

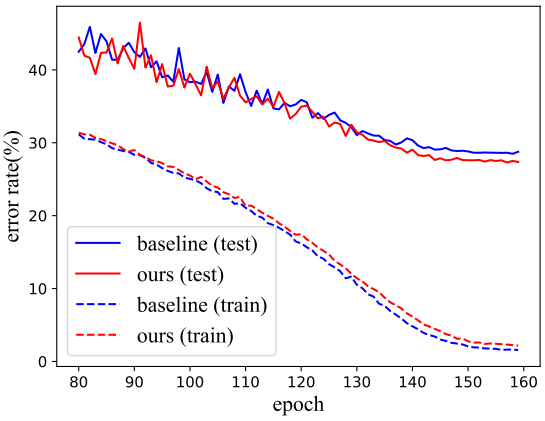

Figure 4. Error rate vs. epoch with CAM-loss and cross entropy loss. We conduct an image classification experiment on CIFAR-100 with a ResNet-56 backbone. The baseline adopts the cross entropy loss while our method adopts CAM-loss.

error rate of ShakeDrop [47] by 0.52% and CutMix [51] by 0.46%. Surprisingly, it also reduces the top-1 error rate of the combination ShakeDrop [47] and CutMix [51] by 0.32%. It means that CAM-loss can further boost the state-of-the-art regularization methods. There is no conflict between CAM-loss and most regularization methods, which is very welcome in practical applications.

Compare with other loss functions. In order to compare CAM-loss with popular loss functions, we conduct classification experiments with ResNet-110 [14] and Wide-ResNet-28-10 [53] on CIFAR-10 and CIFAR-100 following the settings of [42]. Table 3 shows that CAM-loss outperforms all of baseline loss functions. In fact, the routes of CAM-loss and other loss functions are parallel without conflict. They can be used together in an appropriate combine ratio setting. In particular, we compare NegativeCAM [34] with CAM-loss. ResNet-56 [14] and ResNet-110 [14] are adopted on CIFAR-100, and ResNet-50 [14] and ResNet-101 [14] are adopted on ImageNet. We inherit the loss module of Negative-CAM [34] according to the official codes and reproduce the results by utilizing our baseline models. Table 4 shows that CAM-loss consistently outperforms NegativeCAM [34], especially on ImageNet, which means that CAM-loss is more efficient on complex datasets.

Method	CIFAR-10		CIFAR-100	
	ResNet-110	WRN-28-10	ResNet-110	WRN-28-10
CE loss	6.76	3.82	27.68	18.53
center loss [45]	6.38	3.76	26.88	18.50
L-softmax [26]	6.46	3.69	27.03	18.48
SM-softmax [25]	6.49	3.71	26.97	18.40
CAM-loss (ours)	**6.29**	**3.49**	**26.56**	**17.87**

Table 3. Comparison with other loss functions. Top 1 error rate (%) is adopted. Results of CAM-loss are **bold-faced**.

Methods	CIFAR-100		ImageNet	
	ResNet-56	ResNet-110	ResNet-50	ResNet-101
CE loss	28.80	27.68	23.68	22.30
NegativeCAM [34]	27.37	26.76	23.32	22.02
CAM-loss	**27.34**	**26.56**	**22.98**	**21.73**

Table 4. Comparison with NegativeCAM [34] on CIFAR-100 and ImageNet. Top 1 error rate (%) is adopted. Results of CAM-loss are **bold-faced**.

4.3. Transfer learning

We evaluate the generalization ability of CAM-loss under the setting of transfer learning. A ResNet-50 [14] model pre-trained with CAM-loss on ImageNet-1K [5], and then finetune it on CUB [38] and Stanford Dogs [20]. Following [7, 35], we finetune the pre-trained model on the training set for 90 epochs with batch size 8 (CUB) and 16 (Stanford Dogs). SGD optimizer is adopted with an initial learning rate 0.001 and a cosine learning rate decay. We set the weight decay as 5×10^{-4} and momentum as 0.9. Table 5 shows that CAM-loss improves the baseline by 1.1% and 0.7% on CUB [38] and Stanford Dogs [20]. It confirms that CAM-loss has stronger classification capability on novel dataset compared with the cross entropy loss.

