Sharpen Focus: Learning with Attention Separability and Consistency

Lezi Wang[1], Ziyan Wu[2], Srikrishna Karanam[2], Kuan-Chuan Peng[2],
Rajat Vikram Singh[2], Bo Liu[1], and Dimitris N. Metaxas[1]
[1]Rutgers University, New Brunswick NJ
[2]Siemens Corporate Technology, Princeton NJ
{lw462, lb507, dnm}@cs.rutgers.edu, {ziyan.wu, srikrishna.karanam, kuanchuan.peng, singh.rajat}@siemens.com

Abstract

Recent developments in gradient-based attention modeling have seen attention maps emerge as a powerful tool for interpreting convolutional neural networks. Despite good localization for an individual class of interest, these techniques produce attention maps with substantially overlapping responses among different classes, leading to the problem of visual confusion and the need for discriminative attention. In this paper, we address this problem by means of a new framework that makes class-discriminative attention a principled part of the learning process. Our key innovations include new learning objectives for attention separability and cross-layer consistency, which result in improved attention discriminability and reduced visual confusion. Extensive experiments on image classification benchmarks show the effectiveness of our approach in terms of improved classification accuracy, including CIFAR-100 (+3.33%), Caltech-256 (+1.64%), ILSVRC2012 (+0.92%), CUB-200-2011 (+4.8%) and PASCAL VOC2012 (+5.73%).

1. Introduction

Visual recognition has seen tremendous progress in the last few years, driven by recent advances in convolutional neural networks (CNNs) [13, 17]. Understanding their predictions can help interpret models and provide cues to design improved algorithms.

Recently, class-specific attention has emerged as a powerful tool for interpreting CNNs [5, 31, 45]. The big-picture intuition that drives these techniques is to answer the following question- *where is the target object in the image?* Some recent extensions [20] make attention end-to-end trainable, producing attention maps with better localizability. While these methods consider the localization problem, this is insufficient for image classification, where the model needs to be able to tell various object classes apart. Specifically, existing methods produce attention maps corresponding to an individual class of interest that may not be *discriminative* across classes. Our intuition, shown in Figure 1, is that such separable attention maps can lead to improved classification performance. Furthermore, we contend that false classifications stem from patterns across classes which confuse the model, and that eliminating these confusions can lead to better model discriminability. To illustrate this, consider Figure 2 (a), where we use the VGG-19 model [33] to perform classification on the ILSVRC2012 [30] dataset, we collect failure cases and generate the attention maps via Grad-CAM [31] and we show the top-5 predictions. Figure 2 (a) depicts that, while the attention maps of the last feature layer are reasonably well localized, there are large overlapping regions between the attention of the ground-truth class (marked by red bounding boxes) and the false positives, demonstrating the problem, and the need for *discriminative attention*.

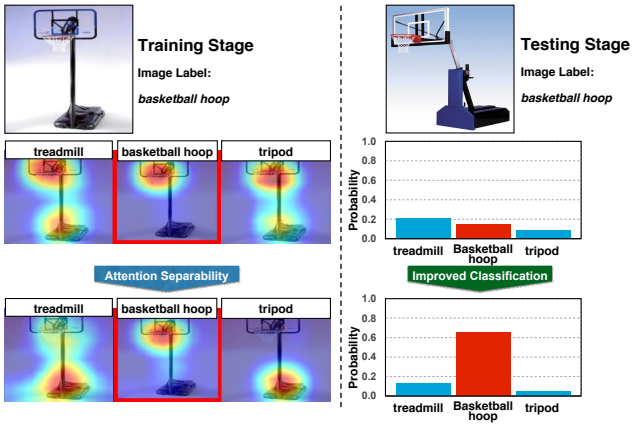

Figure 1. The baseline CNN attends to similar regions, *i.e.* the central areas, when it comes to the relevant pixels for classes "treadmill," "basketball hoop," and "tripod." The CNN with our proposed framework is able to tell the three classes apart and has high confidence to classify the input as "basketball hoop."

To overcome the above attention-map limitations, we need to address two key questions: (a) *can we reduce visual confusion, i.e., make class-specific attention maps*

2380-7504/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCV.2019.00060

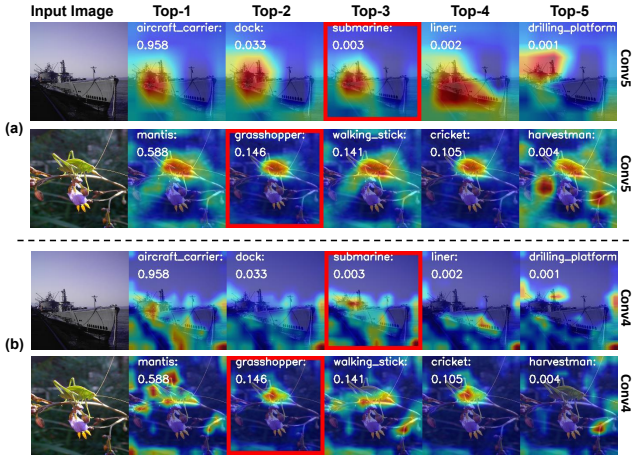

Figure 2. Grad-CAM [31] attention maps of the VGG-19 [33] top-5 predictions. Predictions with red-bounding boxes correspond to the ground-truth class. (a) Ground-truth class attention maps from the last layer (Conv5) have a large overlap with false positives (top-1 predictions). (b) Inner-layer attention maps (Conv4) are more separable than their last-layer counterparts.

separable and discriminative across different classes?, and (b) *can we incorporate attention discriminability in the learning process in an end-to-end fashion?* We answer these questions in a principled manner, proposing the first framework that makes attention maps class discriminative. Furthermore, we propose a new attention mechanism to guide model training towards attention discriminability, which provides end-to-end supervisory signals by explicitly enforcing attention maps of various classes to be separable.

