DANet: Divergent Activation for Weakly Supervised Object Localization

Haolan Xue[†], Chang Liu[†], Fang Wan[†*], Jianbin Jiao[†], Xiangyang Ji[‡] and Qixiang Ye[†♯*]

[†]University of Chinese Academy of Sciences, Beijing, China
[‡]Tsinghua University, Beijing, China. [♯]Peng Cheng Laboratory, Shenzhen, China
{xuehaolan17,liuchang615,wanfang13}@mails.ucas.ac.cn, xyji@tsinghua.edu.cn
{jiaojb,qxye}@ucas.ac.cn,

Abstract

Weakly supervised object localization remains a challenge when learning object localization models from image category labels. Optimizing image classification tends to activate object parts and ignore the full object extent, while expanding object parts into full object extent could deteriorate the performance of image classification. In this paper, we propose a divergent activation (DA) approach, and target at learning complementary and discriminative visual patterns for image classification and weakly supervised object localization from the perspective of discrepancy. To this end, we design hierarchical divergent activation (HDA), which leverages the semantic discrepancy to spread feature activation, implicitly. We also propose discrepant divergent activation (DDA), which pursues object extent by learning mutually exclusive visual patterns, explicitly. Deep networks implemented with HDA and DDA, referred to as DANets, diverge and fuse discrepant yet discriminative features for image classification and object localization in an end-to-end manner. Experiments validate that DANets advance the performance of object localization while maintaining high performance of image classification on CUB-200 and ILSVRC datasets [1].

1. Introduction

Weakly supervised learning refers to methods that utilize training data with incomplete annotations to learn recognition models. Weakly supervised object localization (WSOL) requires solely the image-level annotations indicating the presence or absence of a class of objects in images to learn localization models [39]. It can leverage rich Web images with tags as a data source for model learning.

To tackle the WSOL problem with convolutional neural network (CNN), people resort to the discriminative localization method [39], *i.e.*, learning class activation maps for object localization using excitation back-propagation from image category supervision[36]. In the forward-propagation procedure the convolutional filters in CNN act as object detectors and in the back-propagation procedure the feature maps are excited to produce class activation maps, which identify discriminative regions for specific object classes.

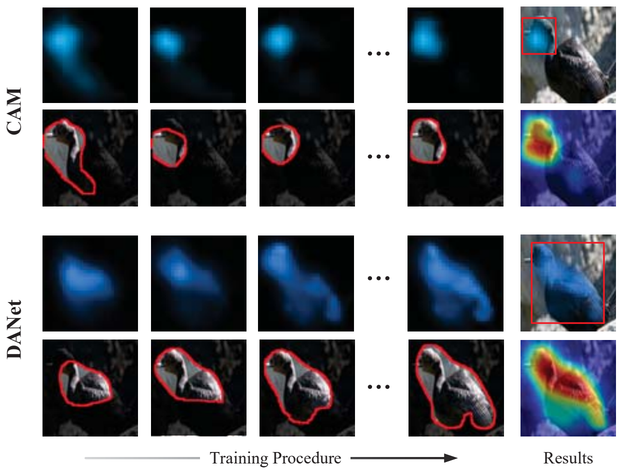


Figure 1: Evolution of the activation maps during training. In the early stages of training, both the CAM [39] and our DANet activate partial objects. Along with the learning procedure, the activated region of CAM shrinks to a small object part while that of our approach diverges to full object extent. (Best viewed in color)

Discriminative localization methods are simple yet efficient for weakly supervised object localization. However, they are usually observed to activate object parts instead of full object extent, as shown in the first row of Fig. 1. Specific activated object parts are capable of minimizing image classification loss, but experience difficultly in optimizing object localization. Existing approach has explored graph

*Corresponding author.
[1]The code is available at https://github.com/xuehaolan/DANet

propagation [40], data argumentation [13], dilated convolution [31], and adversarial erasing [37, 12] to expand class activation maps and pursue full object extent. Nevertheless, most exist approaches address the problem in the way of step-wised or alternative optimization. Theoretically plausible frameworks for localizing full object extent under the constraint of image classification performance remain to be explored.

In this paper, we propose a divergent activation (DA) approach, and target at learning complementary and discriminative visual patterns from the perspective of discrepancy. To this end, we design hierarchical divergent activation (HDA) and discrepant divergent activation (DDA). The HDA is inspired by the image category structure, $i.e.$, images from different categories can be merged by their similarity and assigned with hierarchical class labels. Training classification models with hierarchical class labels can effectively expand visual patterns and provide extra guidance to discriminative localization. The DDA is based on the complementary spatial structure, $i.e.$, an object could be decomposed into spatially exclusive visual patterns. Activating and fusing such visual patterns during training facilitate localizing full object extent, Fig. 1.

Deep networks implemented with DA, referred to as DANets, incorporate image classification and weakly supervised object localization with a joint optimization objective (loss) function. With an end-to-end learning procedure optimizing the objective function, DANets discover complementary and discriminative visual patterns for precise object localization while maintaining the high performance of image classification.

The contributions of this paper include:

(1) We propose a divergent activation (DA) method, and jointly optimize weakly supervised object localization and image classification in a systematic way.

