# Soft Proposal Networks for Weakly Supervised Object Localization

Yi Zhu[1], Yanzhao Zhou[1], Qixiang Ye[†1], Qiang Qiu[2] and Jianbin Jiao[†1]

[1]University of Chinese Academy of Sciences
[2]Duke University
{zhuyi215, zhouyanzhao215}@mails.ucas.ac.cn, {qxye, jiaojb}@ucas.ac.cn, qiang.qiu@duke.edu

## Abstract

*Weakly supervised object localization remains challenging, where only image labels instead of bounding boxes are available during training. Object proposal is an effective component in localization, but often computationally expensive and incapable of joint optimization with some of the remaining modules. In this paper, to the best of our knowledge, we for the first time integrate weakly supervised object proposal into convolutional neural networks (CNNs) in an end-to-end learning manner. We design a network component, Soft Proposal (SP), to be plugged into any standard convolutional architecture to introduce the nearly cost-free object proposal, orders of magnitude faster than state-of-the-art methods. In the SP-augmented CNNs, referred to as Soft Proposal Networks (SPNs), iteratively evolved object proposals are generated based on the deep feature maps then projected back, and further jointly optimized with network parameters, with image-level supervision only. Through the unified learning process, SPNs learn better object-centric filters, discover more discriminative visual evidence, and suppress background interference, significantly boosting both weakly supervised object localization and classification performance. We report the best results on popular benchmarks, including PASCAL VOC, MS COCO, and ImageNet.* [1]

## 1. Introduction

The success of object proposal methods greatly drives the progress of the object localization. With the popularity of deep learning, object detection is evolving from pipelined frameworks [11, 12] to unified frameworks [17, 21, 22], thanks to the unprecedentedly learning capability of convolutional neural networks (CNNs) and abundant object bounding box annotations.

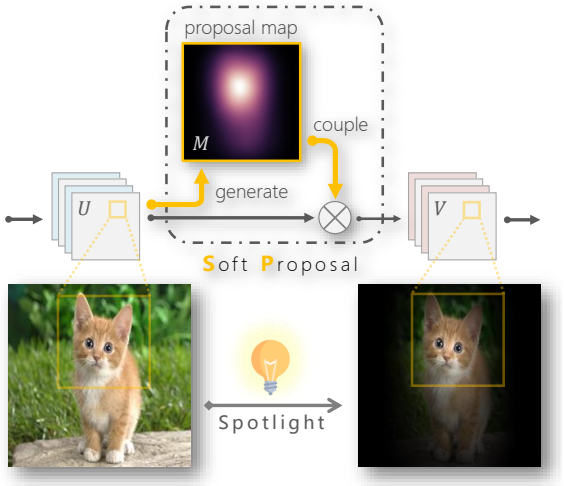



Figure 1. Soft Proposal (SP) module can be inserted after any CNN layer. A proposal map $M$ is generated based on deep feature maps $U$ and then projected back, which results in feature maps $V$. During the end-to-end learning procedure, $M$ iteratively evolves and jointly optimizes with the feature maps to spotlight informative object regions.

Despite the unified frameworks achieve remarkable performance in supervised object detection, they can not be directly applied to weakly supervised object localization where only image-level labels, *i.e.*, the presence or absence of object categories, are available during training.

To tackle the problem of weakly supervised object localization, many of the conventional methods follow a multi-instance learning (MIL) framework by using object proposal methods [5, 8, 14, 31, 34]. The learning objective is designed to choose an instance (a proposal) from each bag (an image with multiple proposals) to minimize the image classification error; however, the pipelined proposal-and-classification method is sub-optimal as the two steps can not be jointly optimized. Recent research [6] demonstrates that the convolutional filters in CNN can be seen as object detectors and their feature maps can be aggregated to produce Class Activation Map (CAM) [36], which specifies

[†]Corresponding Authors
[1]Source code is publicly available at yzhou.work/SPN

the spatial distribution of discriminative patterns for different image classes. This end-to-end network demonstrates a surprising capability to localize objects under weak supervision. However, without the prior knowledge of informative object regions during training, conventional CNNs can be misled by co-occurrence patterns and noisy backgrounds, Fig. 2. The weakly supervised setting increases the importance of high-quality object proposals, but the problem to integrate the proposal functionality into a unified framework for weakly supervised object localization remains open.

In this paper, we design a network component, Soft Proposal (SP), to be plugged into standard convolutional architectures for nearly cost-free object proposal (∼0.9ms per image, 10×faster than RPN [22], 200×faster than EdgeBoxes [37]), Fig. 1. CNNs using SP module are referred to as Soft Proposal Networks (SPNs). In SPNs, iteratively evolved object proposals are projected back on the deep feature maps, and further jointly optimized with network parameters, using image-level labels only. We further apply the SP module to successful CNNs including CNN-S, VGG, and GoogLeNet, and upgrade them to Soft Proposal Networks (SPNs), which can learn better object-centric filters and discover more discriminative visual evidence for weakly supervised localization tasks.

The meaning of the word "soft" is threefold. First of all, instead of extracting multiple materialized proposal boxes, we predict objectness score for each receptive field, based on the deep feature maps. Next, the proposal couples with deep activation in a probabilistic manner, which not only avoids threshold tuning but also aggregates all information to improve performance. Last but not least, the proposal iteratively evolves along with CNN filters updating.

To summarize, the main contributions of this paper are:

- We design a network component, Soft Proposal (SP), to upgrade conventional CNNs to Soft Proposal Networks (SPNs), in which the network parameters can be jointly optimized with the nearly cost-free object proposal.
- We upgrade successful CNNs to SPNs, including CNN-S, VGG16, and GoogLeNet, and improve the state-of-the-art of weakly supervised object localization by a significant margin.