Attention separability and localizability are key aspects of our proposed learning framework for image classification. Non-separable attention maps from the last layer, as shown in Figure 2 (a), prompted us to look "further inside" the CNN and Figure 2 (b) shows attention maps from an intermediate layer. This illustration shows that these inner-layer attention maps are more separable than those from the last layer. However, the inner-layer attention maps are not as well-localized as the last layer. So, another question we ask is- *can we get the separability of the inner-layer attention and the localization of the last-layer attention at the same time?* Solving this problem would result in a "best-of-both-worlds" attention map that is separable and localized, which is our goal. To this end, our framework also includes an explicit mechanism that enforces the ground-truth class attention to be cross-layer consistent.

We conduct extensive experiments on five competitive benchmarks (CIFAR-100 [19], Caltech-256 [12], ILSVRC2012 [30], CUB-200-2011 [36] and PASCAL VOC 2012 [10]), showing performance improvements of 3.33%, 1.64%, 0.92%, 4.8%, and 5.73%, respectively.

In summary, we make the following contributions:

- We propose channel-weighted attention $\mathcal{A}_{ch}$, which has better localizability and avoids higher-order derivatives computation, compared to existing approaches for attention-driven learning.
- We propose attention separation loss L_{AS}, the first learning objective to enforce the model to produce class-discriminative attention maps, resulting in improved attention separability.
- We propose attention consistency loss L_{AC}, the first learning objective to enforce attention consistency across different layers, resulting in improved localization with "inner-layer" attention maps.
- We propose "Improving Classification with Attention Separation and Consistency" (ICASC), the first framework that integrates class-discriminative attention and cross-layer attention consistency in the conventional learning process. ICASC is flexible to be used with available attention mechanisms, *i.e.* Grad-CAM [31] and $\mathcal{A}_{ch}$, providing the learning objectives for training CNN with discriminative and consistent attention, which results in improved classification performance.

2. Related work

Visualizing CNNs. Much recent effort has been expended in visualizing internal representations of CNNs to interpret the model better. Erhan *et al*. [9] synthesized images to maximally activate a network unit. Mahendran *et al*. [24] and Dosovitskiy *et al*. [8] analyzed the visual coding to invert latent representations, performing image reconstruction by feature inversion with an up-convolutional neural network. In [32, 34, 41], the gradient of the prediction was computed w.r.t. the specific CNN unit to highlight important pixels. These approaches are compared in [25, 31]. The visualizations are fine-grained but not class-specific, where visualizations for different classes are nearly identical [31].

Our framework is inspired by recent works [5, 31, 45] addressing class-specific attention. CAM [45] generated class activation maps highlighting task-relevant regions by replacing fully-connected layers with convolution and global average pooling. Grad-CAM [31] solved CAM's inflexibility where without changing the model architecture and retraining the parameters, class-wise attention maps were generated by means of gradients of the final prediction w.r.t. pixels in feature maps. However, we observe that directly averaging gradients in Grad-CAM [31] results in the improper measurement of channel importance, producing substantial attention inconsistency among various feature layers. Grad-CAM++ [5] proposed to introduce higher-order derivatives to capture pixel importance, while its high computational cost in calculating the second-

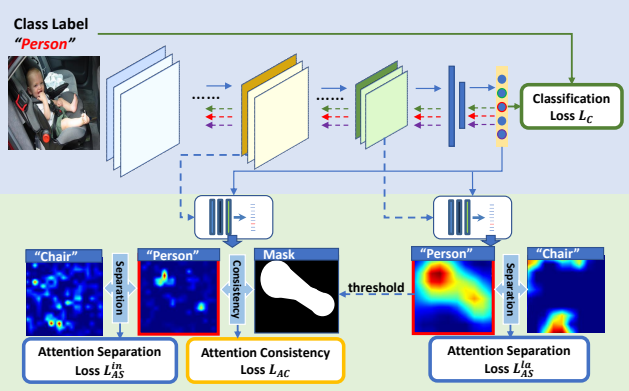

Figure 3. The framework of Improving Classification with Attention Separation and Consistency (ICASC).

and third-order derivatives makes it impractical to be used during training.

Attention-guided network training. Several recent methods [14, 17, 38, 40] have attempted to incorporate attention mechanisms to improve the performance of CNNs in image classification. Wang *et al*. [38] proposed Residual Attention Network, modifying ResNet [13] by adding the hourglass network [26] to the skip-connection, generating attention masks to refine feature maps. Hu *et al*. [14] introduced a Squeeze-and-Excitation (SE) module which used globally average-pooled features to compute channel-wise attention. CBAM [27, 40] modified the SE module to exploit both spatial and channel-wise attention. Jetley *et al*. [17] estimated attentions by considering the feature maps at various layers in the CNN, producing a 2D matrix of scores for each map. The ensemble of output scores was then used for class prediction. While these methods use attention for downstream classification, they do not explicitly use class-specific attention as part of model training for image classification.

