Spatially Attentive Output Layer for Image Classification

Ildoo Kim* Woonhyuk Baek* Sungwoong Kim
Kakao Brain
Seongnam, South Korea
{ildoo.kim, wbaek, swkim}@kakaobrain.com

Abstract

Most convolutional neural networks (CNNs) for image classification use a global average pooling (GAP) followed by a fully-connected (FC) layer for output logits. However, this spatial aggregation procedure inherently restricts the utilization of location-specific information at the output layer, although this spatial information can be beneficial for classification. In this paper, we propose a novel spatial output layer on top of the existing convolutional feature maps to explicitly exploit the location-specific output information. In specific, given the spatial feature maps, we replace the previous GAP-FC layer with a spatially attentive output layer (SAOL) by employing a attention mask on spatial logits. The proposed location-specific attention selectively aggregates spatial logits within a target region, which leads to not only the performance improvement but also spatially interpretable outputs. Moreover, the proposed SAOL also permits to fully exploit location-specific self-supervision as well as self-distillation to enhance the generalization ability during training. The proposed SAOL with self-supervision and self-distillation can be easily plugged into existing CNNs. Experimental results on various classification tasks with representative architectures show consistent performance improvements by SAOL at almost the same computational cost.

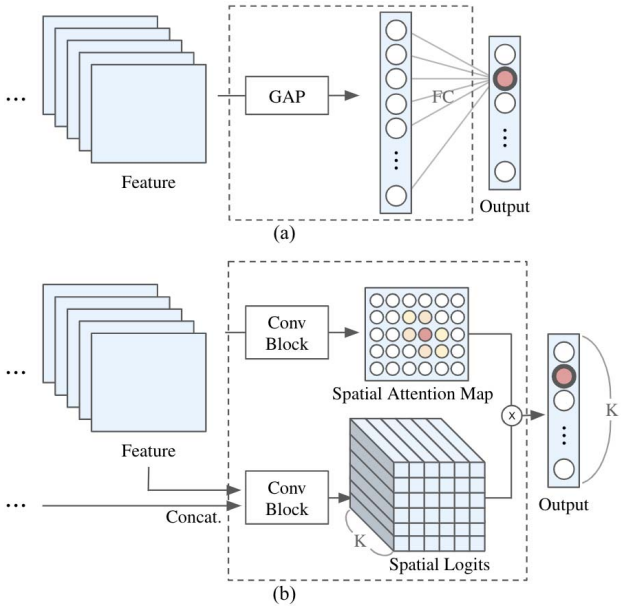

Figure 1: Comparison between (a) the conventional GAP-FC based output layer and (b) the proposed output layer, SAOL. SAOL separately obtains *Spatial Attention Map* and *Spatial Logits* (classification outputs for each spatial location). Then, *Spatial Logits* are weighted averaged by the *Spatial Attention Map* for the final output.

1. Introduction

Deep convolutional neural networks (CNNs) have made great progress in various computer vision tasks including image classification [23, 16], object detection [13, 31, 27], and semantic segmentation [28, 2]. In particular, there have been lots of researches on modifying convolutional blocks and their connections such as depthwise separable convolution [5], deformable ConvNet [7], ResNet [16], and NASNet [48] to improve feature representations. However, in contrast to well-developed convolutional architectures for (multi-scale) spatial feature extraction, the output module to generate the classification logits from the feature maps has been almost unchanged from a standard module that is composed of a global average pooling (GAP) layer and fully-connected (FC) layers. Even though it has shown that CNNs with this feature aggregation can retain its localization ability to some extent [26, 46, 47], in principle, these CNNs have a restriction in full exploitation of benefits from an explicit localization of output logits for image classification.

Recently, the use of localized class-specific responses has drawn increasing attention for image classification, which allows taking the following three main advantages: (1) it can help to interpret the decision making of a CNN

*Contributed equally.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00955

through visual explanation [47, 33, 1]; (2) a spatial attention mechanism can be used for performance improvement by focusing only on the regions that are semantically relevant to the considered labels [21, 38, 36, 10]; and (3) it enables to make use of auxiliary self-supervised losses or tasks based on spatial transformations, which leads to enhanced generalization ability [25, 11, 45, 15, 19, 37].

However, most of the previous methods have obtained spatial logits or attention maps via conventional class activation mapping techniques such as class activation mapping (CAM) [47] and gradient-weighted class activation mapping (Grad-CAM) [33]. They have still utilized the GAP for image-level prediction and thus only located a small part of a target object [25] or attended inseparable regions across classes [37]. While this inaccurate attention mapping hinders its use to improve the classification accuracy, it also has limited an application of self-supervision concerning spatial labeling to maintaining attention consistency under simple spatial transformations such as rotation and flipping [15] or naive attention cropping and dropping [19].

Accordingly, we propose to produce explicit and more precise spatial logits and attention maps as well as to apply useful self-supervision by employing a new output module, called *Spatially Attentive Output Layer* (SAOL). In specific, from the feature maps, we separately obtain the spatial logits (location-specific class responses) and the spatial attention map. Then, the attention weights are used for a weighted sum of the spatial logits to produce the classification result. Figure 1 shows an overall structure of the proposed output layer in comparison to the conventional one.

The proposed output process can be considered as a weighted average pooling over the spatial logits to focus selectively on the target class region. For more accurate spatial logits, we aggregate multi-scale spatial logits inspired by decoder modules used for semantic segmentation [28, 32, 3]. Note that SAOL can generate spatially interpretable attention outputs directly and target object locations during forward propagation without any post-processing. Besides, the computational cost and the number of parameters of the proposed SAOL are almost the same as the previous GAP-FC based output layer.