## 2. Related Work

Weakly supervised object localization problems are often solved with a pipelined approach, *i.e.*, an object proposal method [30, 37] is first applied to decompose images into object proposals, with which a latent variable learning method, *e.g.*, multi-instance learning (MIL), is used to iteratively perform proposal selection and classifier estimation

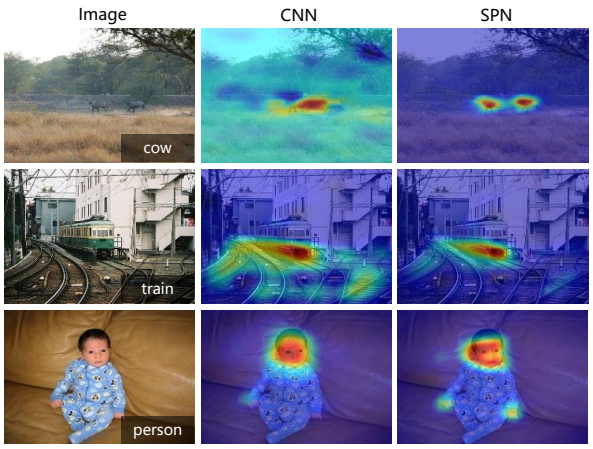



Figure 2. Visualization of Class Activation Maps (CAM) [36] for generic CNN and the proposed SPN. CNNs can be misled by noisy backgrounds, *e.g.*, grass for "cow", and co-occurrence patterns, *e.g.*, rail for "train", and thus miss informative object evidence. In contrast, SPNs focus on informative object regions during training to discover more fine-detailed evidence, *e.g.*, hands for "person", while suppressing background interference. Best viewed in color.

[8, 15, 32, 26, 3, 5, 14]. With the popularity of deep learning, the pipelined approaches have been evolving to end-to-end MIL networks [20, 27] by learning convolutional filters as detectors and using response maps to localize objects.

### 2.1. Object Proposal

Conventional object proposal methods, *e.g.*, Selective Search (SS) [30] and EdgeBoxes (EB) [37], use redundant proposals generated with hand-craft features to hypothesize objects locations. Region Proposal Network (RPN) regresses object locations using deep convolutional features [22], reports the state-of-the-art proposal performance. The success of RPN roots in the localization capability of deep convolutional features; however, such capability is not available until the network is well trained with precise annotations about object locations, *i.e.*, bounding boxes, which limits its applicability to weakly supervised methods.

Our SPN is specified for weakly supervised object localization task with only image-level annotations, *i.e.*, presence or absence of object categories. The key difference between our method to existing ones is that the "soft" proposal is an objectness confidence map instead of materialized boxes. Such a proposal couples with convolutional activation and evolves with the deep feature learning.

### 2.2. Weakly Supervised Localization

**Pipelined methods.** Weakly supervised localization methods often use a stepwise strategy, *i.e.*, first extracting candidate proposals and then learning classification model together with selecting proposals to localize objects. Many approaches have been explored to prevent the learning procedure from getting stuck to a local minimum, *e.g.*, prior

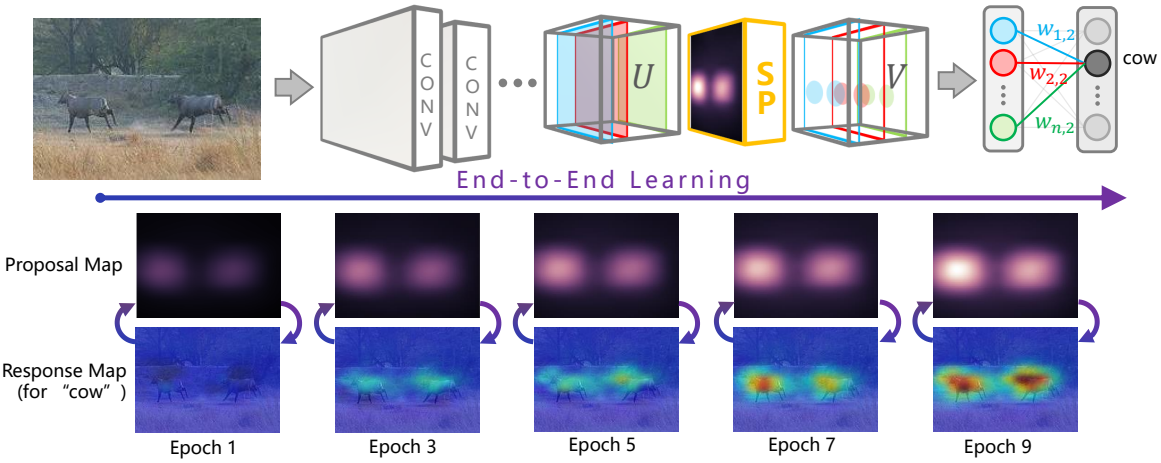



Figure 3. The first row shows the Soft Proposal Network architecture. The second row illustrates the evolution of the proposal map during training epochs (corresponding to the outer loop of Algorithm 1). The third row presents the evolution of the response map for "cow". The proposal map produced by SP module iteratively evolves and jointly optimizes with convolutional filters during the learning phase, leading SPN to discover fine-detailed visual evidence for localization. Best viewed in color.

regularization [3], multi-fold learning [8], and smooth optimization methods [26, 3]. One representative method is WSDDN [5], which significantly improves the object detection performance by performing proposal selection together with classifier learning. ContextLoc [14] updates WSDDN by introducing two context-aware modules which try to expand or contract the fixed proposals in learning procedure to leverage the surrounding context to improve localization. Attention net [29] computes an attention score for each precomputed object proposals. ProNet [27] uses parallel CNN streams for multiple scales to propose possible object regions and then classify these regions via cascaded CNNs.