(2) We design hierarchical divergent activation (HDA) and discrepant divergent activation (DDA) modules, and leverage semantic discrepancy and spatial discrepancy to learn complementary and discriminative visual patterns.

(3) We update popular deep neural networks including VGG16 and GoogLeNet to DANets and advance the performance about weakly supervised object localization.

2. Related Work

Multiple instance learning (MIL) and discriminaitve localization are major WSOL methods. With the MIL method, an image is first decomposed into region proposals, based on which proposal selection and classifier estimation are iteratively performed [29, 35, 4, 1, 23, 27]. With the discriminaitve localization method, deep pixels are activated with excitation back-propagation to cover objects of interest under the supervision of image class labels [40, 15, 8, 38, 3].

2.1. Weakly Supervised Object Localization

Step-wised multiple instance learning. A major WSOL approach is decomposing an image into a "bag" of region proposals (instances) and iteratively selecting high-scored instances from each bag when learning detectors in step-wised manner [4]. MIL has been updated to MIL networks [1] where the convolutional filters behave as detectors to activate regions of interest on the feature maps [27]. Recent approaches have used image segmentation [7], context information [11], online classifier refinement [23], and min-entropy [27, 28] to regularize the MIL procedure. Progressive optimization [35] and clique partition [27] have been explored to enhance object localization.

Benefit from the location prior of region proposals, the step-wised MIL methods are effective to localize object extent. However, they are puzzled by the time-consuming proposal generation procedure. The WeakRPN [24] approach takes a step towards learning region proposal networks, but remains relying on region proposals in the training phrase.

End-to-end discriminative localization. Discriminative localization excites object extent in an end-to-end manner by introducing a global average pooling (GAP) module into the classification network [39]. With the GAP module, convolutional filters behave as detectors to activate discriminative regions on feature maps to localize objects. However, most discriminative localization approaches are observed to activate object parts instead of full object extent. The reason behind the phenomenon lies in that the networks tend to learn the most compact features for image classification while suppressing less discriminative ones [20].

One way to enhance object localization is self-paced learning [38, 10]. For example, the self-produced guidance (SPG) approach uses a classification network to learn high confident regions, and then leverages attention maps to learn the object extent under the guidance of high confident regions. The other way to pursue full object extent is about adversarial erasing and hide-and-seek[12, 15, 8, 13, 37], which first activates the most discriminative regions and then erases them so that less discriminative regions can be activated. Although [37] uses end-to-end learning, it remains a step-wised processing strategy in each training iteration. In this paper, we propose a divergent activation approach, where the discrepant feature maps can be simultaneously activated.

The self-paced and adversarial erasing approaches work a progressive manner, $i.e.$, discovering and fusing discriminative regions. Although practically plausible, they are theoretically sub-optimal as working in a way of heuristic search. The soft proposal network [40] integrates confidence propagation with discriminative localization in an end-to-end manner, but remain falling into progressive optimization instead of joint optimization.

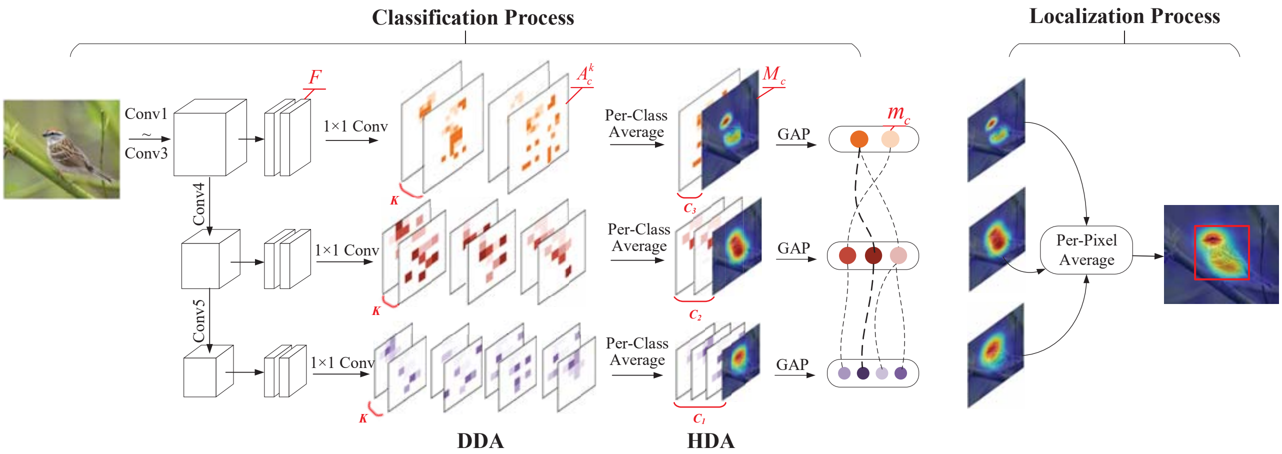


Figure 2: Architecture of the proposed DANet, which leverages the hierarchical features and the hierarchical image class supervision to implement the hierarchical divergent activation (HDA) during classification. It also implements discrepant divergent activation (DDA) by maximizing the spatial discrepancy of feature maps. During the localization process, the discrepant yet complementary visual patterns are fused to diverge object parts to full object extent. (Best viewed in color)

2.2. Classification with Category Hierarchy

Our research is also related to category hierarchy of images, which has been exploited for fine-grained image recognition [34, 32, 14, 30, 2]. The main idea lies in that the discriminative visual patterns of parent classes are different from those of sub-classes. This implies that the activated regions of objects could be expanded if multiple sub-classes are merged.