Our work, to the best of our knowledge, is the first to use class-specific attention to produce supervisory signals for end-to-end model training with attention separability and cross-layer consistency. Furthermore, our proposed method can be considered as an add-on module to existing image classification architectures without needing any architectural change, unlike other methods [14, 17, 38, 40]. While class-specific attention has been used in the past for weakly-supervised object localization and semantic segmentation tasks [6, 20, 39, 43], we model attention differently. The goal of these methods is singular - to make the attention well localize the ground-truth class, while our goal is two-fold - good attention localizability as well as discriminability. To this end, we devise novel objective functions to guide model training towards discriminative attention across different classes, leading to improved classification performance as we show in the experiments section.

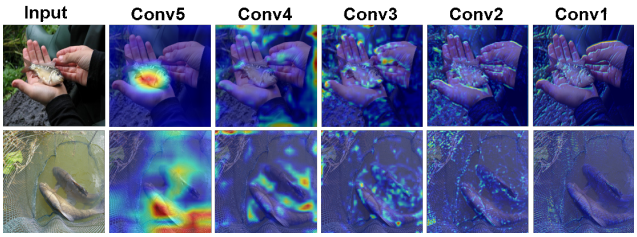

Figure 4. The Grad-CAM [31] attentions of different VGG-19 [33] feature layers for the 'tench' class. In both rows, the target is the fish while the model attention shifts across the layers.

3. Approach

In Figure 3, we propose "Improving Classification with Attention Separation and Consistency" (ICASC), the first end-to-end learning framework to improve model discriminability for image classification via attention-driven learning. The main idea is to produce separable attention across various classes, providing supervisory signals for the learning process. The motivation comes from our observations from Figure 2 that the last layer attention maps computed by the existing methods such as Grad-CAM [31] are not class-separable, although they are reasonably well localized. To address this problem, we propose the attention separation loss L_{AS}, a new attention-driven learning objective to enforce attention discriminability.

Additionally, we observe from Figure 2 that inner layer attention at higher resolution has the potential to be separable, which suggests we consider both intermediate and the last layer attention to achieve separability and localizability at the same time. To this end, we propose the attention consistency loss L_{AC}, a new cross-layer attention consistency learning objective to enforce consistency among inner and last layer attention maps. Both proposed learning objectives require that we obtain reasonable attention maps from the inner layer. However, Grad-CAM [31] fails to produce intuitively satisfying inner layer attention maps. To illustrate this, we depict two Grad-CAM [31] examples in Figure 4, where we see the need for better inner layer attention. To this end, we propose a new channel-weighted attention mechanism $\mathcal{A}_{ch}$ to generate improved attention maps (explained in Sec. 3.1). We then discuss how we use them to produce supervisory signals for enforcing attention separability and cross-layer consistency.

3.1. Channel-weighted attention $\mathcal{A}_{ch}$

Commonly-used techniques to compute gradient-based attention maps given class labels include CAM [45], Grad-CAM [31], and Grad-CAM++ [5]. We do not use CAM because (a) it is inflexible, requiring network architecture modification and model re-training, and (b) it works only for the last feature layer.

Compared to CAM [45], Grad-CAM [31] and Grad-

CAM++ [5] are both flexible in the sense that they only need to compute the gradient of the class prediction score w.r.t. the feature maps to measure pixel importance. Specifically, given the class score Y^c for the class c and the feature map F^k in the k-th channel, the class-specific gradient is determined by computing the partial derivative $(\partial Y^c)/(\partial F^k)$. The attention map is then generated as $\mathcal{A} = ReLU(\sum_k \alpha_k^c F^k)$, where α_k^c indicates the importance of F^k in the k-th channel. In Grad-CAM [31], the weight α_k^c is a global average of the pixel importance in $(\partial Y^c)/(\partial F^k)$:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial Y^c}{\partial F_{ij}^k} \tag{1}$$

where Z is the number of pixels in F^k. Grad-CAM++ [5] further introduces higher-order derivatives to compute α_k^c so as to model pixel importance.

Although Grad-CAM [31] and Grad-CAM++ [5] are more flexible than CAM [45], they have several drawbacks that hinder their use *as is* for our purposes of providing separable and consistent attention guidance for image classification. First, there are large attention shifts among attention maps of different feature layers in Grad-CAM [31] which are caused by negative gradients while computing channel-wise importance. A key aspect of our proposed framework ICASC is to exploit the separability we observe in inner layer attention in addition to good localization from the last layer attention. While we observe relatively less attention shift with Grad-CAM++ [5], the high computational cost of computing higher-order derivatives precludes its use in ICASC since we use attention maps from multiple layers to guide model training in every iteration.

To address these issues, we propose channel-weighted attention $\mathcal{A}_{ch}$, highlighting the pixels where the gradients are positive. In our exploratory experiments, we observed that the cross-layer inconsistency of Grad-CAM [31], noted above, is due to negative gradients from background pixels. In Grad-CAM [31], all pixels of the gradient map contribute equally to the channel weight (Eq. 1). Therefore, in cases where background gradients dominate, the model tends to attend only to small regions of target objects, ignoring regions that are important for class discrimination.

