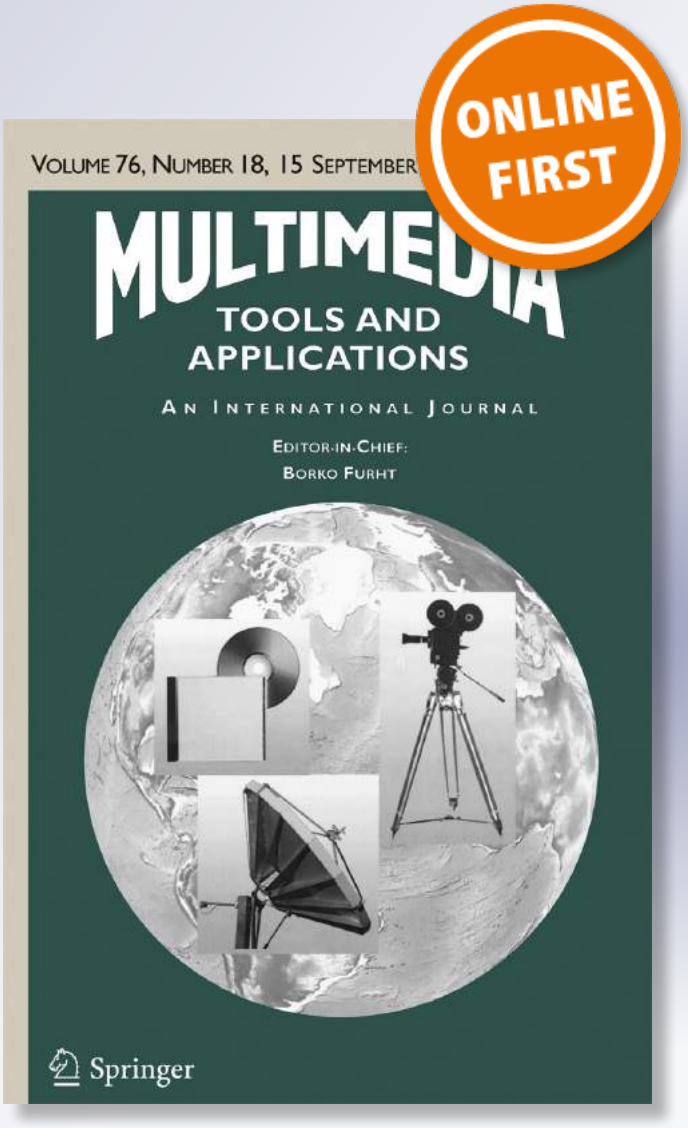
Weakly supervised detection with decoupled attention-based deep representation

Wenhui Jiang, Zhicheng Zhao & Fei Su

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-017-5087-x



Springer

Multimed Tools Appl
DOI 10.1007/s11042-017-5087-x

Weakly supervised detection with decoupled attention-based deep representation

Wenhui Jiang[1] · **Zhicheng Zhao**[1,2] · **Fei Su**[1,2]

Received: 8 March 2017 / Revised: 4 August 2017 / Accepted: 7 August 2017

Abstract Training object detectors with only image-level annotations is an important problem with a variety of applications. However, due to the deformable nature of objects, a target object delineated by a bounding box always includes irrelevant context and occlusions, which causes large intra-class object variations and ambiguity in object-background distinction. For this reason, identifying the object of interest from a substantial amount of cluttered backgrounds is very challenging. In this paper, we propose a decoupled attention-based deep model to optimize region-based object representation. Different from existing approaches posing object representation in a single-tower model, our proposed network decouples object representation into two separate modules, i.e., image representation and attention localization. The image representation module captures content-based semantic representation, while the attention localization module regresses an attention map which simultaneously highlights the locations of the discriminative object parts and down weights the irrelevant backgrounds presented in the image. The combined representation alleviates the impact from the noisy context and occlusions inside an object bounding box. As a result, object-background ambiguity can be largely reduced and background regions can be suppressed effectively. In addition, the proposed object representation model can be seamlessly integrated into a state-of-the-art weakly supervised detection framework, and the entire model can be trained end-to-end. We extensively evaluate the detection performance on the PASCAL VOC 2007, VOC 2010 and

✉ Wenhui Jiang
jiang1st@bupt.edu.cn

Zhicheng Zhao
zhaozc@bupt.edu.cn

Fei Su
sufei@bupt.edu.cn

[1] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China

[2] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China

Published online: 16 August 2017

VOC2012 datasets. Experimental results demonstrate that our approach effectively improves weakly supervised object detection.