To the best of our knowledge, we are the first to integrate proposal step into CNNs and achieve jointly updating among proposal generation, object region selection, and object detector estimation under weak supervision.

**Unified frameworks.** Another line of research shows up in weakly supervised localization uses unified network frameworks to perform both localization and classification. The essence of the method Oquab *et al*. [20] is that the deep feature maps are interpreted as a "bag" of instances, where only the highest responses of feature maps contribute to image label prediction in an MIL-like learning procedure. Zhou *et al*. [36] achieve remarkable localization performance by leveraging a global average pooling layer behind the top convolutional layer to aggregate class-specific activation. In the following works, Zhang *et al*. [35] formulate such a class activation procedure as conditional probability backward propagation along convolutional layers to localize discriminative patterns in generic CNNs. Bency *et al*. [2] propose a heuristic search strategy to hypothesize locations of feature maps in a multi-scale manner and grade the corresponding receptive fields by the classification layer.

The main idea of these methods is that the convolutional filters can behave as detectors to activate locations on the deep feature maps, which provide informative evidence for image classification. Despite the simplicity and efficiency of these networks, they are observed missing useful object evidence, as well as being misled by complex backgrounds. The reason behind this phenomenon can be that the filters learned for common object classes are challenged with object appearance variations and background complexity. Our proposed SPN targets at solving such problems by utilizing image-specific objectness prior and coupling it with the network learning.

## 3. Soft Proposal Network

In this section, we present a network component, Soft Proposal (SP), to be plugged into standard convolutional architectures for nearly cost-free object proposal. CNNs using SP module are referred to as Soft Proposal Networks (SPNs), Fig. 3. Despite the SP module can be inserted after any CNN layer, we apply it after the last convolutional layer where the deep features are most informative. For weakly supervised object localization, SPN has an spatial pooling layer with the output features connected to image labels, as illustrated later.

In the learning procedure of SPN, the Soft Proposal Generation step spotlights potential object locations via performing graph propagation over the receptive fields of deep responses, and the Soft Proposal Coupling step aggregates feature maps with the generated proposal map. With iterative proposal generation, coupling, and activation, SPN performs weakly supervised learning in an end-to-end manner.

### 3.1. Soft Proposal Generation

The proposal map, $M \in \mathbb{R}^{N \times N}$, is an objectness map generated by SP module based on the deep feature maps,

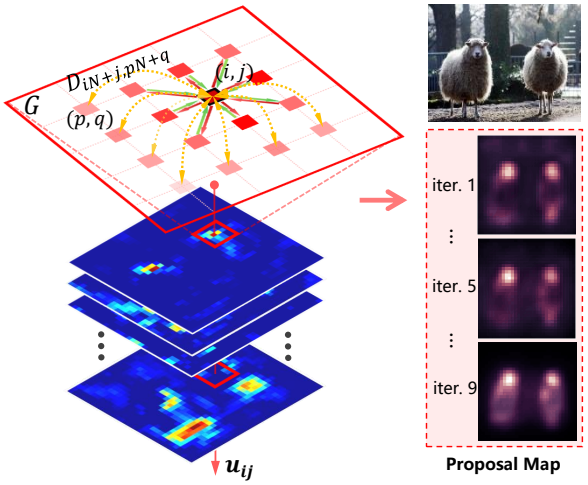



Figure 4. Soft Proposal Generation in a single SPN feedforward pass (corresponding to the inner loop of Algorithm 1). Experimentally, the generation reaches stable in about ten iterations.

Fig. 4. Consider a SP module is inserted after the $l$-th convolutional layer, let $U^l \in \mathbb{R}^{K\times N\times N}$ denote the deep feature maps of the $l$-th convolutional layer, where $K$ is the number of feature maps (channels), $N \times N$ denotes the spatial size of a feature map. Each location $(i, j)$ on $U^l$ has a deep feature vector $\mathbf{u}^l_{ij} = U^l_{:,i,j} \in \mathbb{R}^K$ from all $K$ channels of $U^l$. To generate $M$, a fully connected directed graph $G$ is first constructed by connecting every location on $U^l$, with the weight matrix $D \in \mathbb{R}^{N^2\times N^2}$ where $D_{iN+j,pN+q}$ indicating the weight of edge from node $(i, j)$ to node $(p, q)$.

To calculate the weight matrix $D$, two kinds of objectness measures are utilized: 1). Image regions from the same object category share similar deep features. 2). Neighboring regions exhibit semantic relevance. The objectness confidence are reflected with a dissimilarity measure that combines feature difference and spatial distance, as $D'_{iN+j,pN+q} \triangleq \|\mathbf{u}^l_{ij} - \mathbf{u}^l_{pq}\| \cdot L(i-p, j-q)$, and $L(a, b) \triangleq exp(-\frac{a^2+b^2}{2\epsilon^2})$, where $\epsilon$ is empirically set as $0.15N$ in all experiments. And then the weights of the outbound edges of each node are normalized to 1, *i.e.*, $D_{a,b} = \frac{D'_{a,b}}{\sum_{a=1}^{N} D'_{a,b}}$.

