WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation

Thibaut Durand$^{(1)\star}$, Taylor Mordan$^{(1,2)\star}$, Nicolas Thome$^{(3)}$, Matthieu Cord$^{(1)}$
(1) Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu, 75005 Paris
(2) Thales Optronique S.A.S., 2 Avenue Gay Lussac, 78990 Élancourt, France
(3) CEDRIC - Conservatoire National des Arts et Métiers, 292 rue St Martin, 75003 Paris, France
{thibaut.durand, taylor.mordan, nicolas.thome, matthieu.cord}@lip6.fr

Abstract

This paper introduces WILDCAT, a deep learning method which jointly aims at aligning image regions for gaining spatial invariance and learning strongly localized features. Our model is trained using only global image labels and is devoted to three main visual recognition tasks: image classification, weakly supervised pointwise object localization and semantic segmentation. WILDCAT extends state-of-the-art Convolutional Neural Networks at three major levels: the use of Fully Convolutional Networks for maintaining spatial resolution, the explicit design in the network of local features related to different class modalities, and a new way to pool these features to provide a global image prediction required for weakly supervised training. Extensive experiments show that our model significantly outperforms the state-of-the-art methods.

1. Introduction

Over the last few years, deep learning and Convolutional Neural Networks (CNNs) have become state-of-the-art methods for visual recognition, including image classification [34, 56, 28], object detection [21, 20, 10] or semantic segmentation [8, 42, 9]. CNNs often require a huge number of training examples: a common practice is to use models pre-trained on large scale datasets, *e.g.* ImageNet [53], and to fine tune them on the target domain.

Regarding spatial information, there is however a large shift between ImageNet, which essentially contains centered objects, and other common datasets, *e.g.* VOC or MS COCO, containing several objects and strong scale and translation variations. To optimally perform domain adaptation in this context, it becomes necessary to align informative image regions, *e.g.* by detecting objects [44, 29], parts [68, 69, 70, 35] or context [23, 13]. Although some works incorporate more precise annotations during training, *e.g.* bounding boxes [43, 21], the increased annotation cost prevents its widespread use, especially for large datasets and pixel-wise labeling, *i.e.* segmentation masks [3].

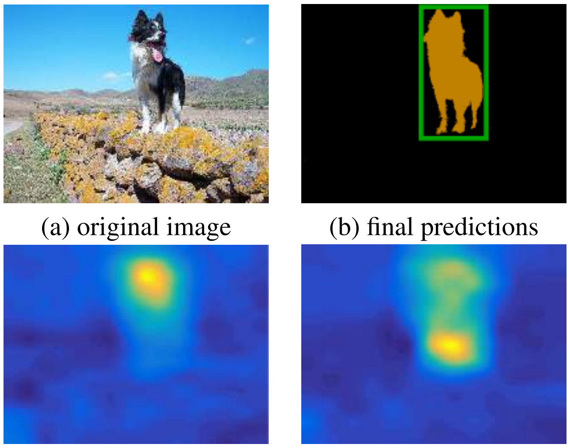
(a) original image (b) final predictions
(c) *dog* heatmap 1 (head) (d) *dog* heatmap 2 (legs)

Figure 1. WILDCAT example performing localization and segmentation (b), based on different class-specific modalities, here head (c) and legs (d) for the *dog* class.

In this paper, we propose WILDCAT (Weakly supervIsed Learning of Deep Convolutional neurAl neTworks), a method to learn localized visual features related to class modalities, *e.g.* heads or legs for a dog – see Figure 1(c) and 1(d). The proposed model can be used to perform image classification as well as weakly supervised pointwise object localization and segmentation (Figure 1(b)).

The overall architecture of WILDCAT (Figure 2) improves existing deep Weakly Supervised Learning (WSL) models at three major levels. Firstly, we make use of the latest Fully Convolutional Networks (FCNs) as back-end module, *e.g.* ResNet [28] (left of Figure 2). FCNs have recently shown outstanding preformances for fully super-

⋆ equal contribution

This research was supported by a DGA-MRIS scholarship.