Keywords Weak supervision · Object detection · Deep learning · Attention model

1 Introduction

Object detection aims to recognize and localize every object instance within a given image. It has a variety of applications such as autonomous vehicular systems [15, 56], object retrieval [21, 54, 55, 57–59], event detection [6, 8, 9, 26, 27] and remote sensing [18]. The state-of-the-art approaches [12, 24, 32, 33] typically assume a large number of precise bounding box annotations is available for training the object detectors. However, manually labeling annotation is time consuming [31, 44]. For this reason, scaling up these object detection algorithms to a large number of object categories is very challenging in practice.

In this paper, we are interested in weakly supervised detection (WSD) where only image-level labels are available for training. WSD alleviates the problem in lack of instance-level annotations. This particular setting is important for large scale practical applications, because labeling image-level annotations requires only a fraction of efforts compared with bounding box annotations, and even readily available from the Internet.

However, detecting objects with only image-level supervision remains a challenging task. Typically, an object region delimited by a bounding box always includes irrelevant context and occlusions, especially for the highly deformable and irregular objects such as animals. The cluttered backgrounds cause large object variations as well as visual ambiguity between target objects and background regions. As a consequence, the learned object detector cannot localize the targets accurately. Convolutional Neural Network (CNN) based methods build region-level object representation with single-tower architecture (i.e., a single-path network as in Fast RCNN [17] and R-FCN [12]) do not solve this problem explicitly. In this context, optimizing object representation [5, 7] to reduce object-background ambiguity is important for a weakly supervised detection algorithm.

In this work, we are inspired from the way humans perform image annotation. While trying to predict the label for a whole image, humans tend to selectively focus on the relevant object parts instead of putting their attention on the entire image at once [22, 42] (see Fig. 1 for examples). Such attentive mechanism not only filters out irrelevant background clutters, but

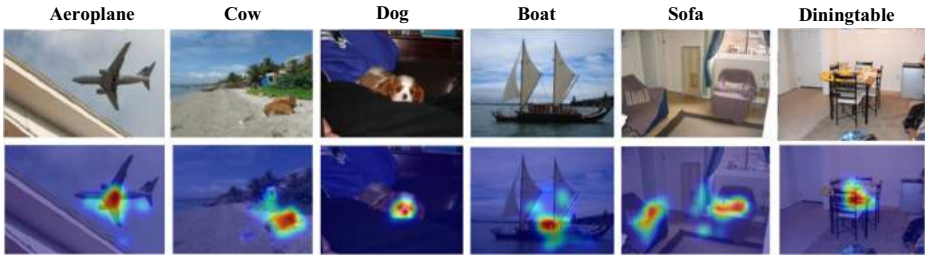


Fig. 1 Illustrations of human visual attentions. Line one: raw images contain certain objects (e.g., boat, cat). Line two: human eye fixations when searching for these objects. The fixation points locate at the discriminative parts of the objects

also provides the possibility of instance-level object localization [49, 50, 52]. Upon this, we propose a decoupled attention-based object representation model to address weakly supervised detection. A brief illustration of our model is shown in Fig. 2. It stands out from the traditional single-tower model [17] because our proposed network decouples object representation into two separate modules, i.e., image representation and attention localization. In particular, the image representation module captures content-based image representation with image feature maps, and the attention localization module regresses an attention map characterizing the locations of discriminative object parts presented in the image. Then the region-level object representation is obtained by applying the attention map on the image feature maps through *attention pooling*. Intuitively, the attention map depresses the activations from background regions in the image feature maps. As a result, the background regions are represented by feature maps with very small activations, therefore are more separable with foreground objects. To train this representation model, we simply connect it to a state-of-the-art weakly supervised object localization model [3]. The learning of both object representation and object localization are supervised by image-level annotation only, and can be trained efficiently end-to-end.

The major contribution of this work is the decoupled attention-based object representation model. Such model provides at least two advantages: 1) For region-level representation, the attention map reduces the impact from cluttered backgrounds, and makes region-level object representation more discriminative. 2) The extracted attention maps allow us to diagnostically visualize the importance of different visual clues to image annotation.