Furthermore, we apply two novel location-specific self-supervised losses based on CutMix [41] to improve the generalization ability. We remark that different from CutMix, which mixes the ground truth image labels proportionally to the area of the combined input patches, the proposed self-supervision utilizes cut and paste of the self-annotated spatial labels according to the mixed inputs. The proposed losses make our spatial logits and attention map more complete and accurate. We also explore a self-distillation by attaching the conventional GAP-FC as well as SAOL and distilling SAOL logits to GAP-FC. This technique can improve performances of the exiting CNNs without changing their architectures at test time.

We conduct extensive experiments on CIFAR-10/100 [22] and ImageNet [8] classification tasks with various state-of-the-art CNNs and observe that the proposed SAOL with self-supervision and self-distillation consistently improves the performances as well as generates more accurate localization results of the target objects.

Our main contributions can be summarized as follows:

- The SAOL on top of the existing CNNs is newly proposed to improve image classification performances through spatial attention mechanism on the explicit location-specific class responses.
- In SAOL, the normalized spatial attention map is separately obtained to perform a weighted average aggregation over the elaborated spatial logits, which makes it possible to produce interpretable attention outputs and object localization results by forward propagation.
- Novel location-specific self-supervised losses and a self-distillation loss are applied to enhance the generalization ability for SAOL in image-level supervised learning.
- On both of image classification tasks and Weakly Supervised Object Localization (WSOL) tasks with various benchmark datasets and network architectures, the proposed SAOL with self-supervision consistently improves the performances. Additionally, ablation experiments show the benefits from the more accurate spatial attention as well as the more sophisticated location-specific self-supervision.

2. Related Work

Class activation mapping. Class activation mapping methods have been popularly used (1) for visualizing spatial class activations to interpret decision making of the final classification output, (2) for incorporating an auxiliary regularization based on it to boost classification performances, or (3) for performing WSOL. Specifically, CAM [47] can obtain an activation map for each class by linearly combining the last convolutional feature maps with the weights associated with that class at the last FC layer. However, CAM needs to replace the FC layer with convolution and GAP to produce the final classification output. On the other hand, Guided Back-propagation [34], Deconvolution [43], and Grad-CAM [33] was proposed for generating class-wise attention maps by using gradients in back-propagation without requiring architectural changes. Grad-CAM++ [1] modified Grad-CAM to localize multiple instances of the same class more accurately using higher-order derivatives. These methods still adapted the GAP for image-level prediction, which often leads to highlighting only on a discriminative but uncompleted part of a target object.

Attention mechanism. Several works have been recently explored the use of attention mechanism for image classification and WSOL [21, 38, 36, 10]. Residual Atten-

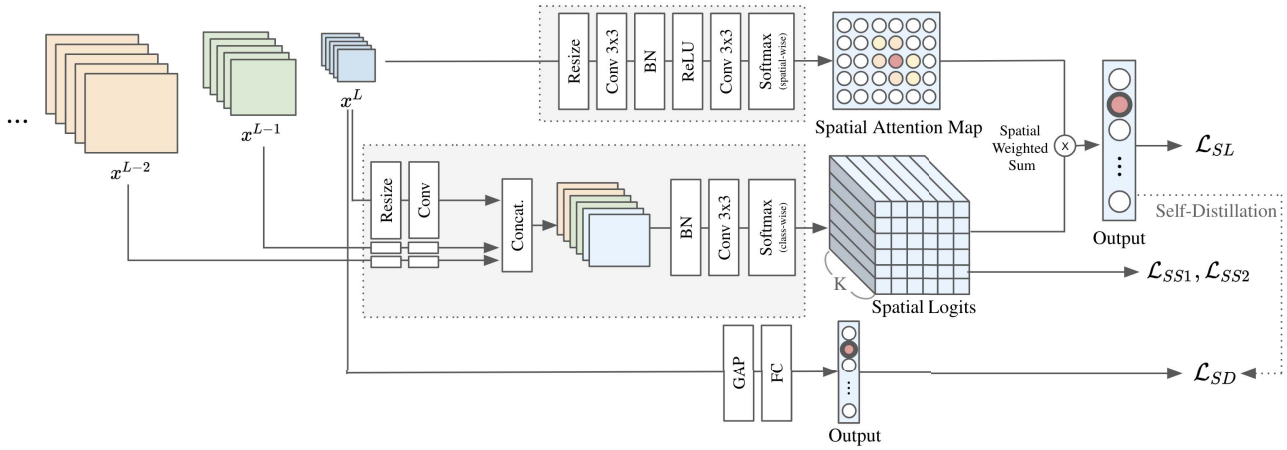

Figure 2: The detailed structure of the proposed SAOL. It produces the spatial attention map and spatial logits, separately. Note that we use additional self-annotated spatial labels to leverage our architecture further. We can also train the conventional GAP-FC based output layer jointly, using self-distillation.

tion Network [36] modified ResNet [16] by stacking multiple soft attention modules that gradually refine the feature maps. Jetley *et al*. [21] proposed a trainable module for generating attention weights to focus on different feature regions relevant to the task of classification at hand. Woo *et al*. [38] introduced a convolutional block attention module that sequentially applies channel and spatial attention modules to refine intermediate feature maps. Attention Branch Network (ABN) [10] designed a separate attention branch based on CAM to generate attention weights and used them to focus on important feature regions. While all of these attention methods refine intermediate feature maps, we apply the attention mechanism on the output layer to directly improve spatial output logits. Girdhar *et al*. [12] introduced a more closely related method based on spatial attention for pooling spatial logits on action recognition tasks. Still, they used simple linear mappings only from the last feature map.