This inspires us designing hierarchical divergent divergence and leveraging the semantic discrepancy to spread visual patterns for object localization. To our best knowledge, this is the first time that the category hierarchy is explored for WSOL.

3. Divergent Activation Network

In this section, we first review and reformulate the discriminative method for WSOL. We then propose the divergent activation (DA) method and incorporate it with discriminative localization in a joint optimization framework.

3.1. Class Activation Map Revisit

We use a discriminative localization model in [37] to extract class activation maps from classification networks. The network is first converted to a fully convolutional network by removing the global pooling layer and transforming the weights of the fully connected layers to 1×1 convolutional filters, Fig. 2.

Let $F \in \mathcal{R}^{P \times P \times N}$ denote the feature maps of CNN, where P defines the resolution of the feature maps and N the channel number. Let $W_c^k \in \mathcal{R}^{1 \times 1 \times N}$ denote the 1×1 convolutional filters, where $c = 1, \cdots C$ denote the class index and $k = 1, \cdots K$ denote the feature map index. The k^{th} activation map, A_c^k, for class c is computed as $A_c^k = F * W_c^k$. The activation maps are then summarized to produce a single class activation map, $M_c = \sum_k A_c^k$.

The class activation maps for all classes are then fed to a global average pooling (GAP) layer to produce logits, $m_c = \frac{\sum_{i,j} M_c(i,j)}{P \times P}$, where (i,j) denotes the spatial location on the activation map . A softmax operation is applied to produce classification results. The output of the softmax layer for class c, p_c, is given by $\frac{exp(m_c)}{\sum_c exp(m_c)}$, and the classification loss function is defined as

$$\arg\min_{\alpha} \mathcal{L}(\alpha), \tag{1}$$

where $\mathcal{L}(\alpha) = -\frac{1}{C}\sum_c y_c log(p_c)$. $y_c \in \{0,1\}$ denotes the label for class c and α the network parameters.

The class activation maps produced by the image classification network are observed shrinking to small object parts, Fig. 1. This phenomenon is attributed to the intrinsic compact nature of the convoltuional features. With the solely objective (loss) function to optimize image classification, the only goal of learning is to capture and represent the relevant visual patterns between input images and object category label y [25]. Since the category label y implicitly determines the relevant and irrelevant features in F, an optimal representation of image would capture the relevant features and compress F by suppressing the irrelevant visual patterns which do not contribute to the prediction of y. Considering the corresponding relationship between feature maps F and the class activation map M defined above, a compressed F produces sparse class activation map M, which indicates the spatial locations of objects from class c.

3.2. Divergent Activation

To expand the compressed features and explore richer visual patterns for object localization, we propose divergent activation (DA) and integrated it with an image classification network. The divergent activation is fulfilled from the perspective of discrepancy learning, and is deployed as hierarchical divergent activation (HDA) and discrepant divergent activation (DDA) modules. The learning procedure is fulfilled by optimizing a joint objective function, as

$$\arg\min_{\alpha}\{\mathcal{L}_H(\alpha)+\lambda\mathcal{L}_D(\alpha)\}, \tag{2}$$

where $\mathcal{L}_H(\alpha)$ denotes the hierarchical classification loss and $\mathcal{L}_D(\alpha)$ the divergent activation loss. λ is the regularization factor.

Hierarchical divergent activation (HDA). For image classification, CNNs learn to discriminate an image class from the others by activating the discriminative visual patterns. Meanwhile, the similar visual patterns between classes are suppressed, as shown by each network branch in Fig. 2. To localize full object extent, the key lies in how to activate the suppressed visual patterns.

It is a common sense that for two classes which are semantically similar, *e.g.*, "dog" and "wolf", there exist many similar visual patterns (object parts). If we merge the similar (child) classes into a parent class and train a classifier for the parent class, *e.g.*, a "dog+wolf" class, those similar visual patterns shared by the child classes are activated if they are discriminative to other parent classes. Recursively, regarding the parent classes as new child classes and merging them to obtain a new parent class, more visual patterns are further activated.

Based on above analysis, we propose hierarchical divergent activation (HDA) to activate the similar regions among classes. Given an image dataset containing C^h classes of objects, *e.g.*, 200 classes of birds in CUB-200-2011 [26], we first merge them into C^{h+1} parent classes based on the semantic similarity among the child classes, and then merge the C^{h+1} classes into C^{h+2} classes, where $C^{h+2} < C^{h+1} < C^h$. On the hierarchical classes, the loss function of HDA is defined as

$$\arg\min_{\alpha}\mathcal{L}_H(\alpha)=\sum_h \mathcal{L}_h(\alpha)=-\sum_h \frac{1}{C^h}\sum_c y_c^h log(p_c^h), \tag{3}$$

where $\mathcal{L}_h$ is the loss of the h^{th} class hierarchy. y_c^h is the label of the c^{th} class where $c \in C^h$ and C^h is the number of the classes in h^{th} class hierarchy.