We evaluate and compare the performance of our model on three challenging datasets: the PASCAL VOC 2007, VOC 2010 and VOC 2012 [14]. For the three datasets, we only use image-level class labels for training. The experimental results show that our method outperforms well-established baselines, especially on highly deformable objects such as cats and irregular objects such as sofas.

The rest of the paper is organized as follows. We briefly review the related studies in Section 2. Then we provide the details of the proposed system in Section 3. In Section 4, we apply the model to the PASCAL VOC datasets and compare the performance with state-of-the-art methods. We draw conclusions in Section 5.

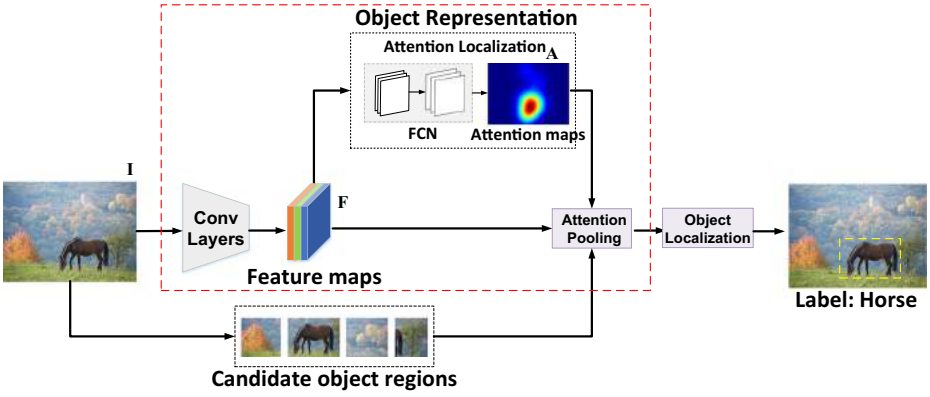


Fig. 2 The overall architecture of our model. The attention-based object representation part (shown in red rectangle) extracts content-based feature maps $\boldsymbol{F}$ and an object-aware attention map $\boldsymbol{A}$. The combined representation improves bounding box-based object representation by alleviating noisy context and occlusions. We apply a state-of-the art object localization model (weakly supervised deep detection network) for subsequent region mining and classification under weak supervision

2 Related work

2.1 Multiple instance learning (MIL)

A number of recent works explore weakly-supervised object detection by using multiple instance learning framework. Under this framework, a set of candidate object regions (including target objects and noisy backgrounds) are first extracted from each image. Then the learning process alternates between estimating CNN-based object detectors and updating the training set using the detectors. However, MIL-based approaches are sensitive to the initialization of the network and the quality of the training samples. To address these issues, Song et al. [40, 41] built a graph-based model to select a subset of visual similar candidate regions among positive images for initialization, and found co-occurring part configurations to improve the model's robustness. Bilen et al. [4] proposed a formulation that jointly learns a discriminative model and enforces the similarity of the selected object regions via a discriminative convex clustering algorithm. Ren et al. [34] introduced a bag-splitting algorithm that iteratively removes negative instances from positive bags for better convergence. Wang et al. [45] improved the robustness of region representation by clustering all candidate regions into groups, then selected object category among groups. Cinbis et al. [11] proposed a multi-fold split scheme to deal with the training data. Shi et al. [37] selected training regions based on size prior. Zhang et al. [51] addressed this problem with self-paced curriculum learning. More recently, Bilen et al. [3] proposed a weakly supervised deep detection network (WSDDN) which consists of two streams for object detection – one stream for region classification and the other for object localization.

Above mentioned models achieve promising performance, however, they share a common drawback: representing an object with a bounding box would inevitably include background clutters and occlusions. The noisy clutters may impact the robustness of the learned object representation. In this paper, we aim at improving weakly supervised detection by proposing a decoupled attention-based object representation. The decoupled architecture has two merits: 1) enables us to distill clutter backgrounds and unrelated visual clues within an object bounding box; 2) provides a discriminative feature representation for object localization. Our object representation model can be integrated with a variety of existing weakly supervised detection models. In this paper, we simply integrate our model with weakly supervised deep detection network (WSDDN), which is a state-of-the-art model for WSD.