CutMix and attention-guided self-supervision. As an efficient and powerful data augmentation method, CutMix [41] was recently developed, and it significantly outperforms over previous data augmentation methods such as Cutout [9] and Mixup [17]. Yet, CutMix cannot guarantee that a randomly cropped patch always has a part of the corresponding target object with the same proportion used for label-mixing. Several recent works derived auxiliary self-supervised losses using attention maps. For example, Guo *et al*. [15] proposed to enhance attention consistency under simple spatial transformations, and Hu *et al*. [19] applied the attention cropping and dropping to data augmentation. Li *et al*. [25] proposed guided attention inference networks that explore self-guided supervision to optimize the attention maps. Especially, they applied an attention mining technique with image cropping to make complete maps; However, these maps are obtained based on Grad-CAM. Zhang *et al*. [45] introduced adversarial learning to leverage complementary object regions found by CAM to discover entire objects. Wang *et al*. [37] presented new learning objectives for enhancing attention separability and attention consistency across layers. Different from these attention-guided self-supervised learning methods, we design a more sophisticated location-specific self-supervision leveraging CutMix.

3. Methods

In this section, we describe the proposed output layer architecture named SAOL and location-specific self-supervised losses and self-distillation loss in detail.

3.1. Spatially Attentive Output Layer

Let $\mathbf{x}$ and $\mathbf{y}$ denote an input image and its one-hot encoded ground truth label, respectively. For CNN-based image classification, an input $\mathbf{X}^0 = \mathbf{x}$ is first fed into successive L convolution blocks $\{\Theta_\ell(\cdot)\}_{\ell=1}^{L}$, where intermediate feature maps $\mathbf{X}^\ell \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$ at the block ℓ is computed by $\mathbf{X}^\ell = \Theta_\ell(\mathbf{X}^{\ell-1})$. Here, H_ℓ, W_ℓ, and C_ℓ are the height, width, and number of channels at the ℓ_{th} block. Then, the final normalized output logits $\hat{\mathbf{y}} \in [0,1]^K$, which can be considered as an output probability distribution over K classes, are obtained by an output layer $O(\cdot)$ such that $\hat{\mathbf{y}} = O(\mathbf{X}^L)$. In specific, the conventional GAP-FC based output layer $O_{\text{GAP-FC}}(\cdot)$ can be formulated as

$$\hat{\mathbf{y}} = O_{\text{GAP-FC}}(\mathbf{X}^L) = \text{softmax}\Big((\bar{\mathbf{x}}_{\text{GAP}}^L)^T \mathbf{W}^{FC}\Big), \quad (1)$$

where $\bar{\mathbf{x}}_{\text{GAP}}^L \in \mathbb{R}^{C_L \times 1}$ denotes the spatially aggregated feature vector by GAP, and $\mathbf{W}^{FC} \in \mathbb{R}^{C_L \times K}$ is the weight matrix of the output FC layer. Here, $(\bar{\mathbf{x}}_{\text{GAP}}^L)_c = \frac{\sum_{i,j}(\mathbf{X}_c^L)_{ij}}{H_\ell W_\ell}$, where $(\mathbf{X}_c^L)_{ij}$ is the $(i,j)_{th}$ element of the c_{th} feature map

$\mathbf{X}_c^L$ at the last block. Instead of this aggregation on the last feature map, our method produces output logits explicitly on each spatial location and then aggregates them selectively through the spatial attention mechanism.

Specifically, the proposed SAOL, $O_{\text{SAOL}}(\cdot)$, first produces *Spatial Attention Map*, $\mathbf{A} \in [0,1]^{H_o \times W_o}$, and *Spatial Logits*, $\mathbf{Y} \in [0,1]^{K \times H_o \times W_o}$, separately. Here, it is noted that we set $H_o = H_L$ and $W_o = W_L$ by default. The attention values are normalized via softmax across the spatial positions while we take softmax on the spatial logits across classes: $\sum_{i,j} \mathbf{A}_{ij} = 1, \forall k$ and $\sum_k (\mathbf{Y}_k)_{ij} = 1, \forall i, j$. Then, we generate the final output logits by a spatially weighted sum of the spatial logits as follows:

$$\hat{\mathbf{y}}_k = O_{\text{SAOL},k}(\mathbf{X}^L) = \sum_{i,j} \mathbf{A}_{ij} (\mathbf{Y}_k)_{ij}, \quad \forall k, \tag{2}$$

where $\hat{\mathbf{y}}_k$ is the output logit of the k_{th} class. These attention weights indicate the relative importance of each spatial position regarding the classification result.

The architecture in SAOL is described in detail in Figure 2. First, to obtain the spatial attention map $\mathbf{A}$, we feed the last convolutional feature maps $\mathbf{X}^L$ into two-layered convolutions followed by the softmax function. At the same time, for the sake of the precise spatial logits, we combine multi-scale spatial logits, motivated by previous decoder modules for semantic segmentation [28, 32, 3]. In specific, at each of the selected blocks, the feature maps are mapped to the intermediate spatial logits through convolutions after resized to the output spatial resolution. Then, a set of the intermediate spatial logits are concatenated and re-mapped to the final spatial logits $\mathbf{Y}$ by another convolution layer and the softmax function. Note that in contrast to CAM [47] and Grad-CAM [33], this SAOL can directly generate spatially interpretable attention outputs or target object locations using $\mathbf{A}$ and $\mathbf{Y}$ in a feed-forward manner. This makes it possible to use location-specific regularizers during training, as presented in the next subsection.