Dataset	CE loss	CAM-loss
CUB [38]	85.6	**86.7**
Stanford Dogs [20]	83.9	**84.6**

Table 5. Accuracies of fine-grained classification tasks under the setting of transfer learning. Top 1 accuracy (%) is adopted. Results of CAM-loss are **bold-faced**.

4.4. Few shot learning

Due to limited data on novel classes, few shot learning relies heavily on the generalization ability of models trained on the base classes. The mainstream few shot image classification methods are evaluated and compared in [4], in which Baseline++ is verified with the competitive classification performance and generalization capability. For simplicity, we add CAM-loss to the Baseline++ method to evaluate the performance improvements under three scenarios: (1) general object recognition, (2) fine-grained image classification, (3) cross-domain adaptation (using Mini-ImageNet [37] as base classes and the 50 validation and 50 novel classes from CUB [38]). We adopt the same setting as [4] except for the hyper parameters of CAM-loss.

Table 6 shows that under the standard 5-way 1-shot and 5-shot protocols, CAM-loss averagely boosts Baseline++ by 7.04% and 4.75% on CUB [38], 2.78% and 1.68% on Mini-ImageNet [37]. The cross-domain result of Baseline++ is also significantly improved by 2.75%. As far as we know, our results of few shot classification on CUB [38]

Method	Mini-ImageNet		CUB		Mini→CUB
	1-shot	5-shot	1-shot	5-shot	5-shot
Baseline++ [4]	52.18	75.86	67.08	84.19	65.88
Baseline++ [4] + CAM-loss	**54.80**	**77.54**	**74.12**	**88.93**	**68.63**

Table 6. Accuracies of few shot classification tasks. Mini→CUB represents the cross-domain. Results of CAM-loss are **bold-faced**.

Setup	Teacher	Student	baseline	KD [15]	AT [52]	CCM
(a)	WRN-28-4 [53]	WRN-16-4 [53]	23.14	21.93	21.77	**21.46**
(b)	WRN-28-4 [53]	WRN-28-2 [53]	25.40	23.12	22.82	**22.50**
(c)	WRN-28-4 [53]	WRN-16-2 [53]	27.94	26.05	25.85	**25.45**
(d)	WRN-28-4 [53]	Resnet-56 [14]	28.80	27.11	26.98	**26.48**
(e)	PyramidNet-200 [13]	WRN-28-4 [53]	20.97	20.08	20.22	**19.93**

Table 7. Performance of various knowledge distillation setups on CIFAR-100. 'WRN' denotes Wide-ResNet for short. Baseline denotes the top 1 error rate (%) of the student network. Results of CCM method are **bold-faced**.

has been very competitive with the state-of-the-art in inductive inference setting.

Why does CAM-loss perform so well in few shot image classification tasks? [16, 50] has repeatedly confirmed that the features of background bring great troubles to few shot image classification. A 5-shot learning example is shown in Figure 5, where the features of yellow grass and green grass are misleading. In query set, the lion samples with yellow grass are misclassified as dog while the dog samples with green grass are misclassified as lion. CAM-loss is an effective method to suppress the expression of background features. That is very helpful to reduce the negative effect of background in few shot image classification tasks, especially in the fine-grained few shot image classification.

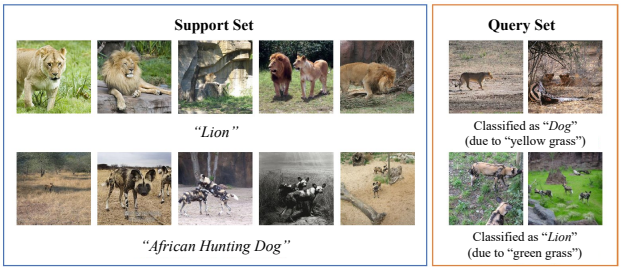

Figure 5. Negative effects of background in few shot image classification. [50]