The essence of HDA lies in that by hierarchically changing the discriminative conditions using child-parent classes, more informative visual patterns are collected and the activation maps diverge from small object parts to full object extent.

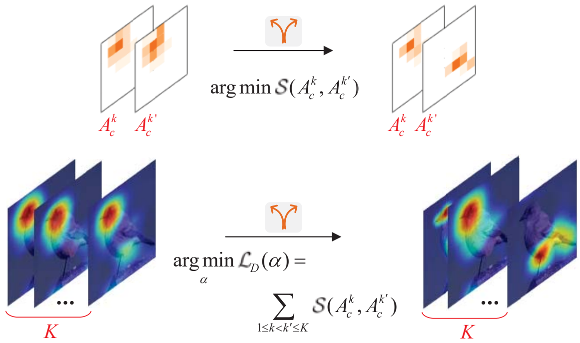


Figure 3: Discrepant divergent activation (DDA) leverages spatial discrepancy of feature maps to learn visual patterns suppressed by image classification.

Discrepant divergent activation (DDA). HDA tends to activate full object extent by fusing complementary semantics from multiple hierarchy levels, but does not consider the spatial complementary of activation maps for a single hierarchy level of objects. We thus further propose the discrepant divergent activation (DDA) to aggregate visual patterns, Fig. 3.

To fulfill this purpose, a single class activation map is first expanded to K activation maps. Specifically for the c^{th} class, we introduce the DDA loss so that the K activation maps are discrepant, as much as possible, with each other. This is equivalent to minimize the similarity among activation maps, A_c , as

$$\arg\min_{\alpha}\mathcal{L}_D(\alpha)=\sum_{1\le k<k'\le K}\mathcal{S}(A_c^k, A_c^{k'}), \tag{4}$$

where A_c^k denotes the k^{th} activation map for the c^{th} class. $\mathcal{S}(A_c^k, A_c^{k'}) = \frac{A_c^k \cdot A_c^{k'}}{\|A_c^k\| \cdot \|A_c^{k'}\|}$ is the cosine similarity between activation map A_c^k and $A_c^{k'}$.

Once Eq. 4 is optimized, the activation maps of class c are most discrepant to each other. If an activation map discovers one object part, the other maps will be forced to activate other spatially exclusive parts. It means that the visual patterns discovered by each two activation maps are different with each other and the activated regions on the maps are complementary.

3.3. Network Implementation

Fully convolutional neural networks implemented DA modules, referred to as DANets, activate and fuse complementary discriminative regions for precise object localization and accurate image classification in an end-to-end manner, Fig. 2. Given a network, multiple scales of feature maps (*i.e.*, convolutional maps of CONV3, CONV4 and CONV5

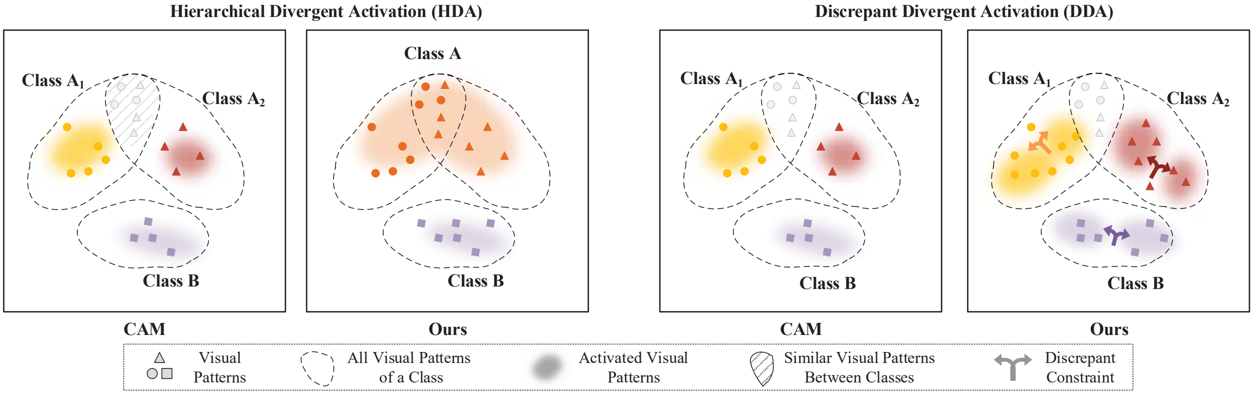


Figure 4: Explanation of the proposed hierarchical divergent activation (HDA) and discrepant divergent activation (DDA). With the HDA module, the parent class (A) can learn more visual patterns to span the feature space. Such visual patterns are suppressed by each child class (A1 or A2) as they are not discriminate against A1 and A2. With the DDA module, the visual patterns learned by each class (A1 or A2) are aggregated. This is because the discrepancy constraint drives learning different but discriminative features for image classification. (Best viewed in color)

in VGG-16) are first extracted to represent hierarchical image categories. Atop the feature maps from each hierarchy, a 1×1 convolutional layer is added to produce K activation maps for each class. The activation maps are then fed to the HDA and DDA modules.