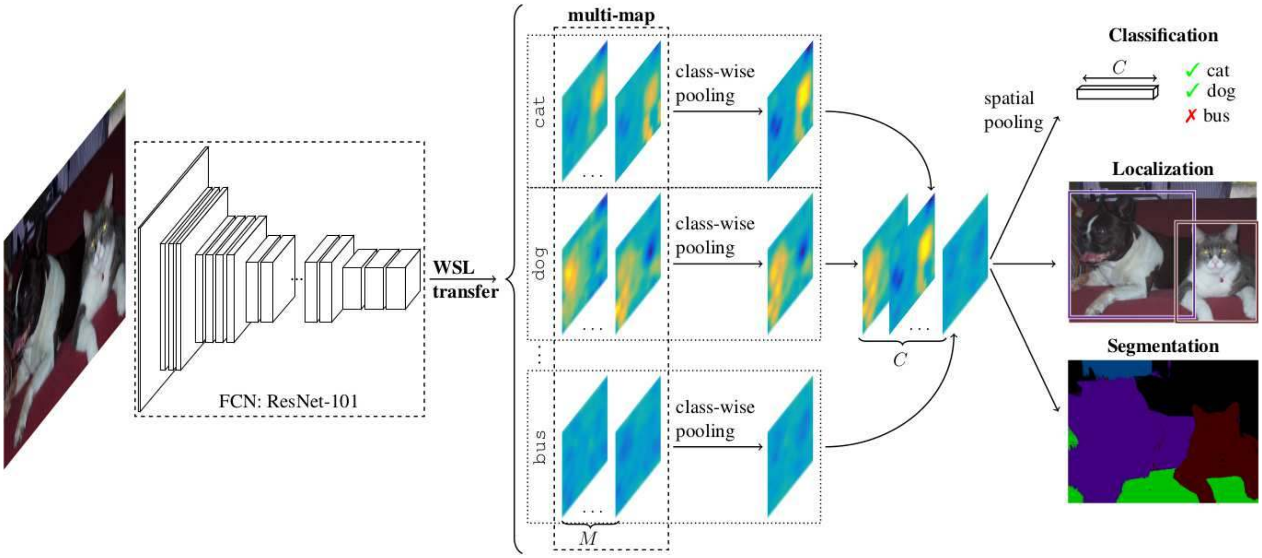

Figure 2. WILDCAT architecture. It is based on FCN ResNet-101 to extract local features from whole images with good spatial resolution (Section 3.1). All regions are encoded into multiple class modalities with a WSL multi-map transfer layer (Section 3.2). Feature maps are then combined separately to yield class-specific heatmaps that can be globally pooled to get a single probability for each class, using a new spatial aggregation module (Section 3.3). WILDCAT is trained with image-level labels in a WSL way and is applied to complex scene understanding, WSL object detection and semantic segmentation (Section 3.4).

vised object detection [10] and semantic segmentation [9], and we adapt their ability to preserve spatial information in our WSL context.

Secondly, we incorporate a new multi-map WSL transfer layer (middle of Figure 2), which explicitly learns multiple localized features related to complementary class modalities, *e.g*. head and legs for dogs in Figure 1. Our multi-map strategy is not specifically designed for any particular kind of feature, *e.g*. part or view-based features, as some approaches are [16, 22, 10, 9].

Finally, we address the problem of aggregating spatial scores into a global prediction, which is a crucial issue for WSL training. We propose a new pooling strategy (right of Figure 2) which generalizes several approaches in the literature, including (top) max pooling [44, 39], global average pooling [70] or negative evidence models [47, 12, 13].

We also present a thorough evaluation of the WILDCAT model on six datasets, reporting outstanding performances on classification, WSL pointwise detection and segmentation tasks.

2. Related Work

Despite excellent performances, deep ConvNets [34, 56, 28] carry limited invariance properties, *i.e*. small shift invariance through pooling layers [62, 55, 7]. This is questionable for object or scene databases with strong scale and translation variations. One option to detect informative image regions is to revisit the Bag of Words (BoW) model [57, 2], by using deep features as local region activations [27, 25, 24] or by designing specific BoW layers, *e.g*. NetVLAD [1].

Another option to gain strong invariance is to consider a Weakly Supervised Learning framework (WSL), where we can explicitly align image regions. An important paradigm for WSL training is Multiple Instance Learning (MIL) [11], which considers an image as a bag of instances (regions). The main issue concerns the aggregation function to pool instance scores into a global prediction. Different strategies have been explored to combine deep models and MIL. Max pooling [44] only selects the most informative region for the MIL prediction. Recent alternatives include Global Average Pooling (GAP) [70], soft max in LSE pooling [58], Learning from Label Proportion (LLP) [65, 36], and top max scoring [39]. Negative evidence models [47, 12, 13] explicitly select regions accounting for the absence of the class. In WILDCAT, we propose to incorporate negative evidence insights, but with a differentiate positive and negative contribution process.

Concerning the WSL localization task, [5] uses label co-occurrence information and a coarse-to-fine strategy based on deep feature maps to predict object locations. ProNet [58] uses a cascade of two networks: the first generates bounding boxes and the second classifies them. Similarly, [6] proposes an specific architecture with two branches dedicated to classification and detection. Another important WSL application is segmentation. Many methods are based on MIL framework: MIL-FCN [49] extends MIL to multi-class segmentation, MIL-Base [50] introduces a soft extension of MIL, EM-Adapt [45] includes an adaptive bias into the MIL framework, and Constrained CNN (CCNN) [48] uses a loss function optimized for any set of linear constraints on the output space of a CNN.

Similarly to WSL, the attention-based models [63, 29, 66, 64] select relevant regions to support decisions. However the WSL methods usually include some structure on the selection process while it is implicit in attention-based approaches.

Different semantic categories are often characterized by multiple localized attributes corresponding to different class modalities (see for example head and legs for the *dog* class in Figure 1). The seminal DPM model [16] including several template regions for decision has been extensively studied [54, 46], optionally incorporating priors, *e.g.* sparsity or diversity, in order to learn sensible models [30, 59]. While [22, 60] are direct generalizations of DPM to CNN, R-FCN [10] improves performances by explicitly learning several part models and using a part-based pooling of features designed for accurate spatial localization and directly inserted at the top of the network. MR-CNN [19] exploits several modalities by modeling objects with a fixed set of few local features (*e.g.* parts, context) and incorporating segmentation cues. Combining different regions has also recently been addressed through explicit context modeling [23], or by modeling region correlations as in RRSVM [61]. For fine-grained recognition, multi-feature detection has been tackled in the fully supervised setting [67, 40, 68], and in WSL [33].