We are motivated by prior work [5, 34, 41] that observes that positive gradients w.r.t. each pixel in the feature map F^k strongly correlate with the importance for a certain class. A positive gradient at a specific location implies increasing the pixel intensity in F^k will have a positive impact on the prediction score, Y^c. To this end, driven by positive gradients, we propose a new channel-weighted attention mechanism $\mathcal{A}_{ch}$:

$$\mathcal{A}_{ch} = \frac{1}{Z} ReLU(\sum_k\sum_i\sum_j ReLU(\frac{\partial Y^c}{\partial F_{ij}^k})F^k) \tag{2}$$

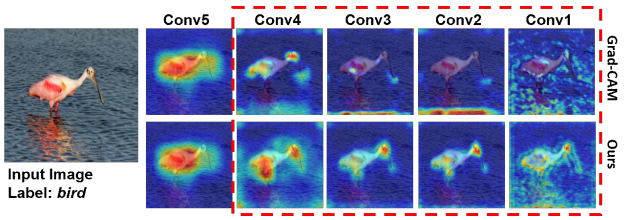

Figure 5. The comparison of attention maps from different VGG-19 [33] layers. Ours has less attention shift than Grad-CAM [31]. In the marked areas, ours attends to the target objects, *i.e.* bird, while Grad-CAM [31] tends to highlight the background pixels.

Our attention does not need to compute higher-order derivatives as in Grad-CAM++ [5], while also resulting in well-localized attention maps with relatively less shift unlike Grad-CAM [31], as shown in Figure 5.

3.2. Attention separation loss L_{AS}

We use the notion of attention separability as a principled part of our learning process and propose a new learning objective L_{AS}. Essentially, given the attention map of a ground-truth class $\mathcal{A}^T$ and the most confusing class $\mathcal{A}^{Conf}$, where $\mathcal{A}^{Conf}$ comes from the non-ground truth class with the highest classification probability, we enforce the two attentions to be separable. We reflect this during training by quantifying overlapping regions between $\mathcal{A}^T$ and $\mathcal{A}^{Conf}$, and minimizing it. To this end, we propose L_{AS} which is defined as:

$$L_{AS} = 2 \cdot \frac{\sum_{ij}(min(\mathcal{A}_{ij}^T, \mathcal{A}_{ij}^{Conf}) \cdot Mask_{ij})}{\sum_{ij}(\mathcal{A}_{ij}^T + \mathcal{A}_{ij}^{Conf})}, \tag{3}$$

where the $\cdot$ operator indicates scalar product, and $\mathcal{A}_{ij}^T$ and A_{ij}^{Conf} represent the $(i,j)^{th}$ pixel in attention maps $\mathcal{A}^T$ and $\mathcal{A}^{Conf}$ respectively. The proposed L_{AS} is differentiable which can be used for model training.

Additionally, to reduce noise from background pixels, we apply a mask to focus on pixels within the target object region for the L_{AS} computation. In Eq. 3, $Mask$ indicates the target object region generated by thresholding the attention map $\mathcal{A}^T$ from the last layer:

$$Mask_{ij} = \frac{1}{1 + exp(-\omega(\mathcal{A}_{ij}^T - \sigma))}, \tag{4}$$

where we empirically choose values of σ and ω to be $0.55 \times max(\mathcal{A}_{ij}^T)$ and 100 respectively.

The intuition of L_{AS} is illustrated in Figure 6. If the model attends to the same or overlapped regions for different classes, it results in visual confusion. We penalize the confusion by explicitly reducing the overlap between the attention maps of the target and the most confusing class. Specifically, we minimize L_{AS}, which is differentiable with values ranging from 0 to 1.

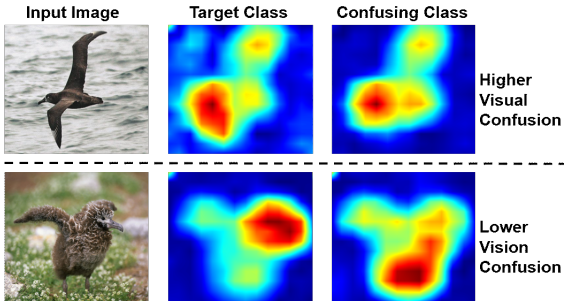

Figure 6. The top row demonstrates higher visual confusion than the bottom row. The two attention maps in the top row have high responses localized at the bird's head, while as shown in the bottom, the ground-truth class attention highlights the bird's head, and the confusing class attention addresses the lower body.

The proposed L_{AS} can be considered an add-on module for training a model without changing the network architecture. Besides applying L_{AS} to the last feature layer, we can also compute L_{AS} for any other layers, which makes it possible for us to analyze model attention at various scales.

While the proposed L_{AS} helps enforce attention separability, it is not sufficient for image classification since inner layer attention maps are not as spatially well-localized as the last layer. We set out to achieve an attention map to be well-localized and class-discriminative, and to this end, we propose a new cross-layer attention consistency objective L_{AC} that enforces the target attention map from an inner layer to be similar to that from the last layer.

3.3. Attention consistency loss L_{AC}

In higher layers (layers closer to output), the model attention captures more semantic information, covering most of the target object [5, 31, 45]. For the intermediate layers with the smaller receptive fields of the convolution kernels, the model attends to more fine-grained patterns as shown in Figures 4 and 5. Compared to higher-layer attention, lower-layer attention contains more noise, highlighting background pixels.

To address these issues, we propose the attention consistency loss L_{AC} to correct the model attention so that the highlighted fine-grained attention is primarily localized in the target region:

$$L_{AC} = \theta - \frac{\sum_{ij}(\mathcal{A}^{in}_{ij} \cdot Mask_{ij})}{\sum_{ij} \mathcal{A}^{in}_{ij}}, \tag{5}$$

where $\mathcal{A}^{in}$ indicates attention maps from the inner feature layers, $Mask_{ij}$ (defined in Eq. 4) represents the target region, and θ is set to 0.8 empirically. As can be noted from Eq. 5, the intuition of L_{AC} is that by exploiting last layer attention's good localizability, we can guide the inner layer attention to be chiefly concentrated within the target region as well. This guidance L_{AC} helps maintain cross-layer attention consistency.