For each hierarchy in HDA, the K activation maps for each class are averaged to generate the activation map of this hierarchy. A global average pooling layer is then used to generate logits, which is followed by the HDA loss defined in Eq. 3 for image classification. In DDA, the activation maps from the same class are first concatenated and the DDA loss, defined in Eq. 4, which minimizes the similarities among the maps is added.

In the training phase, the HDA and DDA modules are jointly optimized[16] with SGD algorithm. In the testing phase, the output classification prediction which comes from the last hierarchy is used to predict the class of an image, Fig. 2. The maps from all hierarchy levels are averaged to form the final activation results and a thresholding approach [38] is then applied to predict the object locations.

3.4. Discussion

From the perspective of representation learning, DANets span the feature space by aggregating visual patterns. As shown in Fig. 4, with the HDA module, the discriminative visual patterns learned by each class (A1 or A2) are united. The parent class (A) can learn visual patterns to span the feature space. Such visual patterns are ignored by a child class (A1 or A2) as they are not discriminative to other child classes. With the DDA module, the discriminative visual patterns learned by each class (A1 or A2) are enriched, as the discrepancy constraint drives learning different but discriminative feature maps for image classification. DANets therefore enhances the representative capacity of features for image classification and object localization, which provides the WSOL problem with a fresh insight.

From the perspective of ensemble learning, DANets actually assemble multiple discrepant learners. Regarding each activation map as a learner for image classification and object localization, the HDA module implements a hierarchical ensemble in the semantic space, while the DDA module implements paralleled ensemble in the feature space. Classical machine learning research suggests that learners to be assembled should "disagree" with each other, as much as possible [19]. The discrepancy incorporated in the HDA and DDA modules therefore shows the general sense to design and assemble learners in deep neural networks.

4. Experiments

4.1. Experimental Setup

Datasets. DANet is evaluated on the commonly used CUB-200-2011 [26] and ILSVRC 2016 [5, 18] datasets. CUB-200-2011 contains 11,788 images of 200 bird species with 5,994 images for training and 5,794 for test. Following the biological taxonomy we divide the 200 species of birds into a three-level hierarchy, which includes 37 families, and 11 orders. For ILSVRC 2016, we use 1.2 million images with 1,000 classes for training, and 5,000 images in the validation set for testing. We apply the off-the-shelf category hierarchy of ILSVRC 2016 dataset from WordNet [17], a language database which structures concepts and how they relate. These hierarchical class labels are obtained from knowledge graphs with taxonomic hierarchy. As for other datasets, a related hierarchy can also be structured from WordNet.

Evaluation metrics. Two metrics are used for WSOL performance evaluation. The first localization metric is suggested by [18]: fraction of images with right prediction of classification of the image labels and 50% IoU with the ground-truth box. The second is the Correct Localization (CorLoc) rate [6], which indicates the localization performance given the class label for each test image.

Experimental details. The proposed DA modules are integrated with the commonly used CNNs including VGGnet [21] and GoogLeNet [22]. Following the settings of previous work [38], we remove the layers after conv5-3 (from pool5 to prob) of the VGG-16 network and the last inception block of GoogLeNet. We then add two convolutional layers with kernel size 3×3, stride 1, pad 1 with 1024 units, and a convolutional layer of size 1×1, stride 1 with 1000 units (200 units for CUB-200-2011). As illustrated in Fig. 2, discrepant activation maps can be conveniently obtained from the feature maps before the GAP layer. Both networks are fine-tuned on the pre-trained weights of ILSVRC [18]. The input images are randomly cropped to 224×224 pixels after being re-sized to 256×256 pixels. For classification, we average the scores from the softmax layer with 10 crops.

4.2. Ablation Studies

The ablation studies on CUB-200-2011 using VGGnet are used to evaluate the effects of the proposed DA modules.

Effect of HDA. As shown in Table 1, HDA reduces the top-1/top-5 *loc. err.* by 5.14%/4.36% compared with the baseline CAM approach, at the cost of little ($\sim$ 1%) classification performance. In Fig. 5, examples of activation maps show the impact of the HDA module. With only the supervision from child class labels, CAM tends to activate object parts, $e.g.$, the bird head. With the introduced image category hierarchy supervisions, the activation maps enrich common visual patterns belonging to the same parent class of birds. For example, the slim body and similar feather color of family *Warbler* is activated by the HDA module, and the activation regions diverge from bird head to bird body. We also do ablation study on number of hierarchy levels with limited hierarchy levels provided by biological taxonomy and obtained 55.85%, 52.80%, and 50.71% loc. err. with one, two and three levels, respectively. It can be seen that loc. err. decreased when more hierarchy supervisions are introduced.

In Table 1,"CAM+multi-loss" refers to applying the same supervision to the feature pyramid of the network in Fig. 2 without using the DA module. It can be seen that both the *cls. err.* and *loc. err.* of "CAM+multi-loss" are worse than that of the baseline CAM approach. This shows that simply updating the backbone network of CAM to a feature pyramid network does not necessarily boost the performance of WSOL. The reason lies in that without DA modules the CAM on the feature pyramid fails activating complementary visual patterns.