When computing local features with deep models, the most naive approach is to rescale each region into a fixed-size vector adapted to the CNN architecture, as done in early works for detection, *e.g.* R-CNN [21], or scene understanding [27, 25, 43, 12]. Since this approach is highly inefficient, there have been extensive attempts for using convolutional layers to share feature computation, for image classification [44, 13, 70], object detection [22, 20, 52] or image segmentation [8, 42]. However, fully connected layers are beneficial in standard deep architectures, *e.g.* AlexNet [34] or VGG [56]. Recently, the huge success of Fully Convolutionnal Networks (FCNs) for image classification, *e.g.* ResNet [28], has been driving successful approaches using FCN for fully supervised object detection [10] and image segmentation [9], which enable complete feature sharing and state-of-the-art performances. Our approach adapts these insights from these latest FCNs to the WSL setting.

3. WILDCAT Model

The overall WIDLCAT architecture (Figure 2) is based on a FCN which is suitable for spatial predictions [42], a multi-map WSL transfer layer encoding modalities associated with classes, and a global pooling for WSL that learns accurate localization. We now delve into each of the three parts of the model.

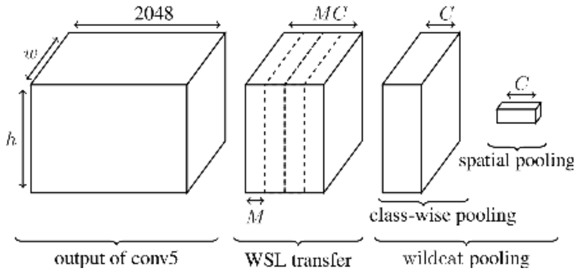

Figure 3. WILDCAT local feature encoding and pooling. Class modalities are encoded with a multi-map WSL transfer layer and pooled separately for all classes. Local features are then aggregated with a global spatial pooling to yield a single score per class.

3.1. Fully convolutional architecture

The selection of relevant information within feature maps is a major issue in WSL. It impacts the localization of the learned representation and the precision of the results (*e.g.* semantic segmentation or object detection). We thus expect the resolution of the feature maps to be a key component for WILDCAT: finer maps keep more spatial resolution and lead to more specific regions (*e.g.* objects, parts).

To this end we exploit the recently introduced FCN ResNet-101 [28] (left of Figure 2) that naturally preserves spatial information throughout the network. It also computes local features from all the regions in a single forward pass, without resizing them. Besides, ResNet architectures are effective at image classification while being parameter- and time-efficient [28]. This kind of architecture has been exploited to speed up computation and to produce accurate spatial predictions in fully supervised setups, *e.g.* in object detection [10] and semantic segmentation [9].

We use the publicly released model pre-trained on ImageNet dataset [53] and remove the last layers (global average pooling and fully connected) to replace them with WSL transfer and wildcat pooling layers (Figure 3) described in the following.

3.2. Multi-map transfer layer

We introduce a multi-map WSL transfer layer that learns multiple class-related modalities, encoded into M feature maps per class through 1×1 convolutions (middle of Figure 2). The modalities are learned in a WSL fashion with only the image-level labels and the transfer layer keeps spatial resolution, key in WSL. We note $w \times h \times d$ the size of *conv5* maps of ResNet-101, which is $\frac{W}{32} \times \frac{H}{32} \times 2048$ for an original image of size $W \times H \times 3$ [28]. The transfer output is then of size $w \times h \times MC$ (Figure 3).

The M modalities aim at specializing to different class-specific features, *e.g.* parts [9, 10] (head and legs of dog in Figure 1) or views [16, 22]. We highlight differences with some specific encoding approaches: position-sensitive RoI pooling in R-FCN [10] forces position-based specialization (relative to the object) while our method can also learn other kind of features, *e.g*. semantic parts (Figure 1). In the same way DPM [16] learns only discriminating parts where our multi-map transfer model can find more general features, *e.g*. context. Furthermore, contrarily to the DPM where a different model is learned for each view, we share most of the computation within the FCN, which is more efficient. We note that when $M = 1$ this reduces to a standard classification layer, *i.e*. into C classes.

3.3. Wildcat pooling

WILDCAT learns from image-level labels so we need a way to summarize all information contained in the feature maps for each class (right of Figure 2). We note that there are no more learned parameters in this pooling layers, which means we can directly interpret and visualize feature maps at this level [70, 10].

We perform this in two steps (Figure 3): a class-wise pooling (Equation (1)) that combines the M maps from the multi-map transfer layer, then a spatial pooling module (Equation (2)) that selects relevant regions within the maps to support predictions. This leads to wildcat pooling, a two-stage pooling operation to compute the score s^c of class c:

$$\begin{cases} \bar{z}^c_{i,j} = \underset{m\in\{1,\dots,M\}}{\mathrm{Cl.\,Pool}} \; z^{c,m}_{i,j} & (1) \\ s^c = \underset{(i,j)\in\{1,\dots,w\}\times\{1,\dots,h\}}{\mathrm{Sp.\,Pool}} \; \bar{z}^c_{i,j} & (2) \end{cases}$$

where z is the output of the transfer layer, $\mathrm{Cl.\,Pool}$ is the chosen class-wise pooling function and $\mathrm{Sp.\,Pool}$ is the spatial aggregation process.

Class-wise pooling. The first step consists in combining the M maps for all classes independently, and is described in Equation (1) with a generic pooling function $\mathrm{Cl.\,Pool}$. We use average pooling in the following. The maps are transformed from $w \times h \times MC$ to $w \times h \times C$ (Figure 3). When $M = 1$ this operation is not needed as each class is already represented by a single map.

We note that even if a multi-map followed by an average pooling is functionally equivalent to a single convolution (*i.e*. $M = 1$), the explicit structure it brings with M modalities has important practical advantages making training easier. We empirically show that $M > 1$ yields better results than regular $M = 1$.