2.2 CNNs for WSD

As another line of research, Oquab et al. [29] showed that CNNs for image classification automatically learn object activation maps, where the target objects can be coarsely localized. However, they do not quantitatively evaluate the performance of object detection. In the follow-up work, Oquab et al. [30] localized the object of interest by examining the maximal response of the object activation maps. Zhou et al. [53] applied a simple thresholding technique to segment the activation maps, and generated object bounding boxes on the largest connected regions in the feature maps. Bency et al. [2] adopted a beam-search strategy to localize objects on the activation maps. In these researches, the activation maps can be learned end-to-end from image-level labels. However, compared with MIL-based models [3, 11, 45], the accuracy of [2, 30] is relatively low, since they do not capture object-level representation in the training process. Moreover, these methods [2, 30] require a separate post-processing step to obtain the final localization.

2.3 Attention-based models

Attention models have been successfully adopted in many computer vision tasks, including digital character detection [1, 28], image captioning [47, 48] and visual questioning answering [38, 46]. The key idea is to assign different weights to different image locations according to their relative importance to the output (categories, captions, answers). Such mechanism is appealing since it adaptively focuses on the region of interest, while ignore the irrelevant image clues. For example, Minh et al. [28] used a kind of hard attention to select image regions for digits recognition. However, their networks are not differentiable. Besides, their networks are relatively simple and are only applied to toy datasets such as MNIST and SVHN. Xu et al. [47] proposed to adaptively combine the feature vectors from different image locations into a single feature vector, which is later utilized to generate image caption. More recently, Xu et al. [46] and Sharma et al. [36] adopted similar strategies to deal with visual question answering and action recognition.

Our work follows the concept of visual attention, since we believe it is beneficial for WSD if we can filter out noisy image regions. To the best of our knowledge, we are the first to apply attention-based models to optimize region-based object representation for weakly supervised object detection.

3 Proposed method

3.1 Overview

The proposed framework is illustrated in Fig. 2. It takes an image $\boldsymbol{I}$ and a set of candidate regions $\boldsymbol{B}$ (extracted with Selective Search [43]) as inputs. It mainly consists of two parts: 1) a decoupled attention-based object representation part and 2) an object localization part. In the object representation part, a fully convolutional net is employed to extract image feature maps as image representation first. Then an attention localization net regresses an attention map characterizing the locations of objects presented in the image. The attention map is further applied on the image feature maps through *attention pooling* to formulate object representations for each candidate region. In the object localization part, we simply apply a state-of-the art model [3] that maps each candidate region into a class probability vector based on the formed representation. The learning of the two parts is supervised by image-level annotation only. In the following subsections, we will explain both parts in more details.

3.2 Object representation

3.2.1 Image representation

Our model first takes an image $\boldsymbol{I}$ (arbitrary size) as the input, and then extracts feature maps $\boldsymbol{F}$ via a stack of convolutional layers. In particular, we use 5 convolutional layers (Conv1 - Conv5) with AlexNet [23] (see Fig. 3 for detailed architecture). $\boldsymbol{F}$ is a 3-D tensor of $W \times H \times D$ dimensions, where D is the number feature maps (i.e., multi-dimensional filters), $W \times H$ represents the spatial resolution. Each location inside $\boldsymbol{F}$ corresponds to a certain subregion of $\boldsymbol{I}$ thus $\boldsymbol{F}$ still retains the crucial spatial information. Because of the employment of pooling layers and the large stride of convolutional layers, the spatial decimation factor of the convolutional layers is 16.