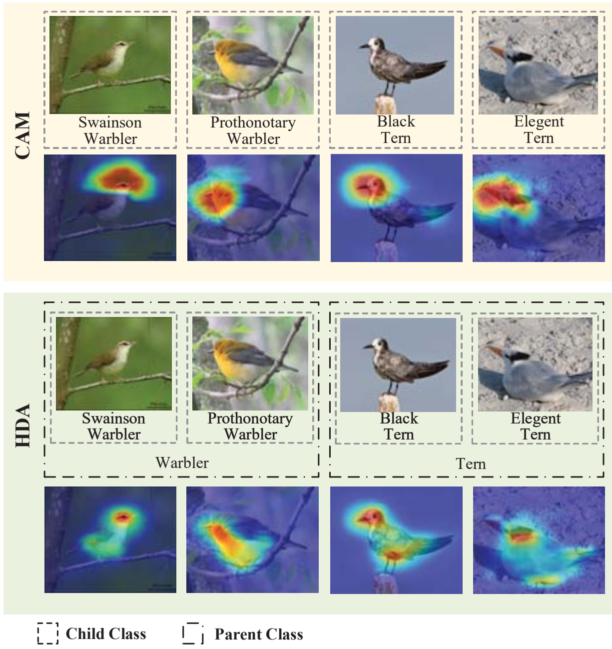
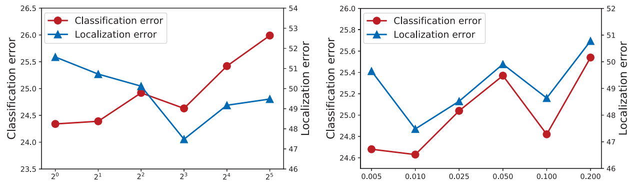


Figure 5: Examples of HDA Maps on CUB-200-2011. The first two rows are supervision and activation maps of CAM and the last two rows are ours. Different frames indicate different labels. With solely child labels provided, CAM focuses only on the most discriminative parts $i.e.$, bird head, while the proposed HDA approach diverges towards full object extent.

	cls. err		loc. err	
Method	top1	top5	top1	top5
CAM [39]	**23.42**	7.47	55.85	47.84
CAM+multi-loss	24.99	8.11	58.58	49.97
HDA	24.13	**6.96**	50.71	43.48
HDA+DDA	24.63	7.73	**47.48**	**38.04**

Table 1: The effect of the proposed hierarchical divergent activation (HDA) and discrepant divergent activation (DDA). Comparing with the baseline CAM approach, DA modules achieve 8.37%/9.80% localization performance gain at the cost of 1.21%/0.26% classification performance. Lower digits indicate better performance.

Effect of DDA. In Fig. 6a, we evaluate the *loc. err.* under different numbers (K) of discrepant activation maps and provides a reference for the selection of K. With too few discrepant maps, it is difficult to produce sufficient spatial discrepancy. With too many discrepant activation maps, the parameters increase significantly, which increases the risk



(a) K (b) λ

Figure 6: Evaluation of DDA parameters, *i.e.*, activation map number K and regularization factor λ, on CUB-200-2011.

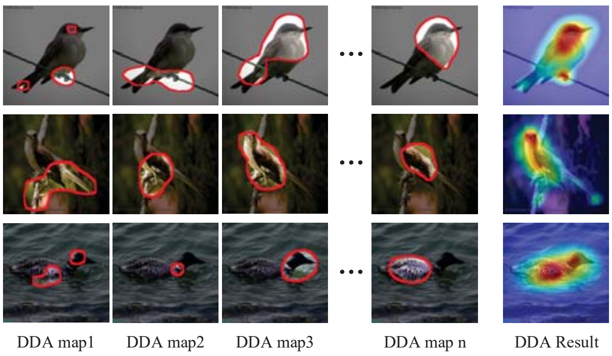


Figure 7: Discrepant activation maps on the cub-200-2011 test set. Discrepant visual patterns (1^{st} to 4^{th} columns) are fused to cover the full object extent (last column). (Best view in color)

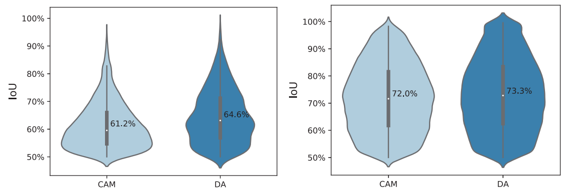
of over-fitting. To alleviate the difficulty of learning additional parameters, we randomly dropout half of the discrepant activation maps in each training mini-batch, which are validated to achieve higher performance and faster network convergence.

In Fig. 6b, we evaluate the regularization factor λ (defined in Eq. 2) and observed that $K = 8$, $\lambda = 0.01$ reports the best performance. With proper parameters, complementary visual patterns are discovered in discrepant activation maps, a combination of these activation maps covers the full object extent, as shown in Fig. 7 and Fig. 9.