With the weight matrix $D$ defining the edge weight between nodes, a graph propagation algorithm, *i.e.*, random walk [18], is utilized to generate the proposal map $M$. The random walk algorithm iteratively accumulates objectness confidence at the nodes that have high dissimilarity with their surroundings. A node receives confidence from inbound directed edges, and then the confidence among the nodes can be diffused along the outbound directed edges which are connected to all other nodes, Fig. 4. In this procedure, a location transfer confidence to others via globally objectness flow, which not only collects local object evidence but also depresses noise regions. For the convenience of random walk operation, we first reshape the 2D proposal map $M$ to a vector with $N^2$ element, initialized with the value $\frac{1}{N^2}$. $M$ is updated with iteratively multiplying with the weight matrix $D$, as

$$M \leftarrow D \times M. \tag{1}$$

The above procedure is a variant of the eigenvector centrality measure [19], which outputs a proposal map to indicate the objectness confidence of each location on the deep feature maps. Note that the weight matrix $D$ is conditional on the deep feature maps $U^l$, and $U^l$ is conditional on the convolutional filters of the $l$-th layer, $W^l$, in the learning procedure. To show such dependency, Eq. 1 is updated as

$$M \leftarrow D\big(U^l(W^l)\big) \times M. \tag{2}$$

The random walk procedure can be seen as a Markov chain that can reach unique stable state because the chain is ergodic, a property which emerges from the fact that the graph $G$ is by construction strongly connected [13]. Given deep feature maps $U$, Eq. 2 usually reaches its stable state in about ten iterations, and the output $M$ is reshaped from a vector to a 2D proposal map $M \in \mathbb{R}^{N\times N}$.

### 3.2. Soft Proposal Coupling

The proposal map generated with the deep feature maps in a weakly supervised manner can be regarded as a kind of objectness map, which indicates possible object regions. From the perspective of image representation, the proposal map spotlights "regions of interest" that are informative to image classification. $M$ can be integrated into the end-to-end learning via SP module, Fig. 1, to aggregate the image-specific discriminative patterns from deep responses.

In the forward propagation of a SP-augmented CNN, *i.e.*, SPN, each feature map of the coupled $V \in \mathbb{R}^{N\times N}$ is the Hadamard product of the corresponding feature map of $U$ and $M$,

$$V_k = U^l_k(W^l) \circ M, _{k=1,2,\ldots,K}, \tag{3}$$

where the subscript $k$ denotes the channel index and "$\circ$" denotes element-wise multiplication. The coupled feature maps $V$ pass forward to predict scores $y \in \mathbb{R}^C$ of $C$ classes, and then the prediction error $E = \ell(y, t)$ of each sample comes out according to the image labels $t$. $\ell(\cdot)$ is the loss function. In the back-propagation procedure of SPN, the gradient is apportioned by $M$, as

$$\begin{aligned} W^l &= W^l + \Delta W(M) \\ \Delta W(M) &= -\eta \frac{\partial E}{\partial W^l}(M) \end{aligned} \tag{4}$$

where $\eta$ is the network learning rate. $\Delta W(M)$ means that $W^l$ is conditional on $M$, as the gradients of filters $\frac{\partial E}{\partial W^l}$ are conditional on $M$, Eq. 7. Since $W^l$ is conditional on $M$, the SPN learns more informative image regions in each image and depresses noisy backgrounds.

**Algorithm 1** Learning SPN with Soft Proposal Coupling

**Input:** Training images with category labels

**Output:** Network parameters, proposal map for each image.

**repeat**
  initial each element in $M$ with $\frac{1}{N^2}$
  **repeat**
    $M \leftarrow D\big(U^l(W^l)\big) \times M$
  **until** stable state reached
  $V = U^l(W^l) \circ M$, feed forward.
  $W^l = W^l + \Delta W(M)$, backward.
  **for** all the convolutional layers $l$ **do**
    $U^l = W^l * U^{l-1}$
  **end for**
**until** Learning converges

Given the Soft Proposal Generation defined by Eq. 2, the Soft Proposal Coupling defined by Eq. 3, and the back propagation procedure defined by Eq. 4, it is clear that $U^l$, $W^l$, and $M$ are conditional on each other. During training, once the convolutional filters $W^l$ changed by Eq. 4, $U^l$ will also change. Once $U^l$ is updated, a random walk procedure, described in Sec. 3.1, is utilized to update the proposal map $M$. The proposal map $M$ helps SPNs to progressively spotlight feature maps $U^l$ and learn discriminative filters $W^l$, thus the proposals and filters are jointly optimized in SPNs, Fig. 3. The procedure is described in Algorithm 1.

### 3.3. Weakly Supervised Activation

The weakly supervised learning task is performed by firstly using an spatial pooling layer to aggregate deep feature maps to a feature vector, and connecting such a feature vector to image categories with a fully connect layer, Fig. 3. Such an architecture uses weak supervision posed from the end of the network, *i.e.*, the image category annotations, to activate potential object regions.

In the forward propagation of SPN, proposal map $M$ is generated by the SP module inserted behind the $l$-th convolutional layer. The feature maps $U^l$ is computed as

$$U_j^l = (\sum_{i \in S_j} U_i^{l-1} * W_{ij}^l + b_j^l) \circ M, \tag{5}$$

where $S_j$ is a selection of input maps, $b_j^l$ is the additive bias, and $W_{ij}^l$ is the convolutional filters between the $i$-th input map in $U^{l-1}$ and the $j$-th output map in $U^l$.

In the backward propagation of SPN, the error propagates from layer $l+1$ to layer $l$ via the $\delta$, as

$$\begin{aligned} \delta^l &= \frac{\partial E}{\partial U^l} = \frac{\partial E}{\partial U^{l+1}} \frac{\partial U^{l+1}}{\partial U^l} \\ &= \delta^{l+1} \frac{\partial [(U^l * W^{l+1} + b^l) \circ M]}{\partial U^l} \\ &= \delta^{l+1} * W^{l+1} \circ M, \end{aligned} \tag{6}$$

which indicates that the proposal map $M$ spotlights not only informative regions on feature maps but also worth-learning locations. Since the $M$ flows along with gradients $\delta$, inserting one SP module after the top convolutional layer can effect all CNN filters.