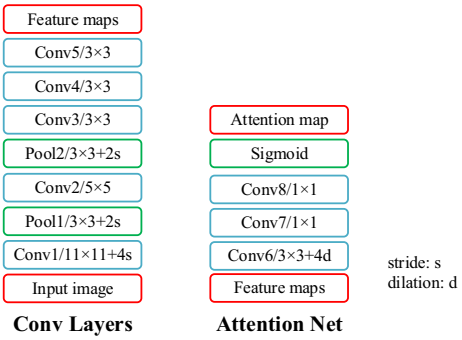


Fig. 3 Detailed architecture of the object representation part

3.2.2 Attention localization

The attention localization module is also a fully convolutional sub-network. Given the image feature maps $\boldsymbol{F}$ as the input, this sub-net predicts a single attention map $\boldsymbol{A}$ which puts more weights on the locations where target objects appear. In the example shown in Fig. 2, only these pixels, which indicate the presence of a horse are highlighted. Specifically, this sub-net mainly contains three convolutional layers (see Fig. 3 for detailed architecture). The first two layers (Conv6 and Conv7) are converted from the fully connected layers from a base model (AlexNet) as in [25]. In order to retain a large spatial resolution of the output attention map, we remove pool5 layer from the original AlexNet architecture, and employ 'atrous algorithm' [10] on Conv6 to compensate for the loss of Field of View (FOV). The last layer (Conv8) reduces the channel dimension to 1. The convolutional operation is padded to ensure $\boldsymbol{A}$ has the same spatial resolution as $\boldsymbol{F}$. In the end, a sigmoid layer is appended to the sub-net to normalize each value to range [0, 1].

3.2.3 Attention pooling

Given the image feature map $\boldsymbol{F}$, the attention map $\boldsymbol{A}$, and one candidate object region b_i, the *attention pooling* layer aggregates $\boldsymbol{F}$ with $\boldsymbol{A}$ inside b_i into ROI attentive feature maps, and subsequently converts the maps into small fix-sized feature maps (See Fig. 4 for visual

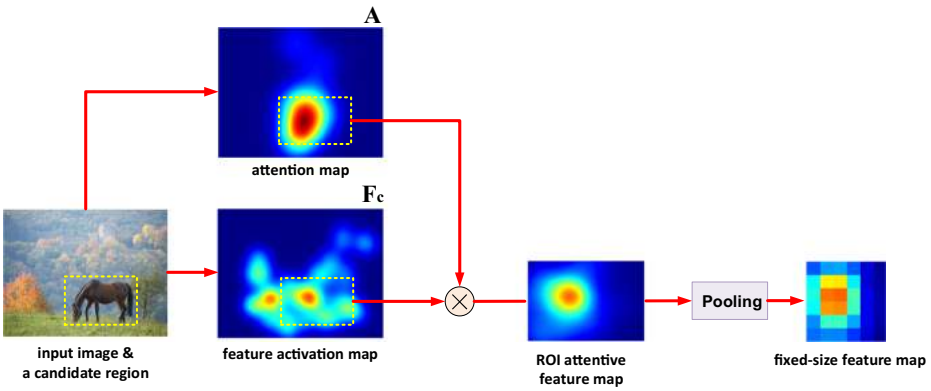


Fig. 4 Visualized example of attention pooling. The attention map highlights the discriminative parts of the target object, therefore makes object representation robust

illustration). More specifically, we divide each candidate region into k × k bins by a regular grid. Assume b_i is with size w × h, a bin is of a size $\frac{w}{k} \times \frac{w}{k}$. Inside the (i,j)-th bin ($0 \le i, j \le k-1$), we define an attention pooling operation that pools over each feature map individually:

$$r_c(i,j) = \max_{(x,y) \in bin(i,j)} \{a(x,y) \cdot f_c(x,y)\} \tag{1}$$

Here $r_c(i,j)$ is the pooled response in the (i,j)-th bin for the c-th feature map, $a(x,y)$ is the value in the (x,y)-th location of the attention map, $f_c(x,y)$ is the value in the (x,y)-th location of the c-th feature map. The (i,j)-th bin spans $\left[\text{floor}\left(i\frac{h}{k}\right) \le x < ceil\left((i+1)\frac{h}{k}\right)\right]$ and $\left[\text{floor}\left(i\frac{w}{k}\right) \le y < ceil\left((i+1)\frac{w}{k}\right)\right]$. The output is fixed-size feature maps of dimension k × k × D. Note that the activation maps of the background regions are down-weighted by $\boldsymbol{A}$ in Eq. (1), the foreground objects and the background regions are more separable in the feature space. Therefore, attention pooling is more discriminative than ROI Pooling [17].