3.4. Overall framework ICASC

We apply the constraints of attention separability and cross-layer consistency jointly as supervisory signals to guide end-to-end model training, as shown in Figure 3. Firstly, we compute inner-layer attention for the loss L^{in}_{AS} with the purpose of enforcing inner-layer attention separability. For example, with ResNet, we use the last convolutional layer in the penultimate block. We empirically adopt this to compute L^{in}_{AS} in consideration of the low-level patterns and semantic information addressed by the inner-layer attention. In Figure 5, this inner-layer attention, with twice resolution as the last layer, highlights more fine-grained patterns while still preserving the semantic information, thus localizing the target object. We also apply the L_{AS} constraint on the attention map from the last layer, giving us L^{la}_{AS}. Secondly, we apply the cross-layer consistency constraint L_{AC} between the attention maps from these two layers. Finally, for the classification loss L_C, we use cross-entropy and multilabel-soft-margin loss for single and multi-label image classification respectively. The overall training objective of ICASC, L, is:

$$L = L_C + L^{in}_{AS} + L^{la}_{AS} + L_{AC} \tag{6}$$

ICASC can be used with available attention mechanisms including Grad-CAM [31] and $\mathcal{A}_{ch}$. We use ICASC$_{Grad-CAM}$ and ICASC$_{\mathcal{A}_{ch}}$ to refer to our framework used with Grad-CAM [31] and $\mathcal{A}_{ch}$ as the attention mechanisms respectively.

4. Experiments

Our experiments contain two parts, (a) evaluating the class discrimination of various attention mechanisms, and (b) demonstrating the effectiveness of the proposed ICASC by comparing it with the corresponding baseline model (having the same architecture) without the attention supervision. We conduct image classification experiments on various datasets, consisting of three parts: generic image classification on CIFAR-100 (D_{CI}) [19], Caltech-256 (D_{Ca}) [12] and ILSVRC2012 (D_I) [30], fine-grained image classification on CUB-200-2011 (D_{CU}) [36], and finally, multi-label image classification on PASCAL VOC 2012 (D_P) [10]. For simplicity, we use the shorthand in the parenthesis after the dataset names above to refer to each dataset and its associated task, and summarize all experimental parameters used in Table 1. We perform all experiments using PyTorch [28] and NVIDIA Titan X GPUs. We use the same training parameters as those in the baselines proposed by the authors of the corresponding papers for fair comparison.

4.1. Evaluating class discriminability

We first evaluate class-discriminability of our proposed attention mechanism $\mathcal{A}_{ch}$ by measuring both localizability

Task	D_{CI}	D_{Ca}	D_I	D_{CU}	D_P
BNA	RN-18	VGG RN-18	RN-18	RN-50 RN-101	RN-18
WD	$5e^{-4}$	$1e^{-3}$	$1e^{-4}$	$5e^{-4}$	$1e^{-3}$
MOM	0.9	0.9	0.9	0.9	0.9
LR	$1e^{-1}$	$1e^{-2}$	$1e^{-1}$	$1e^{-3}$	$1e^{-2}$
BS	128	16	256	10	16
OPM	SGD	CCA	SGD	SGD	CCA
# epoch	160	20	90	90	20
Exp. setting	[13]	[12]	[30]	[35]	[10]

Table 1. Experimental (exp.) settings used in this paper. VGG, RN-18, RN-50, and RN-101 denote VGG-19 [32], ResNet-18 [13], ResNet-50, and ResNet-101, respectively. We use the same parameters as the references in the last row unless otherwise specified, putting more details in the supplementary material. Acronyms: BNA: base network architecture; WD: weight decay; MOM: momentum; LR: initial learning rate; BS: batch size; OPM: optimizer; SGD: stochastic gradient descent [3]; CCA: cyclic cosine annealing [15].

(identifying target objects) and discriminability (separating different classes). We conduct experiments on the PASCAL VOC 2012 dataset. Specifically, with a VGG-19 model trained only with class labels (no pixel-level segmentation annotations), we generate three types of attention maps from the last feature layer: Grad-CAM, Grad-CAM++, and $\mathcal{A}_{ch}$. The attention maps are then used with DeepLab [7] to generate segmentation maps, which are used to report both qualitative (Figure 7 and 8) and quantitative results (Table 2), where we train Deeplab[1] in the same way as SEC [18] is trained in [21], using attention maps as weak localization cues. The focus of our evaluation here is targeted towards demonstrating class discriminability, and segmentation is merely used as a proxy task for this purpose.