Statistical analysis. In Fig. 8, we show the statistical analysis of "correct bounding boxes" which indicates correct classification with over 50% IoU with the ground-truth boxes on CUB and ILSVRC datasets. It can be seen that the proposed DANet enhances the quality of correct bounding boxes on both datasets by improving the IoU rates.

4.3. Comparison with the state-of-the-arts

We compare the proposed DANets with the state-of-the-art approaches on the CUB-200-2011 test set and ILSVRC validation set and report the results in Table 2, Table 3, and Table 4, respectively.



(a) CUB-200-2011 (b) ILSVRC

Figure 8: Statistical analysis of "correct bounding boxes".

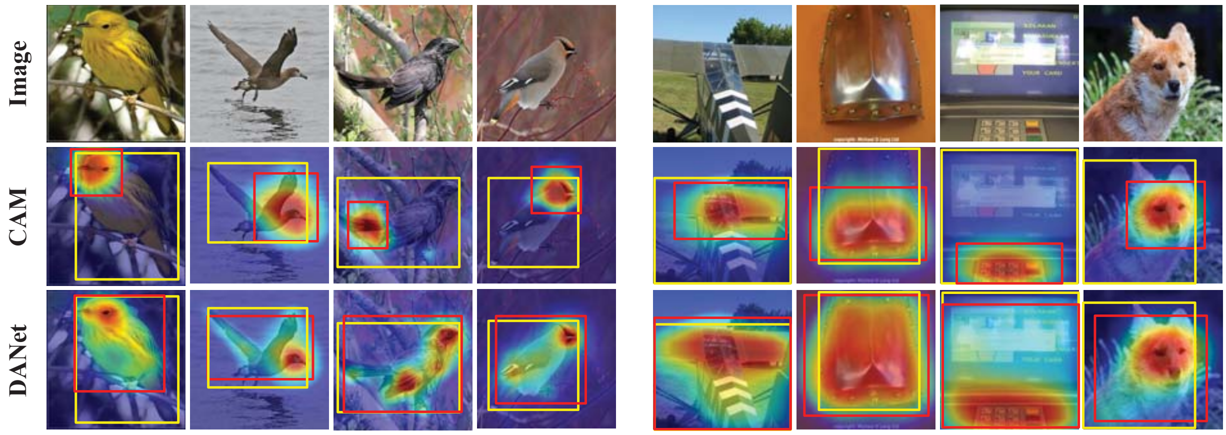
	cls. err.		loc. err.	
Method	top1	top5	top1	top5
GoogLeNet-CAM [39]	**26.2**	**8.5**	58.94	49.34
GoogLeNet-SPG [38]	-	-	53.36	42.28
GoogLeNet-DANet (ours)	28.8	9.4	**50.55**	**39.54**
VGGnet-CAM [39]	**23.4**	**7.5**	55.85	47.84
VGGnet-ACoL [37]	28.1	-	54.08	43.49
VGGnet-SPG [38]	24.5	7.9	51.07	42.15
VGGnet-DANet (ours)	24.6	7.7	**47.48**	**38.04**

Table 2: Performance comparison on the CUB-200-2011 test set. DANets achieve significant localization performance gain over the state-of-the-arts while reporting comparable image classification performance.

On the CUB-200-2011 test set, with a VGGnet backbone, DANet reports $6.60\%/5.45\%$ lower top-1/top-5 *loc. err.* and 3.5% lower top-1 *cls. err.* compared with the adversarial erasing approach (ACoL) approach [37] at the cost of little classification performance. It reports $3.59\%/4.11\%$ lower top-1/top-5 localization error compared with the self-produced guidance (SPG) approach [38] at the cost of 0.1%-0.2% classification performance. With a GoogleLeNet backbone, it reports $2.81\%/2.74\%$ performance gain over the state-of-the-art SPG approach [38]. We also implemented DANet with ResNet-50 and obtained: 18.4% cls. err. and 38.9% loc. err., demonstrating the advantage of DANet with high capacity networks.

On the large-scale ILSVRC dataset, it can be seen that the DANet with a GoogLeNet backbone, simultaneously improves the classification and localization performance comparing with the state-of-the-art ACoL approach [37]. It also reports comparable performance with the state-of-the-art SPG [38] approach. This validates the priority of the proposed joint optimization framework over the stepwised optimization method employed in the compared approaches.

In Table 4, we evaluate the CorLoc performance on the CUB-200-2011 test set. By removing disturbance the from image classification, this metric can explicitly reflect the localization performance. It can be seen that DANet with a



(a) CUB-200-2011

(b) ILSVRC

Figure 9: Comparison with the CAM [39] method. Our method can locate larger object regions to improve localization performance (ground-truth bounding boxes are in yellow and the predicted are in red).

	cls. err.		loc. err.	
Method	top1	top5	top1	top5
VGGnet-Backprop [21]	-	-	61.12	51.46
VGGnet-CAM [39]	33.4	12.2	57.20	45.14
VGGnet-ACoL [37]	32.5	12.0	54.17	40.57
GoogLeNet-Backprop [21]	-	-	61.31	50.55
GoogLeNet	31.9	11.3	60.09	49.34
GoogLeNet-GMP [39]	35.6	13.9	57.78	45.26
GoogLeNet-CAM [39]	35.0	13.2	56.40	43.00
GoogLeNet-HaS-32 [13]	-	-	54.53	-
GoogLeNet-ACoL [37]	29.0	11.8	53.28	42.58
GoogLeNet-SPG [38]	-	-	**51.40**	**40.00**
GoogLeNet-DANet (ours)	**27.5**	**8.6**	52.47	41.72

Table 3: Performance comparison on the large-scale ILSVRC validation set. DANets improve both object localization and image classification performance over the state-of-the-art adversarial erasing approach (ACoL).