3.3 Object localization

Based on the ROI feature maps, the object localization part is designed to map each candidate region into a class confidence vector. We simply apply the object localization part of weakly supervised deep detection network (WSDDN) [3] into our work. As shown in Fig. 5, for each candidate region b_i, we first connect its feature maps to two fully-connected layers (FC6 and FC7) to form a compact semantic feature vector φ_i^{fc7}. Then, the sub-net is branched into two data streams. One stream performs object recognition by associating a class probability score vector $\boldsymbol{p_i}$ for b_i. Specifically, we append an extra fully-connected layers (FC8) to map φ_i^{fc7} to C dimensional outputs φ_i^{fc8} (C is the number of object categories). Afterwards, we apply a Softmax operator to normalize φ_i^{fc8} to class probability vector $\boldsymbol{p_i}$:

$$p_{i,c} = \frac{\exp\left(\varphi_{8i,c}\right)}{\sum_{j=1}^{C} \exp\left(\varphi_{8i,j}\right)} \tag{2}$$

In contrast to the recognition stream, the other stream performs object detection by comparing all image regions and computing a likelihood distribution over them. The layout

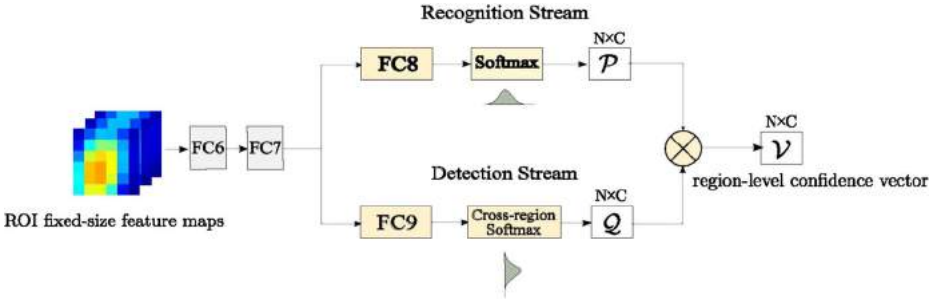


Fig. 5 The detailed architecture of the localization net. It mainly consists of two streams: 1) a recognition stream which estimates a class probability vector for each candidate object region, and 2) a detection stream which selects relevant training samples for each object category

of detection stream is similar to that of the recognition stream, except that a cross-region Softmax operator is utilized to normalize φ_{8i} into a weighting vector $\boldsymbol{q_i}$:

$$q_{i,c} = \frac{\exp\left(\varphi_{9i,c}\right)}{\sum_{h=1}^{N}\exp\left(\varphi_{9h,c}\right)} \tag{3}$$

where N is the total number of candidate regions inside image $\boldsymbol{I}$. The Softmax operator is applied category-by-category across all candidate regions. In principle, the recognition stream predicts which class is associate with a region, whereas the detection stream selects which regions are more likely to contain an informative image fragment.

At last, $\boldsymbol{p_i}$ and $\boldsymbol{q_i}$ are aggregated using element-wise dot product to form region-level confidence vector $\boldsymbol{v_i}$:

$$v_{\boldsymbol{i},\boldsymbol{c}} = p_{i,c} \cdot q_{i,c} \tag{4}$$

3.4 Objective function

We simply assume each category is independent, and define a per-category per-image cross entropy loss $\mathcal{L}$:

$$\mathcal{L}\left(y_{k,c}, \hat{y}_{k,c}\right) = -y_{k,c}\log\hat{y}_{k,c} - \left(1-y_{k,c}\right)\log\left(1-\hat{y}_{k,c}\right) \tag{5}$$

where $y_{k,c} \in \{0, 1\}$ denotes whether the k-th training image contains the c-th object category, $\hat{y}_{k,c}$ is the predicted label which is obtained by aggregating all the region-level confidence vectors $\{\boldsymbol{v_i}\}_{\boldsymbol{i}=1}^{\boldsymbol{N}}$ inside the k-th image (for simplicity of notation we ignore subscript k):

$$\hat{y}_c = \sum_{i=1}^{N} v_{i,c} \tag{6}$$

Denote the parameters of both object representation part and object localization part as vector $\boldsymbol{\theta}$. The the model parameter vector $\boldsymbol{\theta}$ can be learned according to the summed loss:

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{k=1}^{K}\sum_{c=1}^{C}\mathcal{L}\left(y_{k,c}, \hat{y}_{k,c}\right) \tag{7}$$

4 Experiments

4.1 Datasets

We evaluate our model on three representative benchmark datasets, namely the PASCAL VOC2007, VOC2010 and VOC2012 [14]. The VOC 2007 dataset covers 20 different object categories and contains 2501 images for training, 2510 images for validation and 4952 images for testing. VOC 2010 and VOC 2012 shares the same object categories. VOC 2010 includes 4998 images for training, 5105 images for validation and 9637 images for testing. VOC 2012 contains 5717 training images, 5823 validation images and 10,991 test images. These datasets contain both image-level labels and object location annotations. For weak supervision, we only utilize the image-level labels for training. Following [3], we use both train and val splits as the

training set and test split as our test set for VOC 2007 and VOC 2010. For VOC 2012, we use train split as the training set and val split as the test set.