3.2. Self-Supervised Losses

The proposed SAOL performs well when trained even only with the general cross-entropy loss $\mathcal{L}_{CE}$ as our supervised loss such that $\mathcal{L}_{SL} = \mathcal{L}_{CE}(\hat{\mathbf{y}}_{\text{SAOL}}, \mathbf{y})$[1]. However, in order to fully utilize location-specific output information to boost the classification performance, we add two novel spatial losses inspired by CutMix [41] and self-supervised learning methods [11, 24].

CutMix generates a new training sample $(\mathbf{x}', \mathbf{y}')$ by mixing a certain sample $(\mathbf{x}_B, \mathbf{y}_B)$ and a random patch extracted from an another sample $(\mathbf{x}_A, \mathbf{y}_A)$ as follows:

$$\begin{aligned} \mathbf{x}' &= \mathbf{M} \odot \mathbf{x}_A + (\mathbf{1} - \mathbf{M}) \odot \mathbf{x}_B, \\ \mathbf{y}' &= \lambda \mathbf{y}_A + (1 - \lambda) \mathbf{y}_B, \end{aligned} \tag{3}$$

[1]We let $\hat{\mathbf{y}}_{\text{GAP-FC}}$ and $\hat{\mathbf{y}}_{\text{SAOL}}$ denote the final output logits from the GAP-FC based output layer and those from SAOL, respectively.

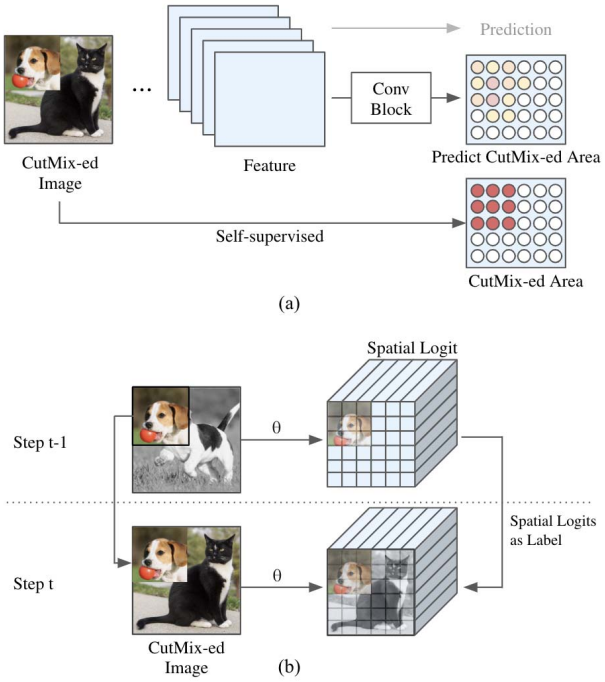

Figure 3: The proposed two self-supervisions based on CutMix for SAOL: (a) $\mathcal{L}_{SS1}$ and (b) $\mathcal{L}_{SS2}$.

where $\mathbf{M}$ denotes a binary mask for cropping and pasting a rectangle region, and λ is a combination ratio sampled using the beta distribution. This label-mixing strategy implies that a cut region should have the meaning as much as the size of the cropped area in the context of its label. However, this assumption would often be incorrect since a randomly cropped patch can fail to capture a part of the corresponding target object, especially when the target object is small.

Specifically, we use two additional self-annotated spatial labels and self-supervised losses, as illustrated in Figure 3. Given a CutMix-ed input image, the first self-supervised loss $\mathcal{L}_{SS1}$ uses $\mathbf{M}$ as an additional ground truth label after resizing to $H_o \times W_o$. We add an auxiliary layer similar to the attention layer to predict $\hat{\mathbf{M}} \in [0,1]^{H_o \times W_o}$. Since $\mathbf{M}$ is the binary mask, the binary cross-entropy loss is used as

$$\mathcal{L}_{SS1} = \mathcal{L}_{BCE}(\hat{\mathbf{M}}, \mathbf{M}). \tag{4}$$

The second self-supervised loss $\mathcal{L}_{SS2}$ we propose is to match the spatial logits in the pasted region of the mixed input with the spatial logits in the cut region of the original data as follows:

$$\mathcal{L}_{SS2} = D_{KL}(\mathbf{M} \odot \mathbf{Y}', \mathbf{M} \odot \mathbf{Y}_A), \tag{5}$$

where D_{KL} represents the Kullback–Leibler divergence[2], and $\mathbf{Y}_A$ denotes the spatial logits of $\mathbf{x}_A$. Since these self-

[2]It is actually the average Kullback–Leibler divergence over spatial positions.

supervisions regularize the network either to identify the specific pasted location or to produce the same spatial logits in the pasted region, these can lead to spatially consistent feature representations and accordingly, improved performances. Note that we update the network through the gradients only from $\mathbf{M} \odot \mathbf{Y}'$.