Spatial pooling. We now introduce our new spatial aggregation method implementing the second, spatial pooling step in Equation (2) for each map c:

$$s^c = \max_{\mathbf{h}\in\mathcal{H}_{k^+}} \frac{1}{k^+}\sum_{i,j} h_{i,j}\bar{z}^c_{i,j} + \alpha\left(\min_{\mathbf{h}\in\mathcal{H}_{k^-}} \frac{1}{k^-}\sum_{i,j} h_{i,j}\bar{z}^c_{i,j}\right) \quad (3)$$

where $\mathcal{H}_k$ is such that $\mathbf{h} \in \mathcal{H}_k$ satisfies $h_{i,j} \in \{0, 1\}$ and $\sum_{i,j} h_{i,j} = k$. It consists in selecting for each class c the k^+ (resp. k^-) regions with the highest (resp. lowest) activations from input $\bar{z}^c$. The output s^c for class c of this layer is the weighted average of scores of all the selected regions. We only consider regions defined by single neurons in the convolutional feature maps.

k^+	k^-	α	Pooling
1	0	0	Maximum [44]
k / ρn	0	0	Top instances [39] / LLP [65]
n	0	0	Average [70]
k	k	1	WELDON [13]

Table 1. Generalization of wildcat spatial pooling to other existing MIL approaches with corresponding parameters. n is the total number of regions, ρ is the proportion of positive labels in LLP, k is an arbitrary number of regions to choose.

Several similar MIL approaches have been used but our proposed model generalizes them in numerous of ways. The corresponding parameters are described in Table 1. The standard max-pooling MIL approach [44] is obtained with only one element, and both top instance model [39], Learning with Label Proportion [65] and global average pooling [70] can be obtained with more. Drawing from negative evidence [47, 12, 13] we can incorporate minimum scoring regions to support classification and our spatial pooling function can reduce to the kMax+kMin layer of [13].

Maximum and minimum scoring regions both are important for good results [12, 13], but do not bring the same kind of information. We explore relative weighting of both types of regions by introducing a factor α which trades off relative importance between both terms. We hypothesize that maximum scoring regions are more useful for classification as they directly support the decision, while minimum scoring regions essentially act as regularization. With $\alpha < 1$ WILDCAT should focus more on discriminating regions and then better localize features than with $\alpha = 1$.

Discussion

WILDCAT architecture is composed of a transfer layer followed by pooling. Since there are no parameters to learn in the pooling module, the transfer layer performs classification and it is easy to visualize heatmaps with direct localization of discriminating regions. We note that this kind of architecture is reversed in [70] where pooling is performed before the last fully connected layer, as in the original ResNet architecture [28] for example. However this order requires an unnatural way of visualizing class-specific heatmaps [70].

It is shown in [70] that if the spatial aggregation method is linear, *e.g*. global average pooling, then the order of both layers is not important, but the two configurations can behave differently with a non linear pooling function such as

wildcat spatial pooling. The difference is more significant when $k^+ + k^-$ is low, *i.e.* when wildcat spatial pooling really differs from global average pooling. We evaluate the impact of this design choice and of the chosen pooling function in the experiments and show that our architecture yields better results.

3.4. WILDCAT applications

Training phase. Our WILDCAT model is based on the backbone architecture ResNet-101 [28]. We initialize it from a model pre-trained on ImageNet [53] and train it with Stochastic Gradient Descent (SGD) with momentum with image-level labels only. All the layers of the network are fine tuned. The input images are warped to a square size at a given scale. We use a multi-scale setup where a different model is learned for each scale and they are combined with Object Bank [38] strategy.

WILDCAT is designed to learn from image-level supervision only: the same training procedure is used for image classification, weakly supervised pointwise object detection and weakly supervised semantic segmentation. When learning WILDCAT, the gradients are backpropagated through the wildcat layer only within the $k^+ + k^-$ selected regions, all other gradients being discarded [13]. The selection of right regions for backpropagation is key to learn precisely localized features without any spatial supervision [58].

Inference phase. Predictions differ according to the task at hand. For image classification, prediction simply takes the single-value output of the network (like in training). Object detection and semantic segmentation require spatial predictions so we extract the class-specific maps before spatial pooling to keep spatial resolution. They are at resolution $\frac{1}{32}$ with respect to the input image for ResNet-101 architecture [28]. For weakly supervised pointwise object detection, we extract the region (*i.e.* neuron in the feature map) with maximum score for each class and use it for point-wise localization, as it is done in [44, 5]. For weakly supervised semantic segmentation we compute the final segmentation mask either by taking the class with maximum score at each spatial position independently or by applying a CRF for spatial prediction as is common practice [8, 48].

4. Classification Experiments

We evaluate WILDCAT for classification tasks. Our model is implemented with Torch7 (http://torch.ch/). To show the robustness of our method in very different recognition contexts, we evaluate it on six datasets: object recognition (VOC 2007 [14], VOC 2012 [15]), scene categorization (MIT67 [51] and 15 Scene [37]), and visual recognition where the context plays an important role (MS COCO [41], VOC 2012 Action [15]). The performances on MIT67, 15 Scene, VOC 2007 and 2012 are evaluated following the standard protocol. On MS COCO dataset (resp. VOC 2012 Action), we follow the protocol of [44] (resp. [13]). Detailed information is available in section 1 of Supplementary. We first compare our model to state-of-the-art methods, then we analyze our contributions.