Figure 7 shows that $\mathcal{A}_{ch}$ (ours) has better localization for the two classes, "Bird" and "Person" compared to Grad-CAM and Grad-CAM++. In "Bird," both Grad-CAM and Grad-CAM++ highlight false positive pixels in the bottom-left area, whereas in "Person," Grad-CAM++ attends to a much larger region than Grad-CAM and $\mathcal{A}_{ch}$. Figure 8 qualitatively demonstrates better class-discriminative segmentation maps using $\mathcal{A}_{ch}$. In Figure 8 top row, as expected for a single object, all methods, including $\mathcal{A}_{ch}$, show good performance localizing the sheep. The second row shows that Grad-CAM covers more noise pixels of the grassland, while $\mathcal{A}_{ch}$ produces similar results as Grad-CAM++, both of which are better than Grad-CAM in identifying multiple instances of the same class. Finally, for multi-class images in the last row, $\mathcal{A}_{ch}$ demonstrates superior results when compared to both Grad-CAM and Grad-CAM++. Specifically, $\mathcal{A}_{ch}$ is able to tell

[1]https://github.com/tensorflow/models/tree/master/research/deeplab

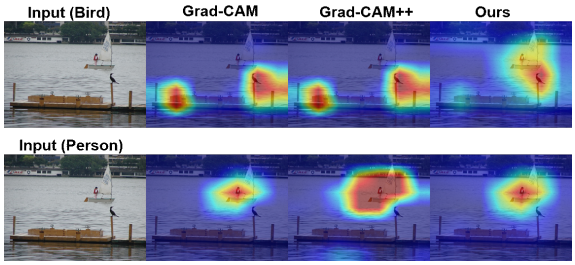

Figure 7. Multi-class attention maps ('bird' and 'person').

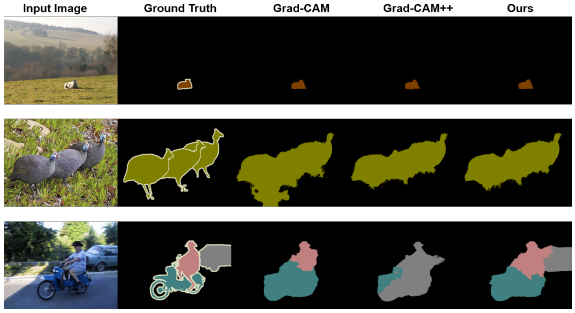

Figure 8. Segmentation masks generated from attention maps by DeepLab [7] (best view in color, zoom in). From left to right: the Input Image, Ground Truth, Grad-CAM, Grad-CAM++ and ours.

Attention Mechanism	Score
Grad-CAM [31]	56.65
Grad-CAM++ [5]	51.70
$\mathcal{A}_{ch}$ (ours)	**57.97**

Table 2. Results on Pascal VOC 2012 segmentation validation set.

Method	Top-1	Δ
ResNet-50	81.70	-
+ L_{AS}^{in}	85.15	3.45
+ L_{AS}^{in} + L_{AC}	85.77	4.07
+ L_{AS}^{in} + L_{AS}^{la} + L_{AC}	**86.20**	4.50

Table 3. Ablation study on CUB-200-2011 (Δ=performance improvement; "Top-1": top-1 accuracy (%)).

the motorcycle, the person, and the car apart in the last row. We also obtain the quantitative results and report the score from the Pascal VOC Evaluation server in Table 2, where $\mathcal{A}_{ch}$ outperforms both Grad-CAM and Grad-CAM++. The qualitative and quantitative results show that $\mathcal{A}_{ch}$ localizes and separates target objects better than the baselines, motivating us to use $\mathcal{A}_{ch}$ in ICASC, which we evaluate next.

4.2. Evaluating L_{AS} and L_{AC} for image classification

4.2.1 Ablation study

Table 3 shows an ablation study with the CUB-200-2011 dataset, which provides a challenging testing set given its fine-grained nature. We use the last convolutional layer in

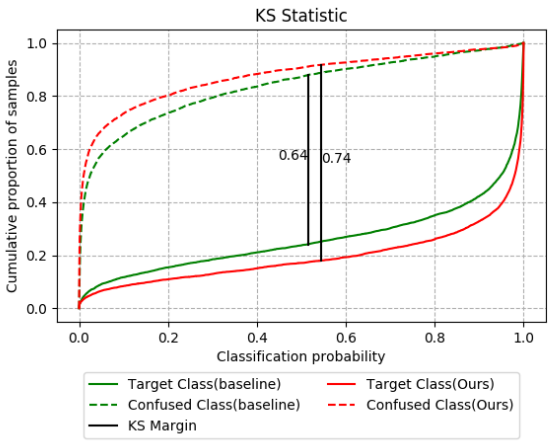

Figure 9. The KS-Chart on the CUB-200-2011 testing set. "Ours" stands for ResNet-50 + L_{AS}^{in} + L_{AS}^{la} + L_{AC} in Table 3.

the penultimate block of ResNet-50 for computing L_{AS}^{in} and the last layer attention map for L_{AS}^{la}. We see that L_{AS}^{in} + L_{AS}^{la} + L_{AC} achieves the best performance. The results show that the attention maps from the two different layers are complementary: last-layer attention has more semantic information, well localizing the target object, and inner layer attention with higher resolution provides fine-grained details. Though the inner-layer attention is more likely to be noisy than the last layer, L_{AC} provides the constraint to guide the inner-layer attention to be consistent with that of the last layer and be concentrated within the target region.

We quantitatively measure the degree of visual confusion reduction with our proposed learning framework. Specifically, as shown in Figure 9, we compute Kolmogorov-Smirnov (KS) statistics [1] on the CUB-200-2011 testing set, measuring the degree of separation between the ground-truth (Target) class and the most confusing (Confused) class distributions [23]. We rank non-ground truth classes in descending order according to their classification probabilities and determine the most confusing class as the one ranked highest. In Figure 9, for the baseline model, the largest margin is 0.64 at the classification probability 0.51 whereas our proposed model has a KS margin of 0.74 at the classification probability 0.55. This demonstrates that our model is able to recognize 10% more testing samples with higher confidence when compared to the baseline.