3.3. Self-Distillation Loss

Since one can insert the proposed SAOL in the existing CNNs, we utilize both the previous GAP-FC based output layer and SAOL, as shown in Figure 2, during training. Specifically, we come up with knowledge transfer from SAOL to the existing output layer. For this, we devise a self-distillation loss $\mathcal{L}_{SD}$ with the two final output logits separately obtained by the two output layers from a given input image, as follows:

$$\mathcal{L}_{SD} = D_{KL}(\hat{\mathbf{y}}_{\text{SAOL}}, \hat{\mathbf{y}}_{\text{GAP-FC}}) + \beta \mathcal{L}_{CE}(\hat{\mathbf{y}}_{\text{GAP-FC}}, \mathbf{y}), \quad (6)$$

where β is the relative weight between the two loss terms, which was similarly used in other self-distillation methods [44, 24]. We set $\beta = 0.5$. At test-time, we take only one of the two output modules to produce the classification result. If we select the GAP-FC based output layer, we can improve the classification performances of the exiting CNNs without computational tax at test time, although it is negligible.

In the end, the final loss $\mathcal{L}$ that we use during training is defined as

$$\mathcal{L} = \mathcal{L}_{SL} + \mathcal{L}_{SS1} + \mathcal{L}_{SS2} + \mathcal{L}_{SD}, \quad (7)$$

where further improvement may be possible using different ratios of losses.

4. Experiments

We evaluate our SAOL with self-supervision and self-distillation compared to the previous methods. We first study the effects of our proposed method on several classification tasks in Section 4.1. Then, to conduct a quantitative evaluation for the obtained attention map, WSOL experiments were performed in Section 4.2.

All experiments were implemented in PyTorch [30], by modifying the official CutMix source code[3]. For a fair comparison, we tried not to change the hyper-parameters from baselines such as CutMix [41] and ABN [10]. We simultaneously trained both of SAOL and the GAP-FC based output layer via the proposed self-distillation loss in an end-to-end manner. At test time, we obtained the classification results by either SAOL or the GAP-FC based output layer.

4.1. Image Classification Tasks

4.1.1 CIFAR-10, CIFAR-100 Classification

The first performance evaluation for image classification is carried out on CIFAR-10 and CIFAR-100 benchmark [22], one of the most extensively studied classification tasks. We used the same hyper-parameters for Wide-ResNet [42] from AutoAugment [6]. ResNet and DenseNet models were trained with the same settings for ABN [10] to compare each other. For PyramidNet200 (widening factor $\bar{\alpha} = 240$), we used the same hyper-parameters used in CutMix [41], except for the learning rate and its decay schedule. We used 0.1 as the initial learning rate for cosine annealing schedule [29]. While our baselines did not obtain much better results with this slight change, the proposed SAOL achieved noticeable performance improvements. Every experiment was performed five times to report its average performance.

Table 1 and Table 2 compare the baseline and the proposed method on CIFAR-10 and CIFAR-100, respectively. The proposed SAOL outperformed the baseline consistently across all models except DenseNet-100. In addition, in most cases for CIFAR-10, SAOL gave clear improvements over self-distilled GAP-FC. However, our self-distilled GAP-FC was also consistently better than the baseline. This means that even without spatial supervision such as object localization label, SAOL can learn spatial attention appropriately and eventually performs better than averaging features. This consistent improvement was also retained when we additionally used CutMix during training.

We also compare SAOL with recently proposed ABN [10]. There are similarities between the two methods in respect of using the attention map. However, SAOL uses the attention map to aggregate spatial output logits. In contrast, ABN makes use of the attention mechanism only on the last feature maps and adapts the previous GAP-FC layer. For ResNet-110 and DenseNet-100, we trained models with the same hyper-parameters used in ABN. ResNet-110 and DenseNet-100 with ABN achieved the accuracies of 95.09%, 95.83% on CIFAR-10 and 77.19%, 78.37% on CIFAR-100, respectively. These results indicate that models with SAOL perform much better than models with ABN. We emphasize that ABN also requires more computations. To be specific, ResNet-110 with ABN requires 5.7 GFLOPs, while ResNet-110 with SAOL only requires 2.1 GFLOPs. As the original ResNet-110 computes as much as 1.7 GFLOPs, not only SAOL is more effective and efficient than ABN, but also it provides a way to keep the amount of computation intact through self-distillation.

4.1.2 ImageNet Classification

We also evaluate SAOL on ILSVRC 2012 classification benchmark (ImageNet) [8] which consists of 1.2 million natural images for training and 50,000 images for validation of 1,000 classes. We used the same hyper-parameters with CutMix [41]. For faster training, we just changed the batch size to 4,096 with a linearly re-scaled learning rate and a gradual warm-up schedule, as mentioned in [14]. We also replaced all convolutions in SAOL with depthwise-

[3] https://github.com/clovaai/CutMix-PyTorch

Model	Baseline GAP-FC	Ours SAOL	Ours $^{\text{self-distilled}}$ GAP-FC
Wide-ResNet 40-2 [42]	94.80	**95.33 (+0.53)**	95.31 (+0.51)
Wide-ResNet 40-2 + CutMix [41]	96.11	**96.44 (+0.33)**	96.44 (+0.33)
Wide-ResNet 28-10 [42]	95.83	**96.44 (+0.61)**	96.42 (+0.59)
Wide-ResNet 28-10 + CutMix [41]	97.08	**97.37 (+0.29)**	97.36 (+0.28)
ResNet-110 [16]	93.57*	**95.18 (+1.61)**	95.06 (+1.49)
ResNet-110 + CutMix [41]	95.77	**96.21 (+0.44)**	96.17 (+0.40)
DenseNet-100 [20]	**95.49***	95.31 (-0.18)	95.35 (-0.14)
DenseNet-100 + CutMix [41]	95.83	**96.27 (+0.44)**	96.19 (+0.36)
PyramidNet200 + ShakeDrop [40]	97.13	**97.33 (+0.20)**	97.31 (+0.18)
PyramidNet200 + ShakeDrop + CutMix [41]	97.57	**97.93 (+0.36)**	97.92 (+0.35)

Table 1: Classification Top-1 accuracies (%) on CIFAR-10. Results from the original papers are denoted as *.