4.5. Knowledge distillation

We conduct knowledge distillation experiments on CIFAR-100 [21], and choose KD [15] and AT [52] as the baseline methods. For KD [15], we set the hyper parameter temperature as 4 and combine ratio of loss terms as 0.5. For AT [52], we choose the best strategy that is to set $(\beta)l_{\text{ce}} + (1-\beta)l_{\text{kd}} + \gamma l_{\text{at}}$ as loss function where l_{at} adopts l_2 distance, β is set as 0.5 and γ is set as 10. For CCM, we set β as 0.5, γ as 1 following Eq. (10) and (12). For generality of the experiments, we adpoted various teacher/student pairs with the same depth (WRN-28-4/WRN-28-2), different depth (WRN-28-4/WRN-16-2, WRN-28-4/WRN-16-4), different type (WRN-28-4/ResNet-56, PyramidNet-200/WRN-28-4). Table 7 shows that CCM consistently outperforms the two baseline methods. Further thinking, CAM-loss can also be seen as a self-distillation strategy, that is, the supervision information comes from the network itself rather than the teacher network.

5. Conclusion

In this paper, we have proposed a novel loss function CAM-loss to boost the performance of CNN classification models. Essentially, it constrains the feature maps with the spatial information from CAMs. A model trained with CAM-loss is inclined to express the features of target category and suppress those of non-target categories, which is effective to enforce intra-class compactness and inter-class separability. As an independent loss function, it can be easily combined with mainstream regularization methods to improve their performance in image classification tasks. Strong generalization capability makes it outstanding in transfer learning and few shot learning tasks. Based on CAM-loss, we also propose a novel CCM knowledge distillation method, which matches different knowledge between teacher and student. In future, we will study the applications of CAM-loss to more generic visual tasks.

Acknowledgments

This work is supported in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grants 2018AAA0100701, the National Natural Science Foundation of China under Grants 61906106 and 62022048, and Beijing Academy of Articial Intelligence.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019.

[2] Kyungjune Baek, Minhyun Lee, and Hyunjung Shim. Psynet: Self-supervised approach to object localization using point symmetric transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10451–10459, 2020.

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018.

[4] Weiyu Chen, Yencheng Liu, Zsolt KiraAuthors, Yu Chiang, and Huang Jiabin. A closer look at few-shot classification. In *Proceedings of the IEEE International Conference on Learning Representations Worshops*, 2019.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[7] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision*, pages 70–86, 2018.

[8] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019.

[9] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, pages 10727–10737, 2018.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[11] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 729–739, 2019.

[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742, 2006.

[13] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5927–5935, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[16] Ruibing Hou, Hong Chang, MA Bingpeng, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems*, pages 4003–4014, 2019.

[17] Andrew G Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[19] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pages 646–661. Springer, 2016.

[20] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization*, volume 2, 2011.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[23] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018.

[24] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7260–7268, 2019.

[25] Xuezhi Liang, Xiaobo Wang, Zhen Lei, Shengcai Liao, and Stan Z Li. Soft-margin softmax for deep classification. In *International Conference on Neural Information Processing*, pages 413–421. Springer, 2017.

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, volume 2, page 7, 2016.

[27] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019.

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[31] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[34] Guolei Sun, Salman Khan, Wen Li, Hisham Cholakkal, and Fahad Shahbaz. Fixing localization errors to improve image classification. 2020.

[35] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision*, pages 805–821, 2018.

[36] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. *arXiv preprint arXiv:1806.05236*, 2018.

[37] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[39] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1306, 2018.

[40] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[41] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[42] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. In *Advances in Neural Information Processing Systems*, pages 12614–12623, 2019.

[43] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision*, pages 434–450, 2018.

[44] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018.

[45] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[47] Yoshihiro Yamada, Masakazu Iwamura, Takuya Akiba, and Koichi Kise. Shakedrop regularization for deep residual learning. *arXiv preprint arXiv:1802.02375*, 2018.

[48] Seunghan Yang, Yoonhyung Kim, Youngeun Kim, and Changick Kim. Combinational class activation maps for weakly supervised object localization. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2941–2949, 2020.

[49] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[50] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000*, 2020.

[51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[52] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[53] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[54] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[55] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4262–4270, 2018.

[56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.