4.1. Comparison with state-of-the-art methods

We compare WILDCAT with several state-of-the-art object classification models. The parameters of our model are fixed at $M = 4$ and $\alpha = 0.7$. The results for object classifications (Table 2) show that WILDCAT outperforms all recent methods by a large margin. We can point out a large improvement compared to deep features computed on the whole image with ResNet-101 [28]: 5.2 pt on VOC 2007 and 4.2 pt on VOC 2012. Note that these differences directly measure the relevance of the proposed WSL method, because WILDCAT is based on ResNet-101. We also compare our model to region selection approaches: DeepMIL [44], WELDON [13] and RRSVM [61]. Although using multiple regions as in [44, 13, 61] is important, we show here that we can further significantly improve performances by learning multiple modalities per category.

Method	VOC 2007	VOC 2012
VGG16 [56]	89.3	89.0
DeepMIL [44]	-	86.3
WELDON [13]	90.2	-
ResNet-101 (*) [28]	89.8	89.2
ProNet [58]	-	89.3
RRSVM [61]	92.9	-
SPLeaP [35]	88.0	-
WILDCAT	**95.0**	**93.4**

Table 2. Classification performances (MAP) on object recognition datasets. We used VOC evaluation server to evaluate on VOC 2012. (*) means that results are obtained with online code https://github.com/facebook/fb.resnet.torch.

In Table 3, we compare WILDCAT results for scene categorization with recent global image representations used for image classification: deep features [71, 28], and global image representation with deep features computed on image regions: MOP CNN [25] and Compact Bilinear Pooling [18]. Again, WILDCAT gets the best results, showing the capacity of our model to seek discriminative part regions, whereas background and non-informative parts are incorporated into image representation with other approaches. We also compare WILDCAT to existing part-based models including negative evidence during training [47] and non-linear part classifiers combined with part-dependent soft pooling [35]. WILDCAT also outperforms recent WSL models with different spatial pooling strategies: 17 pt with respect to GAP GoogLeNet [70] which uses a global average pooling and 6 pt with respect to WELDON [13] which uses a kMax+kMin pooling. This validates the relevance of our spatial pooling.

Method	15 Scene	MIT67
CaffeNet Places [71]	90.2	68.2
MOP CNN [25]	-	68.9
Negative parts [47]	-	77.1
GAP GoogLeNet [70]	88.3	66.6
WELDON [13]	94.3	78.0
Compact Bilinear Pooling [18]	-	76.2
ResNet-101 (*) [28]	91.9	78.0
SPLeaP [35]	-	73.5
WILDCAT	**94.4**	**84.0**

Table 3. Classification performances (multi-class accuracy) on scene datasets.

Finally, we report the excellent performances of WILDCAT on context datasets in Table 4. We compare our model to ResNet-101 deep features [28] computed on the whole image and recent WSL models for image classification: DeepMIL [44], WELDON [13] and ProNet [58]. WILDCAT outperforms ResNet-101 by 8 pt on both datasets, again validating our WSL model in this context.

Method	VOC 2012 Action	MS COCO
DeepMIL [44]	-	62.8
WELDON [13]	75.0	68.8
ResNet-101 (*) [28]	77.9	72.5
ProNet [58]	-	70.9
WILDCAT	**86.0**	**80.7**

Table 4. Classification performances (MAP) on context datasets.

4.2. Further analysis

We detail the impact of our contributions on three datasets: VOC 2007, VOC 2012 Action and MIT67. We present results for an input image of size 448 × 448 and $k^+ = k^- = 1$, but similar behaviors are observed for other scales and larger k^+ and k^-. By default, our model parameters α and M are fixed to 1.

Deep structure. Firstly, to validate the design choice of the proposed WILDCAT architecture, we evaluate two different configurations (see discussion before Section 3.4):
(a) conv5 + conv + pooling (our architecture);
(b) conv5 + pooling + conv (architecture proposed in [70]).These two configurations are different for the nonlinear WILDCAT pooling scheme described in Section 3.3, and their comparison is reported in Table 5. We can see that our architecture (a) leads to a consistent improvement over architecture (b) used in GAP [70] on all three datasets, *e.g.* 1.7 pt on VOC07.

Method	VOC07	VOC12Action	MIT67
Architecture (a)	**89.0**	**78.9**	**69.6**
Architecture (b)	87.3	77.5	68.1

Table 5. Classification performances for architectures (a) and (b).

Note that the strategy of architecture (a) has a very different interpretation from (b): (a) classifies each region independently and then pools the region scores, whereas (b) pools the output of the convolution maps and then performs image classification on the pooled space.

Impact of parameter α. We investigate the effect of the parameter α on classification performances. From the results in Figure 4, it is clear that incorporating negative evidence, *i.e.* $\alpha > 0$, is beneficial for classification, compared to standard max pooling, *i.e.* $\alpha = 0$. We further note that using different weights for maximum and minimum scores, *i.e.* $\alpha \neq 1$, yields better results than with $\alpha = 1$ from [13], with best improvement of 1.6 pt (resp. 2 and 1.8) with $\alpha = 0.6$ (resp. 0.7 and 0.8) on VOC 2007 (resp. VOC 2012 Action and MIT67). This confirms the relevance of using a relative weighting for negative evidence. Moreover our model is robust with respect to the value of α.