Model	Baseline GAP-FC	Ours SAOL	Ours $^{\text{self-distilled}}$ GAP-FC
Wide-ResNet 40-2 [42]	74.73	**76.50 (+1.77)**	76.18 (+1.45)
Wide-ResNet 40-2 + CutMix [41]	78.21	**79.53 (+1.32)**	79.04 (+0.83)
Wide-ResNet 28-10 [42]	80.13	80.89 (+0.76)	**81.16 (+1.03)**
Wide-ResNet 28-10 + CutMix [41]	82.41	**83.71 (+1.30)**	83.71 (+1.30)
ResNet-110 [16]	75.86*	77.15 (+1.29)	**77.23 (+1.37)**
ResNet-110 + CutMix [41]	77.94	**78.02 (+0.08)**	77.94 (+0.00)
DenseNet-100 [20]	**77.73***	76.84 (-0.89)	76.25 (-1.48)
DenseNet-100 + CutMix [41]	78.55	**79.25 (+0.70)**	78.90 (+0.35)
PyramidNet200 + ShakeDrop [40]	84.43	84.72 (+0.29)	**84.95 (+0.52)**
PyramidNet200 + ShakeDrop + CutMix [41]	86.19	86.95 (+0.76)	**87.03 (+0.84)**

Table 2: Classification Top-1 accuracies (%) on CIFAR-100. Results from the original papers are denoted as *.

Model	Baseline GAP-FC	Ours SAOL	Ours $^{\text{self-distilled}}$ GAP-FC
ResNet-50 [16]	76.32 / 92.95*	**77.11 / 93.59**	76.66 / 93.25
ResNet-50 + CutMix [41]	78.60 / 94.10*	**78.85 / 94.24**	78.09 / 94.00
ResNet-101 [16]	78.13 / 93.71*	**78.59 / 94.25**	78.22 / 93.82
ResNet-101 + CutMix [41]	79.83 / 94.76*	**80.49 / 94.96**	80.24 / 94.84
ResNext-101 [39]	78.82 / 94.43*	**79.23 / 95.03**	79.23 / 94.97
ResNext-101 + CutMix [41]	80.53 / 94.97*	**81.01 / 95.15**	80.81 / 95.03
ResNet-200 [16]	78.50 / 94.20	**79.31 / 94.54**	78.92 / 94.37
ResNet-200 + CutMix [41]	80.70 / 95.20	**80.82 / 95.19**	80.73 / 95.21

Table 3: ImageNet classification Top-1 / Top-5 accuracies (%). Results from the original papers are denoted as *.

separable convolutions [18] to reduce computations. We found that in many situations, this convolution change made a marginal difference in performances.

Table 3 shows performances with diverse architectures. We quoted results from the CutMix paper except for ResNet-200, which was not tested by CutMix. We trained all models with the same hyper-parameters for a fair comparison. Our results indicate that models with SAOL outperformed the models with GAP-FC consistently. For example, ResNet-101 architecture trained with CutMix regularization scored 79.83% of top-1 accuracy, which is improved from 78.13% without CutMix. For both cases, SAOL further improves the model by 0.46% and 0.66% without and with CutMix, respectively. We remark that adding our SAOL requires 6% more computations only (from 7.8 GFLOPs to 8.3 GFLOPs), which is efficient compared to the previous methods. As shown in Figure 4, SAOL performed better than both of Residual Attention Network [36] and ABN [10], especially even with much smaller computational cost.

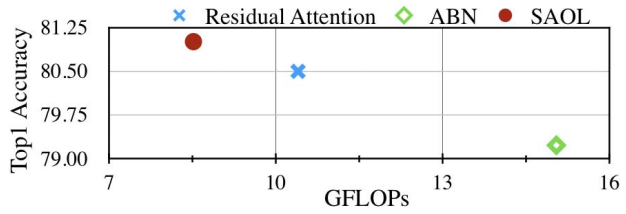

Figure 4: Comparison of different attention models on ImageNet. Attention layers are added on the same ResNet-200 backbone. Our model (SAOL) outperforms previous methods [10, 36] using negligible computational overhead.

	WResNet 40-2	WResNet 28-10
Conv Block 3	75.68	79.99
Conv Block 2+3	76.18	80.70
Conv Block 1+2+3	**76.50**	**80.89**

Table 4: Performance comparisons on CIFAR-100 according to different combinations of feature blocks used for producing the spatial logits. WResNet stands for Wide-ResNet. Wide-ResNet has three convolutional blocks, and we denote the ith block as Conv Block i.

4.1.3 Ablation Study

In this section, we conduct ablation experiments for many factors in SAOL to measure their contributions towards our outperforming results.