Once $\delta^l$ is calculated, we can immediately compute the gradients for filters as

$$\begin{aligned} \frac{\partial E}{\partial W_{ij}^l} &= \sum_{p,q} (\delta_j^l)_{pq} (\mathbf{x}_i^{l-1})_{pq} \\ &= \sum_{p,q} (\delta_j^{l+1} * W_{j\cdot}^{l+1})_{pq} M_{pq} (\mathbf{x}_i^{l-1})_{pq}, \end{aligned} \tag{7}$$

and compute the gradients for bias as

$$\begin{aligned} \frac{\partial E}{\partial b_{ij}^l} &= \sum_{p,q} (\delta_j^l)_{pq} \\ &= \sum_{p,q} (\delta_j^{l+1} * W_{j\cdot}^{l+1})_{pq} M_{pq}, \end{aligned} \tag{8}$$

where $W_{j\cdot}^{l+1}$ denotes the filters of layer $l+1$ that are used to calculate $U_j^l$, and $(\mathbf{x}_i^{l-1})_{pq}$ denotes the patch centered $(p,q)$ on $U_i^{l-1}$. With Eq. 7 and Eq. 8, the proposal map $M$ which indicates the objectness confidence of an image combines with the gradient maps in the weakly supervised activation procedure, driving SPN to learn more useful patterns.

For weakly supervised object localization, we calculate the response map $R_c$ for the $c$-th class, similar to [36], $R_c = \sum_k w_{k,c} \cdot \hat{U}_k \circ M$ where $\hat{U}_k$ is the $k$-th feature map of the last convolutional layer, $w_{k,c}$ is the weight value of the fully connected layer which connects the $c$-th output node and the $k$-th feature vector, Fig. 3.

## 4. Experiment

We upgrade state-of-the-art CNN architectures, *e.g.*, VGG16 and GoogLeNet, to SPNs, and evaluate them on popular benchmarks. In Sec. 4.1, we compare SPN with conventional object proposal methods, showing that it can generate high-quality proposals with negligible computational overhead. In Sec. 4.2, on a weakly supervised point-based object localization task, we demonstrate SPNs can learn better object-centric filters, which produce precise responses on class-specific objects. In Sec. 4.3, SPNs are further tested on a weakly supervised object bounding box localization task, validating its capability of discovering more fine-detailed visual evidence in complex cluttered scenes. In Sec. 4.4, the significant improvement of classification performance on PASCAL VOC [10] (20-classes, ~10k images), MS COCO [16] (80-classes, ~160k images), and ImageNet [23] (1000-classes, ~1300k images), shows the superiority of SPNs beyond weakly supervised object localization tasks[2]. We train SPNs using SGD with cross-

[2]Please refer to supplementary materials for more results.

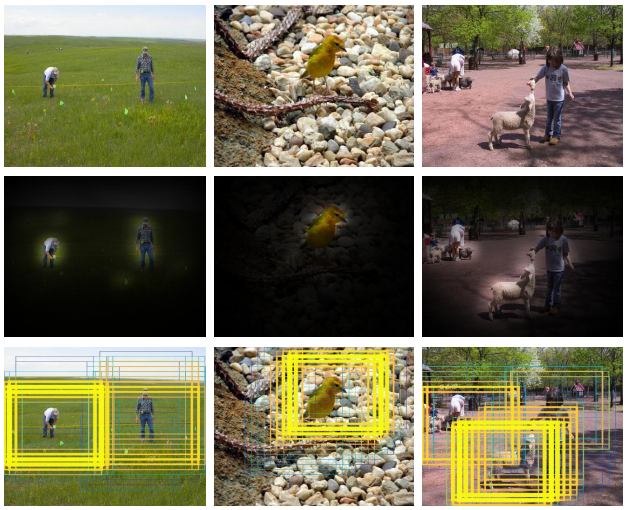


Figure 5. Proposal examples. The first row presents input images. The second row presents proposal coupled images, by composing the proposal map with the original images. The third row shows top-100 scored receptive fields according to the proposal map. Best viewed in color.

| Method | ObjectEnergy(%) | Time(ms) |
|---|---|---|
| Selective Search [30] | 53.7 | 2000 |
| EdgeBoxes [37] | 58.8 | 200 |
| RPN (supervised) [22] | **63.3** | 10.5 |
| SPN (weakly supervised) | 62.2 | **0.9** |

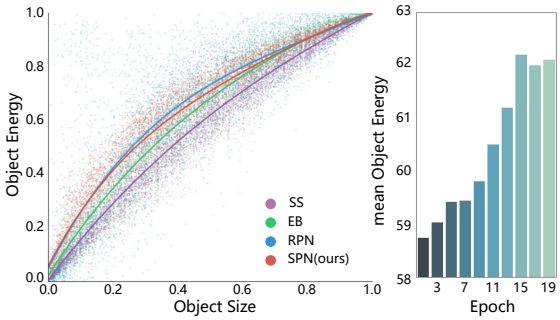
Table 1. Proposal quality evaluation on VOC2007 test set. The Object Energy in the second column indicates the percentage of spotlighted object areas. Note that RPN is learned with object bounding box annotations (supervised) while SPN is learned with image label annotations (weakly supervised). The third column describes the average time cost per image. RPN and SPN are tested with a NVIDIA Tesla K80 GPU while Selective Search and EdgeBoxes are tested on CPU due to algorithm complexity.

entropy loss. We use a weight decay of 0.0005 with a momentum of 0.9 and set the initial learning rate to 0.01.