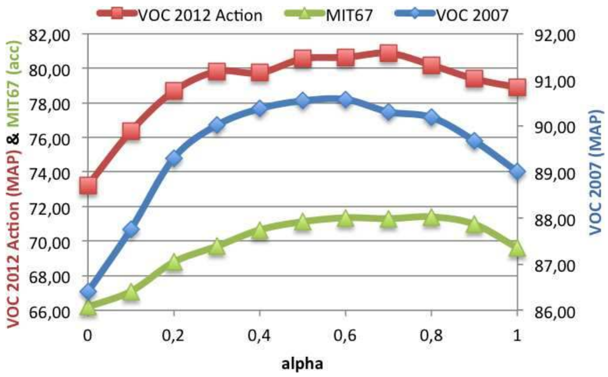

Figure 4. Analysis of parameter α.

Number of modalities. Another important hyperparameter of our model is the number of modalities (M) used in the multi-map transfer layer. The performances for different values of M are reported in Table 6. Explicitly learning multiple modalities, *i.e.* $M > 1$, yields large gains with respect to a standard classification layer, *i.e.* $M = 1$ [13]. However encoding more modalities than necessary (*e.g.* $M = 16$) might lead to overfitting since the performances decrease. The best improvement is 3.5 pt (resp. 4.3 and 3.5) with $M = 8$ (resp. 8 and 12) on VOC 2007 (resp. VOC 2012 Action and MIT 67). Examples of heatmaps for the same category are shown in Figure 6.

Ablation study. We perform an ablation study to illustrate the effect of each contribution. Our baseline is a WSL transfer with $M = 1$ and the spatial pooling with $\alpha = 1$. The

M	1	2	4	8	12	16
VOC 2007	89.0	91.0	91.6	**92.5**	92.3	92.0
VOCAction	78.9	81.5	82.1	**83.2**	83.0	82.7
MIT67	69.6	71.8	72.0	72.8	**73.1**	72.9

Table 6. Analysis of multi-map transfer layer.

results are reported in Table 7. From this ablation study, we can draw the following conclusions:

– Both $\alpha = 0.7$ and $M = 4$ improvements result in large performance gains on all datasets;

– Combining $\alpha = 0.7$ and $M = 4$ improvements further boost performances: 0.4 pt on VOC 2007, 0.8 pt on VOC 2012 Action and 0.8 on MIT67. This shows the complementarity of both these contributions.

max+min	α=0.7	$M = 4$	VOC07	VOCAc	MIT67
✓			89.0	78.9	69.6
✓	✓		90.3	80.9	71.3
✓		✓	91.6	82.1	72.0
✓	✓	✓	**92.0**	**82.9**	**72.8**

Table 7. Ablation study on VOC 2007, VOC 2012 Action (VOCAc) and MIT67. The results are different from results of section 4.1 because only one scale is used for this analysis.

5. Weakly Supervised Experiments

In this section, we show that our model can be applied to various tasks, while being trained from global image labels only. We evaluate WILDCAT for two challenging weakly supervised applications: pointwise localization and segmentation.

5.1. Weakly supervised pointwise localization

We evaluate the localization performances of our model on PASCAL VOC 2012 *validation* set [15] and MS COCO *validation* set [41]. The performances are evaluated with the point-based object localization metric introduced by [44]. This metric measures the quality of the detection, while being less sensitive to misalignments compared to other metrics such as IoU [15], which requires the use of additional steps (*e.g.* bounding box regression).

WILDCAT localization performances are reported in Table 8. Our model significantly outperforms existing weakly supervised methods. We can notice an important improvement between WILDCAT and MIL-based architecture DeepMIL [44], which confirms the relevance of our spatial pooling function. In spite of its simple and multipurpose architecture, our model outperforms by a large margin the complex cascaded architecture of ProNet [58]. It also outperforms the recent weakly supervised model [5] by 3.2 pt (resp. 4.2 pt) on VOC 2012 (resp. MS COCO), which use a more complex strategy than our model, based on search-trees to predict locations.

Method	VOC 2012	MS COCO
DeepMIL [44]	74.5	41.2
ProNet [58]	77.7	46.4
WSLocalization [5]	79.7	49.2
WILDCAT	**82.9**	**53.4**

Table 8. Pointwise object localization performances (MAP) on PASCAL VOC 2012 and MS COCO.

Note that since the localization prediction is based on classification scores, good classification performance is important for robust object localization. In Figure 5, we evaluate the classification and localization performances with respect to α on VOC 2012. Both classification and localization curves are very similar. The best localization performances are obtained for $\alpha \in [0.6, 0.7]$, and the improvement between $\alpha = 1$ and $\alpha = 0.7$ is 1.6 pt. We can note that the worst performance is obtained for $\alpha = 0$, which confirms that the contextual information brought by the minimum is useful for both classification and localization.

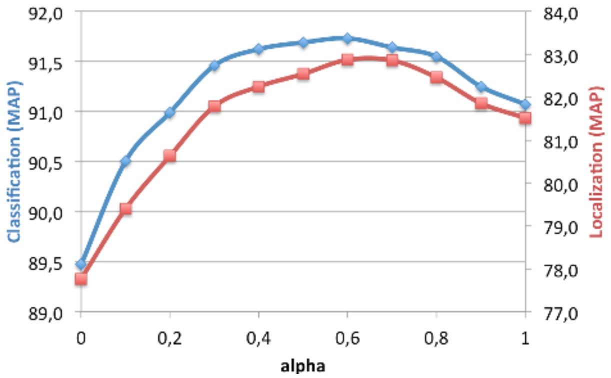

Figure 5. Classification and localization performances with respect to α on VOC 2012.