Effectiveness of Multi-level Feature Aggregation for Spatial Logits. SAOL uses features not only from the last convolution block but from multiple intermediate blocks for producing the spatial logits. In detection and segmentation tasks, majority of works [3][32][2][27] similarly used multiple feature layers in a decoder to be more size-invariant. We experimented on CIFAR-100 to verify performance changes according to different numbers of features to be combined to generate the spatial logits for SAOL, and Table 4 shows the obtained results. Performances tend to be improved with more feature layers for spatial logits.

Effectiveness of Self-Supervision. To verify the benefits from the proposed two self-supervised losses, we conducted experiments with Wide-ResNet 40-2 on CIFAR-10 and CIFAR-100 (C-100), and the results are shown in Table 5. Similar to the baseline model, SAOL was also improved with the original CutMix regularization alone. However, additional incorporating $\mathcal{L}_{SS1}$ or $\mathcal{L}_{SS2}$ further enhanced the performances. Using both of self-supervised losses with SAOL led to the best performance.

Note that we also tried to use $\mathcal{L}_{SS1}$ on the baseline. For this, we attached an auxiliary layer on the last convolution block to produce a spatial map predicting the CutMix region and trained the original image classification loss and

	CIFAR-10	C-100
Baseline (GAP-FC)	94.80	74.73
Baseline + CutMix	96.11	78.21
Baseline + CutMix + $\mathcal{L}_{SS1}$	96.04	78.14
SAOL	95.33	76.50
SAOL + CutMix	96.21	78.44
SAOL + CutMix + $\mathcal{L}_{SS1}$	96.19	78.92
SAOL + CutMix + $\mathcal{L}_{SS2}$	96.30	78.60
SAOL + CutMix + $\mathcal{L}_{SS1}$ + $\mathcal{L}_{SS2}$	**96.44**	**79.53**

Table 5: Influences of CutMix and its additional self-supervised losses for Wide-ResNet 40-2 on CIFAR-10/100.

$\mathcal{L}_{SS1}$ jointly. As a result, the use of $\mathcal{L}_{SS1}$ did not improve the performance of the baseline. We conjecture that SAOL worked well with $\mathcal{L}_{SS1}$ since it tried to learn the attention map for classification outputs simultaneously. We leave a more detailed investigation of this for future work.

Effectiveness of Self-Distillation. We also conducted experiments on CIFAR-100 to measure the effectiveness of our self-distillation. Instead of distilling outputs from SAOL, the standard cross-entropy (CE) loss was solely applied to the GAP-FC auxiliary layer during training. The results are shown in Table 6. Irrespective of the selected output layer at test time, training both of SAOL and the GAP-FC based output layer with the same CE loss led to performance drop compared to the use of our self-distillation loss $\mathcal{L}_{SD}$, even though it still outperformed the baseline. This indicates that the knowledge transfer from robust SAOL to the conventional GAP-FC based output layer by our self-distillation is beneficial to performance improvement.

	WResNet 40-2		WResNet 28-10	
	SAOL	GAP-FC	SAOL	GAP-FC
Baseline	N/A	74.73	N/A	80.13
CE	75.75	75.28	80.36	80.21
$\mathcal{L}_{SD}$	**76.50**	**76.18**	**80.89**	**81.16**

Table 6: Evaluation on the effectiveness of self-distillation.

4.2. Weakly-Supervised Object Localization Task

To evaluate the spatial attention map by SAOL quantitatively, we performed experiments with ResNet-50 models for the tasks of WSOL. We followed the evaluation strategy of the existing WSOL method [47]. A common practice in WSOL is to normalize the score maps using min-max normalization to have a value between 0 and 1. The normalized output score map can be binarized by a threshold, then the largest connected area in the binary mask is chosen. Our model was modified to enlarge the spatial resolutions of the spatial attention map and spatial logits to be 14×14 from 7×7 and finetuned ImageNet-trained model. The obtained

Model	Method	GFLOPs	Backprop.	CUB200-2011 Loc Acc (%)	ImageNet Loc Acc (%)
ResNet-50 [16]	CAM [47]	4.09	O	49.41*	46.30*
ResNet-50 [16] + CutMix [41]	CAM [47]	4.09	O	54.81*	47.25*
ResNet-50 [16]	ABN [10]	7.62	X	56.91	44.65
ResNet-50 [16] + CutMix [41]	SAOL (Ours)	4.62	X	52.39	45.01

Table 7: Weakly supervised object localization results on CUB200-2011 test set and ImageNet validation set. The asterisk * indicates that the score is from the original paper.