### 4.1. Proposal Quality

On the VOC2007 dataset, we assess the quality of proposals by an Object Energy metric defined below. For the compared Selective Search [30], EdgeBoxes [37] and RPN [22] methods, the energy value of a pixel is the sum of scores of the proposal boxes that cover the pixel. Therefore, all objectness values in an image constitute an energy map that indicates the informative object regions predicted by the method. For the SPN, we produce Object Energy maps by rescaling proposal maps to the image size, Fig. 5. We further normalize each energy map and compute the sum of Object Energy of pixels those fall into ground-truth bounding boxes as the Object Energy.

It can be seen from the definition that the Object Energy values range in $[0.0, 1.0]$, which indicates how many informative object areas in the image are spotlighted by the method. The second column in Tab. 1 demonstrates that the proposals generated by SPN are of high-quality. The Object Energy of SPN proposals is significantly larger than those of Selective Search and EdgeBoxes, which usually produce redundant proposals and cover many background regions. Surprisingly, The Object Energy of SPN proposals obtained by weakly supervised learning is comparable to that of supervised RPN method (62.2% vs. 63.2%). It can be seen in Fig. 6(a) that the proposed SPN can spotlight small objects significantly better than the Selective Search and EdgeBoxes methods, despite that the proposal maps are based on low-resolution deep feature maps. Fig. 6(b) demonstrates that the SPN proposals can iteratively evolve and jointly optimize with network filters during the end-to-end training. Moreover, the implementation of SPN is simple and naturally compatible with GPU parallelization. It can be seen from the third column of Tab. 1 that the proposed SP module can introduce weakly supervised object proposal to CNNs in a nearly cost-free manner.




(a) Object Energy distribution (b) Object Energy evolution

Figure 6. (a) Object Energy curves. The x-coordinate is the ratio between the object area to the image size, and y-coordinate is the Object Energy. The curves are produced by using a 3-polynomial regression on the dots, each of which denotes an image. (b) Evolution of Object Energy during the learning procedure. Best viewed zooming on screen.

### 4.2. Pointing Localization

**Pointing without prediction.** To evaluate whether the proposed SPN can learn more discriminative filters which are effective to produce accurate response maps, we test it on the weakly supervised pointing task. We select three successful CNNs, including CNN-S [7], VGG16 [25], and GoogLeNet [28] and upgrade them to SPNs by inserting the SP module after their last convolution layers, Fig. 3. All SPNs are fine-tuned on the VOC2007 training set with same hyper-parameters, and we calculate the response maps as described in Sec. 3.3 with ground-truth labels for pointing localization. Following the setting of c-MWP [35], a state-of-the-art method, we calculate the accuracy of pointing lo-

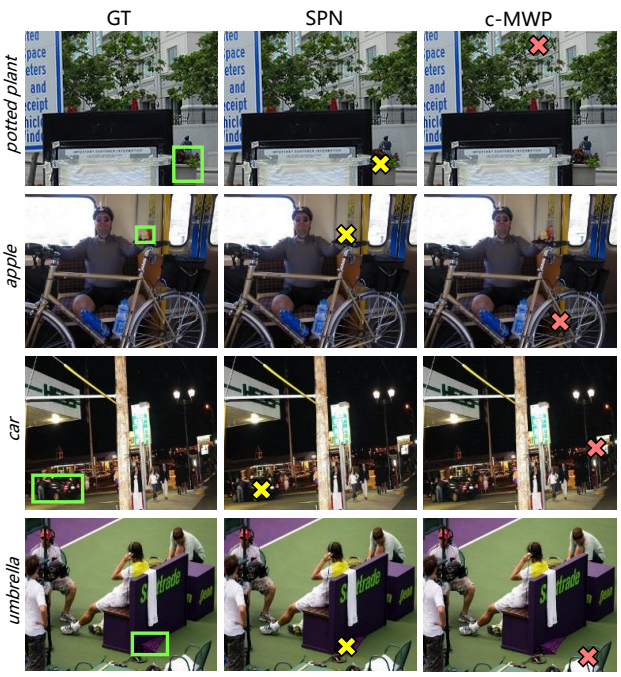



Figure 7. Examples of pointing localization, which shows that SPN is effective in complex scenes: a) Noisy co-occurrence patterns, *e.g.*, leaves for "potted plant". b) Small objects, *e.g.*, "apple" in hand. c) Cluttered backgrounds, *e.g.*, "car" on the street. d) Infrequent form, *e.g.*, closed "umbrella". Best viewed in color.

calization as below: a hit is counted if the pixel of maximum response falls in one of the ground truth bounding boxes of the cued object category within 15 pixels tolerance. Otherwise, a miss is counted. We measure the per-class localization accuracy by $Acc = \frac{Hits}{Hits+Misses}$. The overall results are the mean value of per-class point localization accuracy.

For the VOC2007 dataset, we use two test sets, *i.e.*, **All** and Difficult (**Diff.**) [35]. **All** means the overall test set and **Diff.** means a difficult subset which has mixed categories and contains small objects. As shown in Tab. 2, upgrading conventional CNNs to SPNs brings significant performance improvement. Specifically, the SP-VGGNet outperforms c-MWP by 7.5% (87.5 % vs 80.0 %) for **All** and 11.3% (78.1% vs 66.8%) for **Diff.**. The SP-GoogLeNet outperforms c-MWP by 3.1% and 6.8% for **All** and **Diff.**, respectively. The significant improvement of pointing localization performance validates the effectiveness of the SP module for guiding SPNs to learn better object-centric filters, which can pick up accurate object responses.