5.2. Weakly supervised segmentation

We evaluate our model on the PASCAL VOC 2012 image segmentation dataset [15], consisting of 20 foreground object classes and one background class. We train our model with the *train* set (1,464 images) and the extra annotations provided by [26] (resulting in an augmented set of 10,582 images), and test it on the *validation* set (1,449 images). The performance is measured in terms of pixel Intersection-over-Union (IoU) averaged across the 21 categories. As in existing methods, we add a fully connected CRF (FC-CRF) [32] to post-process the final output labeling.

Segmentation results. The result of our method is presented in Table 9. We compare it to weakly supervised methods that use only image labels during training. We can see that WILDCAT without CRF outperforms existing weakly supervised models by a large margin. We note a

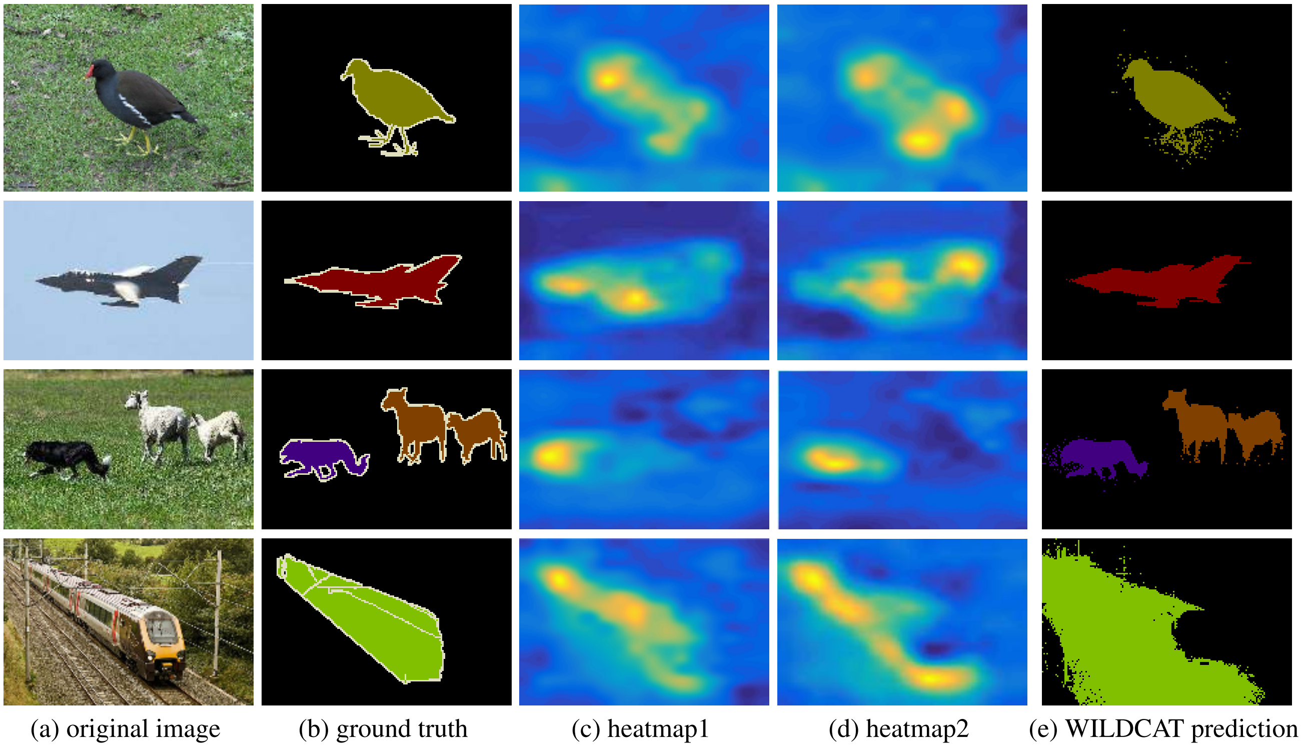

Figure 6. Segmentation examples on VOC 2012. Our prediction is correct except for the train (last row) where our model aggregated rails and train regions. For objects as *bird* or *plane*, one can see how two heatmaps (heatmap1 (c) and heatmap2 (d) representing the same class: respectively *bird, aeroplane, dog* and *train*) succeed to focus on different but relevant parts of the objects.

large gain with respect to MIL models based on (soft-)max pooling [49, 50], which validates the relevance of our pooling for segmentation. The improvement between WILDCAT with CRF and the best model is 7.1 pt. This confirms the ability of our model to learn discriminative and accurately localized features. We can note that all the methods evaluated in Table 9 have comparable complexity.

Method	Mean IoU
MIL-FCN [49]	24.9
MIL-Base+ILP+SP-sppxl [50]	36.6
EM-Adapt +FC-CRF [45]	33.8
CCNN + FC-CRF [48]	35.3
WILDCAT	39.2
WILDCAT + FC-CRF	**43.7**

Table 9. Comparison of weakly supervised semantic segmentation methods on VOC 2012.

With a quite more complex strategy, the very recent paper [31] presents impressive results (50.7 MIoU). The training scheme in [31] incorporates different terms, which are specifically tailored to segmentation: one enforces the segmentation mask to match low-level image boundaries, another one incorporates prior knowledge to support predicted classes to occupy a certain image proportion. In contrast, WILDCAT uses a single model which is trained in the same manner for the three tasks, *i.e*. classification, localization and segmentation.