Method	CorLoc
GoogLeNet-CAM [39]	55.1
GoogLeNet-Friend or Foe[33]	56.51
GoogLeNet-DANet (ours)	67.03
VGGnet-ACoL [37]	54.1
VGGnet-CAM [39]	56.0
VGGnet-SPG [38]	58.9
VGGnet-TSC [9]	65.5
VGGnet-DANet (ours)	**67.7**

Table 4: CorLoc rate on the CUB-200-2011 test set. Larger number indicates better performance.

VGGnet backbone respectively outperforms ACoL [37] and SPG [38] up to 13.6% (67.7% vs. 54.1%) and 8.8% (67.7% vs. 58.9%). It also outperforms the other state-of-the-art approaches with significant margins.

5. Conclusion

In this paper, we proposed a simple yet effective divergent activation (DA) approach for weakly supervised object localization. We designed hierarchical divergent activation (HDA) and discrepant divergent activation (DDA) modules and unified them with the deep learning framework, leading to DANets. We also defined a joint objective function so that the DA loss can be simultaneously optimized with the image classification loss. During learning, DANets diverge object parts into full object extent and significantly improve the performance of weakly supervised object localization while maintaining the high performance of image classification. The underlying reality lies in that the DA modules span the feature space by learning complementary visual patterns while DANets implement a special kind of learner ensemble by maximizing the discrepancy between learners. This provides fresh insights to the challenging weakly supervised learning problem.

Acknowledgments

The authors are very grateful to the support by NSFC grant 61836012, 61771447, and 61671427, and Beijing Municipal Science and Technology Commission grant Z181100008918014.

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2846–2854, 2016.

[2] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proc. ACM Multimedia Conf. (ACM MM)*, pages 2023–2031, 2018.

[3] Junsuk Choe, Joo Hyun Park, and Hyunjung Shim. Improved techniques for weakly-supervised object localization. *arXiv preprint arXiv:1802.07888*, 2018.

[4] Ramazan Gokberk Cinbis, Jakob J. Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(1):189–203, 2017.

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

[6] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100(3):275–293, 2012.

[7] Ali Diba, Vivek Sharma, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 914–922, 2017.

[8] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 642–651, 2017.

[9] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *AAAI Conf. Arti. Intell. (AAAI)*, pages 4075–4081, 2017.

[10] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Adv. in Neural Inf. Process. Syst. (NIPS)*, pages 547–557, 2018.

[11] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 350–365, 2016.

[12] Dahun Kim, Donghyeon Cho, and Donggeun Yoo. Two-phase learning for weakly supervised object localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3534–3543, 2017.

[13] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3524–3533, 2017.

[14] Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, and Honglak Lee. Hierarchical novelty detection for visual object recognition. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1034–1042, 2018.

[15] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9215–9223, 2018.

[16] Shaohui Lin, Rongrong Ji, Chao Chen, Dacheng Tao, and Jiebo Luo. Holistic cnn compression via low-rank decomposition with knowledge transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[17] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[18] 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*, 115(3):211–252, 2015.

[19] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3723–3732, 2018.

[20] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *CoRR*, abs/1703.00810, 2017.

[21] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1–9, 2015.

[23] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2843–2851, 2017.

[24] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan L. Yuille. Weakly supervised region proposal network and object detection. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 352–368, 2018.

[25] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

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

[27] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1297–1306, 2018.

[28] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, DOI:10.1109/TPAMI.2019.2898858, 2019.

[29] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category

learning. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 431–445, 2014.

[30] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 2399–2406, 2015.

[31] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7268–7277, 2018.

[32] Saining Xie, Tianbao Yang, Xiaoyu Wang, and Yuanqing Lin. Hyper-class augmented and regularized deep learning for fine-grained image classification. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2645–2654, 2015.

[33] Zhe Xu, Dacheng Tao, Shaoli Huang, and Ya Zhang. Friend or foe: Fine-grained categorization with weak supervision. *IEEE Transactions on Image Processing*, 26(1):135–146, 2017.

[34] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 2740–2748, 2015.

[35] Qixiang Ye, Tianliang Zhang, Wei Ke, Qiang Qiu, Jie Chen, Guillermo Sapiro, and Baochang Zhang. Self-learning scene-specific pedestrian detectors using a progressive latent model. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 509–518, 2017.

[36] Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 1084–1102, 2016.

[37] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1325–1334, 2018.

[38] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pages 597–613, 2018.

[39] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2921–2929, 2016.

[40] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1841–1850, 2017.