We made multiple observations in Tab. 2. 1). SP-VGGNet has better performance than SP-GoogLeNet on pointing localization. The reason can be that the receptive fields of SP-VGGNet are smaller than that of SP-GoogLeNet. Without much overlap between receptive fields, the objectness propagation in SP module can be more effective. 2). The accuracy improvement on **Diff.** is larger

| Method | CNN-S | VGG16 | GoogLeNet |
|---|---|---|---|
| Center | 69.5/42.6 | 69.5/42.6 | 69.5/42.6 |
| Grad [24] | 78.6/59.8 | 76.0/56.8 | 79.3/61.4 |
| Deconv [33] | 73.1/45.9 | 75.5/52.8 | 74.3/49.4 |
| LRP [1] | 68.1/41.3 | - | 72.8/50.2 |
| CAM [36] | - | - | 80.8/61.9 |
| MWP [35] | 73.7/52.9 | 76.9/55.1 | 79.3/60.4 |
| c-MWP [35] | 78.7/61.7 | 80.0/66.8 | 85.1/72.3 |
| SPN | **81.8/66.7** | **87.5/78.1** | **88.2/79.1** |

Table 2. Pointing localization accuracy (%) on VOC2007 test set (**All/Diff.**). **Center** is a baseline method which uses the image centers as estimation of object centers.

| Method | mAP (%) | |
|---|---|---|
| Dataset | VOC | COCO |
| Oquab *et al.* [20] | 74.5 | 41.2 |
| Sun *et al.* [27] | 74.8 | 43.5 |
| Bency [2] | 77.1 | 49.2 |
| SPN | **82.9** | **55.3** |

Table 3. Mean Average Precision (mAP) of location prediction on VOC2012 val. set and COCO2014 val. set.

than that on **All**, which shows that the proposal functionality of SPNs is particularly effective in cluttered scenes.

**Pointing with prediction.** We further test SPN on a more challenging pointing-with-prediction task. The task requires the network output not only the correct prediction of the presence/absence of the object categories in test images, but also the correct pointing localization of objects, *i.e.*, the point of maximum response falls in one of the ground truth bounding boxes within 18 pixels tolerance [20].

We upgrade a pre-trained VGG16 model to SPN and respectively fine-tune it on VOC2012 and COCO2014 dataset for 20 epochs. Results are reported in Tab. 3. Without multi-scale setting, SPN outperforms the state-of-the-art method [2] by a significant margin (5.8% mAP for VOC2012, 6% mAP for COCO2014). This evaluation demonstrates that the Soft Proposal module endows CNNs accurate localization capability while keeping its classification ability. In Sec. 4.4, we will show that upgrading CNNs to SPNs can even improve the classification performance.

### 4.3. Bounding Box Localization

Although without object-level annotations involved in the learning phase, our method can also be used to estimate object bounding boxes with the help of response maps. We calculate each response map with ground truth labels and convert them to binary maps with the mean value as thresholds. We then rescale them to the original image size and extract the tightest box covering the foreground pixels as the predicted object bounding box.

The Correct Localization (CorLoc) metric [9] is used

| Method | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | persn | plant | sheep | sofa | train | tv | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bilen *et al.* [4] | 66.4 | 59.3 | 42.7 | 20.4 | 21.3 | 63.4 | 74.3 | 59.6 | 21.1 | 58.2 | 14.0 | 38.5 | 49.5 | 60.0 | 19.8 | 39.2 | 41.7 | 30.1 | 50.2 | 44.1 | 43.7 |
| Wang *et al.* [31] | 80.1 | 63.9 | 51.5 | 14.9 | 21.0 | 55.7 | 74.2 | 43.5 | 26.2 | 53.4 | 16.3 | 56.7 | 58.3 | 69.5 | 14.1 | 38.3 | 58.8 | 47.2 | 49.1 | 60.9 | 48.5 |
| Cinbis *et al.* [8] | 65.3 | 55.0 | 52.4 | **48.3** | 18.2 | 66.4 | **77.8** | 35.6 | 26.5 | **67.0** | 46.9 | 48.4 | 70.5 | 69.1 | 35.2 | 35.2 | **69.6** | 43.4 | 64.6 | 43.7 | 52.0 |
| WSDDN [5] | 65.1 | 58.8 | 58.5 | 33.1 | **39.8** | 68.3 | 60.2 | 59.6 | **34.8** | 64.5 | 30.5 | 43.0 | 56.8 | 82.4 | 25.5 | **41.6** | 61.5 | 55.9 | 65.9 | 63.7 | 53.5 |
| ContextLoc [14] | 83.3 | **68.6** | 54.7 | 23.4 | 18.3 | **73.6** | 74.1 | 54.1 | 8.6 | 65.1 | 47.1 | 59.5 | 67.0 | **83.5** | 35.3 | 39.9 | 67.0 | 49.7 | 63.5 | **65.2** | 55.1 |
| **SP-VGGNet** | **85.3** | 64.2 | **67.0** | 42.0 | 16.4 | 71.0 | 64.7 | **88.7** | 20.7 | 63.8 | **58.0** | **84.1** | **84.7** | 80.0 | **60.0** | 29.4 | 56.3 | **68.1** | **77.4** | 30.5 | **60.6** |

Table 4. Correct Localization rate (CorLoc [9]) on the positive trainval images of the VOC2007 dataset (%).