4.2.2 Generic image classification

Tables 4-6 (in all tables, △ indicates performance improvement of our method over baseline) show that the models trained with our proposed supervisory principles outperform the corresponding baseline models with a notable margin. The most noticeable performance improvements are observed with the CIFAR-100 dataset in Table 4, which shows that, without changing the network architecture, the top-1 accuracy of ResNet-110 with our proposed supervision outperforms the baseline model by 3.33%. Our supervised ResNet-110 also outperforms the one with stochastic depth and even the much deeper model with 164 layers. As can be observed from the qualitative results in Figure 10, ICASC$_{\mathcal{A}_{ch}}$ equips the model with discriminative attention where the ground-truth class attention is separable from the confusing class, resulting in improved prediction.

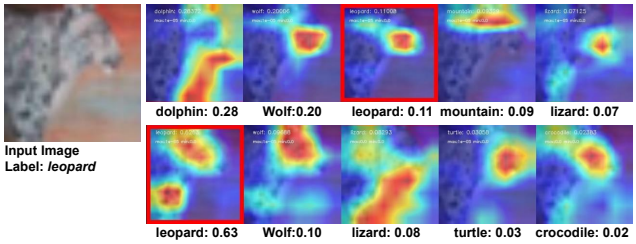

Figure 10. Qualitative results with CIFAR-100. We show top-5 predictions with classification scores given by ResNet-110 (top row) and ResNet-110 + ICASC$_{\mathcal{A}_{ch}}$ (bottom row).

Method	Top-1	△
ResNet-110 [16]	72.78	-
ResNet-110 with Stochastic Depth [16]	75.42	-
ResNet-164 (pre-activation) [16]	75.63	-
ResNet-110 + ICASC$_{Grad-CAM}$	74.02	1.24
ResNet-110 + ICASC$_{\mathcal{A}_{ch}}$	**76.11**	**3.33**

Table 4. Image classification results on CIFAR-100.

Method	N=30			N=60		
	Top-1	Top-5	△	Top-1	Top-5	△
RN-18 [13]	76.77	92.48	-	80.01	94.12	-
RN-18 + ICASC$_{\mathcal{A}_{ch}}$	**78.01**	**92.87**	1.24	**81.32**	**94.57**	1.31
VGG-19 [33]	74.52	90.05	-	78.16	92.17	-
VGG-19 + ICASC$_{\mathcal{A}_{ch}}$	**75.60**	**90.85**	1.08	**79.80**	**93.25**	1.64

Table 5. Results on Caltech-256. "Top-5": top-5 accuracy (%). "RN-18": ResNet-18. "N": # of training images per class. We follow [12] to randomly select 30 or 60 training images per class.

Method	Top-1	Top-5	△
ResNet-18 [13]	69.51	88.91	-
ResNet-18 + ICASC$_{\mathcal{A}_{ch}}$	**69.90**	**89.71**	0.39
ResNet-18 + tenCrop [13]	72.12	90.58	-
ResNet-18 + tenCrop + ICASC$_{\mathcal{A}_{ch}}$	**73.04**	**90.65**	0.92

Table 6. Results on ILSVRC2012.

4.2.3 Fine-grained Image Recognition

For fine-grained image recognition, we evaluate our approach on the CUB-200-2011 dataset [36], which

Method	No Extra Anno.	1-Stage	Top-1	Δ
ResNet-50 [35]	✓	✓	81.7	-
ResNet-101 [35]	✓	✓	82.5	0.8
MG-CNN [37]	✘	✘	83.0	1.3
SPDA-CNN [42]	✘	✓	85.1	3.4
RACNN [11]	✓	✓	85.3	3.6
PN-CNN [4]	✘	✘	85.4	3.7
RAM [22]	✓	✘	86.0	4.3
MACNN + 2parts [44]	✓	✓	85.4	3.7
ResNet-50 + MAMC [35]	✓	✓	86.2	4.5
ResNet-101 + MAMC [35]	✓	✓	**86.5**	4.8
ResNet-50 + ICASC$_{\mathcal{A}_{ch}}$	✓	✓	86.2	4.5
ResNet-101 + ICASC$_{\mathcal{A}_{ch}}$	✓	✓	**86.5**	4.8

Table 7. Results on CUB-200-2011. "No Extra Anno." means not using extra annotation (bounding box or part) in training. "1-Stage" means the training is done in one stage.

contains 11788 images (5994/5794 for training/testing) of 200 bird species. We show the results in Table 7. We observe that training with our learning mechanism boosts the accuracy of the baseline ResNet-50 and ResNet-101 by 4.8% and 4.0% respectively. Our method achieves the best overall performance against the state-of-the-art. Furthermore, with ResNet-50, our method outperforms even the method that uses extra annotations (PN-CNN) by 0.8%.

ICASC$_{\mathcal{A}_{ch}}$ has better flexibility compared to the other methods in Table 7. The existing methods are specifically designed for fine-grained image recognition where, according to prior knowledge of the fine-grained species, the base network architectures (BNA) are modified to extract features of different objects parts [35, 42, 44]. In contrast, ICASC$_{\mathcal{A}_{ch}}$ needs no prior knowledge and works for generic image classification without changing the BNA.

4.2.4 Multi-class Image Classification

We conduct multi-class image classification on the PASCAL VOC 2012 dataset, which contains 20 classes. Different from the above generic and fine-grained image classification where each image is associated with one class label, for each of the 20 classes, the model predicts the probability of the presence of an instance of that class in the test image. As our attention is class-specific, we can seamlessly adapt our pipeline from single-label to multi-label classification. Specifically, we apply the one-hot encoding to corresponding dimensions in the predicted score vector and compute gradients to generate the attention for multiple classes. As for the most confusing class, we consistently determine it as the non-ground truth class with the highest classification probability.