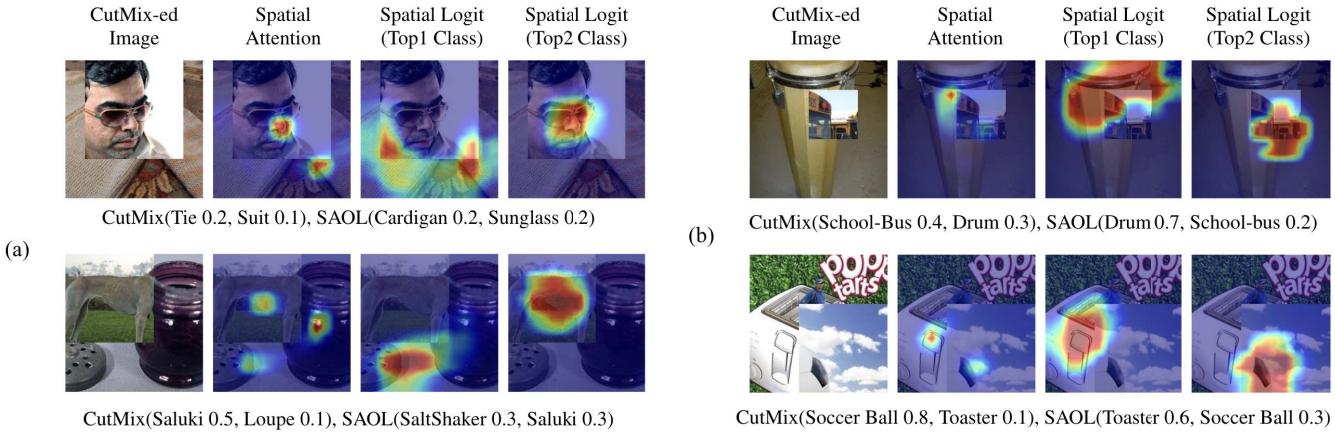

Figure 5: Qualitative analysis of attention maps by SAOL with ResNet-50. From the left: CutMix-ed image, spatial attention map, heatmap of spatial output logit for top-2 classes. (a) Examples that previous CutMix model [41] failed to correctly predict objects with top-2 classes' scores. (b) Examples that previous CutMix model predicted small objects over-confidently.

spatial attention map and spatial logits are combined as an elemental-wise product to yield a class-wise spatial attention map.

As the result are shown in Table 7, our method achieves competitive localization accuracy on ImageNet and CUB200-2011 [35], compared to previous well-performing methods [4, 41]. It is noticeable that our competitive method requires much fewer computations to generate an attention map for object localization. While it is common to use CAM [47], burdensome backward-pass computations are unavoidable. Recently proposed ABN [10] can produce an attention map with the single forward pass; however, it modifies the backbone network with a computationally-expensive attention mechanism. SAOL adds much less computational taxes while it performs competitively. We also emphasize that our results were obtained without any sophisticated post-processing, which is required by many WSOL methods. Utilizing sophisticated post-processing as well as training with a larger attention map may improve the result further.

Figure 5 visualizes the spatial attention map and the spatial logits obtained by SAOL on CutMix-ed image. Our spatial attention map focuses on the regions corresponding to the general concept of objectness. On the other hand, the spatial output logits show class-specific activation maps which have high scores on the respective target object regions. In the situation where two objects are mixed, the attention map by SAOL localizes each object well, and moreover its scores reflect the relative importance of each object more accurately.

5. Conclusion

We propose a new output layer for image classification, named spatially attentive output layer (SAOL). Outputs from the novel two branches, spatial attention map and spatial logits, generate the classification outputs through an attention mechanism. The proposed SAOL improves the performances of representative architectures for various tasks, with almost the same computational cost. Moreover, additional self-supervision losses specifically designed for SAOL also improve the performances further. The attention map and spatial logits produced by SAOL can be used for weakly-supervised object localization (WSOL), and it shows promising results not only for WSOL tasks but also towards interpretable networks. We will continue this research to develop better decoder-like output structures for image classification tasks and to explore a more sophisticated use of self-annotated spatial information without human labor.

References

[1] 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 (WACV)*, 2018. 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834 – 848, 2018. 1, 7

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 4, 7

[4] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2219–2228, 2019. 8

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1251–1258, 2017. 1

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 113–123, 2019. 5

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 764–773, 2017. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255. Ieee, 2009. 2, 5

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[10] 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 (CVPR)*, pages 10705–10714, 2019. 2, 3, 5, 6, 7, 8

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2, 4

[12] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*, pages 34–45, 2017. 3

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014. 1

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[15] 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 (CVPR)*, pages 729–739, 2019. 2, 3

[16] 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 (CVPR)*, pages 770–778, 2016. 1, 3, 6, 8

[17] Yann N. Dauphin David Lopez-Paz Hongyi Zhang, Moustapha Cisse. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 3

[18] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6

[19] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019. 2, 3

[20] 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 (CVPR)*, pages 4700–4708, 2017. 6

[21] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H.S. Torr. Learn to pay attention. In *International Conference on Learning Representations*, 2018. 2, 3

[22] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2, 5

[23] 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. 1

[24] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Rethinking data augmentation: Self-supervision and self-distillation. *arXiv preprint arXiv:1910.05872*, 2019. 4, 5

[25] 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 (CVPR)*, pages 9215–9223, 2018. 2, 3

[26] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations*, 2014. 1

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition (CVPR), pages 2117–2125, 2017. 1, 7

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015. 1, 2, 4

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019. 5

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4, 7

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 4

[34] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations*, 2015. 2

[35] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 8

[36] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3156–3164, 2017. 2, 3, 6, 7

[37] Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N. Metaxas. Sharpen focus: Learning with attention separability and consistency. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 512–521, 2019. 2, 3

[38] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2, 3

[39] 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 (CVPR)*, pages 1492–1500, 2017. 6

[40] Y. Yamada, M. Iwamura, T. Akiba, and K. Kise. Shakedrop regularization for deep residual learning. *IEEE Access*, 7:186126–186136, 2019. 6

[41] 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 (ICCV)*, pages 6023–6032, 2019. 2, 3, 4, 5, 6, 8

[42] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5, 6

[43] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 2

[44] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3713–3722, 2019. 5

[45] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1325–1334, 2018. 2, 3

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *International Conference on Learning Representations*, 2015. 1

[47] 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 (CVPR)*, pages 2921–2929, 2016. 1, 2, 4, 7, 8

[48] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8697–8710, 2018. 1