Figure 8. Bounding box localization results on the VOC2007 test set. By activating fine-detailed evidence like arm or leg for "person", paw for "cat", and texture fragments for "sofa", the estimated bounding boxes are more precise than those by WSDDN.

to evaluate the bounding box localization performance. It can be seen in Tab. 4 that the mean CorLoc of our method outperforms the state-of-the-art ContextLoc method [14] by about 5%. Surprisingly, on the "dog", "cat", "horse", and "person" classes, SPN outperforms the compared method up to 20-30%. It can be seen from Fig. 8 that the conventional method tends to use the most discriminative part for each category, *e.g.*, faces, while SPN can discover more fine-detailed object evidence, *e.g.*, hands and legs, thanks to the objectness prior introduced by the SP module. On the "sofa" and "table" classes, our method outperforms other methods by 10%, demonstrating the capability of SPN to correctly localize the occluded objects, Fig. 8, which shows that the graph propagation in the Soft Proposal Generation step helps to find object fragments of similar appearance.

### 4.4. Image Classification

Although to predict the presence/absence of object categories in an image does not require accurate located and comprehensive visual cues, the proposal functionality of SPNs which highlights informative regions while suppressing disturbing backgrounds during training should also benefit the classification performance.

We use GoogLeNetGAP [36], a simplified version of GoogLeNet, as the baseline. By inserting SP module after the last convolution layer, the GoogLeNetGAP is upgraded to a SPN. The SPN is trained on the ILSVRC2014 dataset, *i.e.*, ImageNet, for 90 epochs with the SGD method. It can be seen in the second column of Tab. 6 that the SPN significantly outperforms the baseline GoogLeNetGAP by 1.5%, which shows that the SPNs can learn more informative feature representation. We then fine-tune each trained model on COCO2014 and VOC2007 by 50 and 20 epochs to assess the generalization capability of SPN. As shown in the third column of Tab. 6. SP-GoogLeNetGAP surpasses the baseline by a large margin, *e.g.*, 4.5% on VOC2007. This further demonstrates that the weakly supervised object proposal is effective for both localization and classification.

| Method | CAM | c-MWP | MWP | Fb[35] | SPN |
|---|---|---|---|---|---|
| Error (%) | 48.1 | 57.0 | 38.7 | 38.8 | **36.3** |

Table 5. Bounding box localization errors on ILSVRC2014 val. set.

| Method | ImageNet | COCO | VOC |
|---|---|---|---|
| GoogLeNetGAP[36] | 35.0/13.2 | 54.4 | 83.4 |
| SP-GoogLeNetGAP | **33.5/12.7** | **56.0** | **84.2** |

Table 6. Classification results. The second column is the top-1/top-5 error rate (%) on ILSVRC2014 val. set. The third and fourth column are mAP (%) on VOC2007 test set and COCO val. set.

## 5. Conclusions

In this paper, we proposed a simple yet effective technique, Soft Proposal (SP), to integrate nearly cost-free object proposal into CNNs for weakly supervised object localization. We designed the SP module to upgrade conventional CNNs, *e.g.*, VGG and GoogLeNet, to Soft Proposal Networks (SPNs). In SPNs, iteratively evolved object proposals are generated based on the deep feature maps then projected back, leading filters to discover more fine-detailed evidence through the unified learning procedure. SPNs significantly outperforms state-of-the-art methods on weakly supervised localization and classification tasks, demonstrating the effectiveness of coupling object proposal with network learning.

## Acknowledgements

The authors are very grateful for support by NSFC grant 61671427, BMSTC grant Z161100001616005.

## References


[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 7

[2] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. S. Manjunath. Weakly supervised localization using deep feature maps. In *European Conference on Computer Vision (ECCV)*, pages 714–731, 2016. 3, 7

[3] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference (BMVC)*, volume 3, 2014. 2, 3

[4] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1081–1089, 2015. 8

[5] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2846–2854, 2016. 1, 2, 3, 8

[6] Z. Bolei, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference (BMVC)*, 2014. 6

[8] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(1):189–203, 2017. 1, 2, 3, 8

[9] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision (IJCV)*, 100(3):275–293, 2012. 7, 8

[10] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 5

[11] R. B. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 1

[12] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 1

[13] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Neural Information Processing Systems (NIPS)*, pages 545–552, 2006. 4

[14] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision (ECCV)*, pages 350–365, 2016. 1, 2, 3, 8

[15] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *Neural Information Processing Systems (NIPS)*, pages 1189–1197, 2010. 2

[16] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 5

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37, 2016. 1

[18] L. Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2:1–46, 1993. 4

[19] M. E. Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008. 4

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–694, 2015. 2, 3, 7

[21] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 1

[22] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 1, 2, 6

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5

[24] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations (ICLR Workshop)*, 2013. 7

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[26] H. O. Song, R. B. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *International Conference on Machine Learning (ICML)*, pages 1611–1619, 2014. 2, 3

[27] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. D. Bourdev. Pronet: Learning to propose object-specific boxes for cascaded neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3493, 2016. 2, 3, 7

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 6

[29] E. W. Teh, M. Rochan, and Y. Wang. Attention networks for weakly supervised object localization. In *British Machine Vision Conference (BMVC)*, 2016. 3

[30] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 2, 6

[31] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision (ECCV)*, pages 431–

445, 2014. 1, 8

[32] Q. Ye, T. Zhang, Q. Qiu, B. Zhang, J. Chen, and G. Sapiro. Self-learning scene-specific pedestrian detectors using a progressive latent model. *CoRR*, abs/1611.07544, 2016. 2

[33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2014. 7

[34] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han. A self-paced multiple-instance learning framework for co-saliency detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 594–602, 2015. 1

[35] J. Zhang, Z. L. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision (ECCV)*, pages 543–559, 2016. 3, 6, 7, 8

[36] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1, 2, 3, 5, 7, 8

[37] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, pages 391–405, 2014. 2, 6