For evaluation, we report the Average Precision (AP) from the PASCAL Evaluation Server [10]. We also compute the AUC score via scikit-learn python module [29] as an additional evaluation metric [2]. Table 8 shows that ResNet-18 [13] with $\mathcal{A}_{ch}$ outperforms the baseline by 5.73%.

Method	AUC Score	AP (%)	△
ResNet-18 [13]	0.976	77.44	-
ResNet-18 + ICASC$_{\mathcal{A}_{ch}}$	**0.981**	**83.17**	5.73

Table 8. Results on Pascal VOC 2012.

Pascal VOC 2012	Top-1	Caltech-256	Top-1
ResNet-18	77.44	ResNet-18	80.01
+ ICASC$_{Grad-CAM}$	82.12	+ ICASC$_{Grad-CAM}$	80.28
+ ICASC$_{\mathcal{A}_{ch}}$	**83.17**	+ ICASC$_{\mathcal{A}_{ch}}$	**81.32**
CUB-200-2011	Top-1	ILSVRC2012	Top-1
ResNet-50	81.70	ResNet-18	69.51
+ ICASC$_{Grad-CAM}$	85.45	+ ICASC$_{Grad-CAM}$	69.84
+ ICASC$_{\mathcal{A}_{ch}}$	**86.20**	+ ICASC$_{\mathcal{A}_{ch}}$	**69.90**

Table 9. Comparing baseline, ICASC$_{Grad-CAM}$ and ICASC$_{\mathcal{A}_{ch}}$.

4.2.5 Comparing attention mechanisms

We compare the image classification performance when ICASC is trained with Grad-CAM [31] and $\mathcal{A}_{ch}$. As can be noted from the results in Table 4 and 9, the higher Top-1 accuracy of ICASC$_{\mathcal{A}_{ch}}$ shows that our attention mechanism provides better supervisory signals for model training than Grad-CAM [31]. Additionally, even ICASC with Grad-CAM still outperforms the baseline, further validating our key contribution of attention-driven learning for reducing visual confusion. The proposed ICASC is flexible to be used with any existing attention mechanisms as well, while resulting in improved classification performance.

5. Conclusions

We propose a new framework, ICASC, which makes class-discriminative attention a principled part of training a CNN for image classification. Our proposed attention separation loss and attention consistency loss provide supervisory signals during training, resulting in improved model discriminability and reduced visual confusion. Additionally, our proposed channel-weighted attention has better class discriminability and cross-layer consistency than existing methods (e.g. Grad-CAM [31]). ICASC is applicable to any trainable network without changing the architecture, giving an end-to-end solution to reduce visual confusion. ICASC achieves performance improvements on various medium-scale, large-scale, fine-grained, and multi-class classification tasks. While we select last two feature layers which contain most semantic information to generate the attention maps, ICASC is flexible w.r.t. layer choices for attention generation, and we plan to study the impact of various layer choices in the future.

References

[1] *Kolmogorov–Smirnov Test*, pages 283–287. Springer New York, New York, NY, 2008. 7

[2] Alexander Binder, Klaus-Robert Müller, and Motoaki Kawanabe. On taxonomies for multi-class image categorization. *International Journal of Computer Vision*, 99(3):281–301, 2012. 8

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 6

[4] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. In *British Machine Vision Conference*, 2014. 8

[5] 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)*, pages 839–847. IEEE, 2018. 1, 2, 3, 4, 5, 6

[6] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *BMVC*, 2017. 3

[7] 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. *TPAMI*, 40(4):834–848, April 2018. 6

[8] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016. 2

[9] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. 2

[10] Mark Everingham, Luc Van Gool, Chris Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2, 5, 6, 8

[11] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, volume 2, page 3, 2017. 8

[12] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 2, 5, 6, 7

[13] 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. 1, 3, 6, 7, 8

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[15] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free. *ICLR*, 2017. 6

[16] 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. 7

[17] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H. S. Torr. Learn to pay attention. In *International Conference on Learning Representations*, 2018. 1, 3

[18] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016. 6

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

[20] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. *CVPR*, 2018. 1, 3

[21] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. 6

[22] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. In *ICCV*, 2017. 8

[23] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *ICLR*, 2017. 7

[24] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 2

[25] Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *European Conference on Computer Vision*, pages 120–135. Springer, 2016. 2

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 3

[27] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 3

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 5

[29] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 8

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2, 5, 6

[31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-CAM: Visual explanations from deep networks

via gradient-based localization. In *ICCV*, pages 618–626, 2017. 1, 2, 3, 4, 5, 6, 8

[32] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshop*, 2014. 2, 6

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1, 2, 3, 4, 7

[34] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations Workshop*, 2015. 2, 4

[35] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, 2018. 6, 8

[36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5, 7

[37] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision*, pages 2399–2406, 2015. 8

[38] 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*, pages 3156–3164, 2017. 3

[39] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE CVPR*, volume 1, page 3, 2017. 3

[40] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018. 3

[41] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2, 4

[42] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1143–1152, 2016. 8

[43] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE CVPR*, 2018. 3

[44] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Int. Conf. on Computer Vision*, volume 6, 2017. 8

[45] 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. 1, 2, 3, 4, 5