Qualitative Results. In Figure 6, we show predicted segmentation masks for four images. Compared to ground truth ((b) column), we can see that our predicted segmentation masks ((e) column) are always relevant, except for the last example where the rails and the train are glued together. The heatmaps from the same class (columns (c) and (d)) show different modalities learned by our model. When successful, they focus on different parts of the objects. For example, on the first row, the heatmap (c) focuses on the head of the bird whereas the heatmap (d) focuses on the legs and the tail.

6. Conclusion

We propose WILDCAT, a new weakly supervised learning dedicated to learn discriminative localized visual features by using only image-level labels during training. Extensive experiments have shown the effectiveness of WILDCAT on three main visual recognition tasks: image classification, for which we report outstanding performances on six challenging datasets, and WSL localization and segmentation, using a single and generic training scheme for all tasks.

Future works include adapting WILDCAT for semantic applications where localized features are crucial, *e.g*. Visual Question Answering [64, 4] or Visual Grounding [17].

References

[1] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016.

[2] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araujo. Pooling in image representation: the visual codeword point of view. *Computer Vision and Image Understanding*, 2012.

[3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the Point: Semantic Segmentation with Point Supervision. In *ECCV*, 2016.

[4] H. Ben Younes, R. Cadene, M. Cord, and N. Thome. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *arXiv*, 2017.

[5] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. S. Manjunath. Weakly supervised localization using deep feature maps. In *ECCV*, 2016.

[6] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[7] M. Blot, M. Cord, and N. Thome. Max-min convolutional neural networks for image classification. In *IEEE International Conference on Image Processing (ICIP)*, 2016.

[8] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *ICLR*, 2015.

[9] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016.

[10] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[11] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 1997.

[12] T. Durand, N. Thome, and M. Cord. MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking. In *ICCV*, 2015.

[13] T. Durand, N. Thome, and M. Cord. WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks. In *CVPR*, 2016.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. In *PAMI*, 2010.

[17] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *arXiv:1606.01847*, 2016.

[18] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact Bilinear Pooling. In *CVPR*, 2016.

[19] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, 2015.

[20] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[22] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In *CVPR*, 2015.

[23] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *CVPR*, 2015.

[24] H. Goh, N. Thome, M. Cord, and J.-H. Lim. Learning Deep Hierarchical Visual Feature Coding. *IEEE Transactions on Neural Networks and Learning Systems*, 2014.

[25] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014.

[26] B. Hariharan, P. Arbelaez, L. D. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[29] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial Transformer Networks. In *NIPS*, 2015.

[30] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[31] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[32] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*. 2011.

[33] J. Krause, T. Gebru, J. Deng, L.-J. Li, and F.-F. Li. Learning features and parts for fine-grained recognition. In *ICPR*, 2014.

[34] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[35] P. Kulkarni, F. Jurie, J. Zepeda, P. Pérez, and L. Chevallier. Spleap: Soft pooling of learned parts for image classification. In *ECCV*, 2016.

[36] K.-T. Lai, F. X. Yu, M.-S. Chen, and S.-F. Chang. Video event detection by inferring temporal instance labels. In *CVPR*, 2014.

[37] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[38] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification. In *NIPS*, 2010.

[39] W. Li and N. Vasconcelos. Multiple Instance Learning for Soft Bags via Top Instances. In *CVPR*, 2015.

[40] D. Lin, X. Shen, C. Lu, and J. Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *CVPR*, 2015.

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[42] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*, 2015.

[43] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *CVPR*, 2014.

[44] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[45] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. In *ICCV*, 2015.

[46] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb. Reconfigurable models for scene recognition. In *CVPR*, 2012.

[47] S. N. Parizi, A. Vedaldi, A. Zisserman, and P. F. Felzenszwalb. Automatic discovery and optimization of parts for image classification. In *ICLR*, 2015.

[48] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. In *ICCV*, 2015.

[49] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Multi-Class Multiple Instance Learning. In *ICLR (Workshop)*, 2015.

[50] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.

[51] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[52] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 2015.

[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[54] F. Sadeghi and M. F. Tappen. Latent pyramidal regions for recognizing scenes. In *ECCV*, 2012.

[55] W. Shang, K. Sohn, D. Almeida, and H. Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *ICML*, 2016.

[56] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[57] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[58] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. Bourdev. ProNet: Learning to Propose Object-Specific Boxes for Cascaded Neural Networks. In *CVPR*, 2016.

[59] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *ICCV*, 2013.

[60] L. Wan, D. Eigen, and R. Fergus. End-to-end integration of a convolution network, deformable parts model and non-maximum suppression. In *CVPR*, 2015.

[61] Z. Wei and M. Hoai. Region Ranking SVM for Image Classification. In *CVPR*, June 2016.

[62] J. Weng, N. Ahuja, and T. S. Huang. Cresceptron: a self-organizing neural network which grows adaptively. In *International Joint Conference on Neural Networks (IJCNN)*, 1992.

[63] Xu, Ba, Kiros, Cho, Courville, Salakhutdinov, Zemel, and Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[64] H. Xu and K. Saenko. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *ECCV*, 2016.

[65] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S.-F. Chang. $\propto$svm for learning with label proportions. In *ICML*, 2013.

[66] Zhang, Lin, Brandt, Shen, and Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.

[67] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas. SPDA-CNN: Unifying Semantic Part Detection and Abstraction for Fine-grained Recognition. In *CVPR*, 2016.

[68] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014.

[69] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose Aligned Networks for Deep Attribute Modeling. In *ECCV*, 2014.

[70] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016.

[71] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In *NIPS*, 2014.