4.2 Evaluation metrics.

We apply two different metrics to evaluate localization performance in this section. First, we quantify localization performance in the training set with the Correct Localization (CorLoc) measure [13]. CorLoc is the percentage of images in which the top bounding-box returned by the model correctly localizes an object of the target class. Second, using mean average precision (mAP) in the test set, we measure the detection performance. For both metrics, we consider that a bounding box is correct if it has an intersection-over-union (IoU) ratio of at least 0.5 with a ground-truth annotation.

4.3 Implementation details

Backbone architecture Our network is built with Caffe [20]. We adopt AlexNet [23] as the backbone architecture. It is pre-trained on ImageNet [35] to initialize the convolutional layers and the two fully connected layers (FC6 and FC7). Conv6 and Conv7 are converted and initialized from FC6 and FC7 respectively. As explained in Section 3, we employ the 'atrous algorithm' on Conv6 to keep a large FOV. We also subsample parameters from FC6 and FC7 — changing the kernel size of Conv6 from 6 × 6 to 3 × 3 and the filter size of Conv6 and Conv7 from 4096 to 1024 — to reduce the model complexity. The rest layers are initialized randomly as in [17].

Training We fine-tune the networks on the target datasets. Each mini-batch contains all the ROIs from one image. We use a weight decay of 0.0005 and a momentum of 0.9. We also adopt multi-scale training. Specifically, the longer side of images is resized to a random scale s ($s \in \{480,576,688,864,1200\}$) while the aspect ratio is kept unchanged. We also apply random horizontal flips to the images for data augmentation. The training process lasts for 20 epochs. All the layers are fine-tuned with the learning rate 0.0005 for the first ten epochs and 0.00005 for the last ten epochs. Fine-tuning our network on PASCAL VOC 2007 takes about 8 h on a NVIDIA Titan Black GPU.

Inference At test time we take the region-level confidence vector $\boldsymbol{v_i}$ as the score. As in [3], we average the outputs of 10 images (i.e. the 5 scales as in training and their flips). The results are post-processed by bounding box voting [16] and non-maximum suppression (NMS) using a threshold of 0.6 IoU.

4.4 Compare with state-of-the-art methods

We compare the detection results of our method with recent state-of-the-art approaches, including MIL-based methods [4, 11, 34, 41, 45] and other CNN-based models [2, 30]. We also remove the attention net as the baseline, which is equivalent to the implementation of [3]. Our re-implementation results in 30.6% in mAP, very close to 31.5% reported in [32]. For fair comparisons, we do not include methods that utilize extra supervisions beyond image-level labels.

In Table 1, we compare the localization results of our method with the state-of-the-arts in terms of CorLoc. On the VOC 2007 trainval set, we achieve a significant improvement by 2.0% over the strong baseline, and 2.2% -10.3% over other competitors. Table 2 shows the detection average precision (AP) performance on the VOC 2007 test set. Our method achieves 32.9% mAP, and outperforms all the alternatives. In particular, we outperform the baseline by 2.4%. We notice that our method significantly outperforms the baseline in several specific categories, e.g., boat (+7.6%), cat (+2.5%), dog (+11.8%), table (+4.0%) and sofa (+18.4%). One common property among these categories is that they are either highly deformable (cats and dogs) or with irregular shape (tables and sofas). As a result, these objects could not be tightly delineated by bounding boxes. The significant improvements on these categories verify that our proposed attention-based model effectively optimizes region-based object representation.

In Table 3, we present our results on the PASCAL VOC 2010 and VOC 2012 datasets. Evaluations are performed via the PASCAL VOC evaluation server. Our method achieves mAP of 31.0% on VOC 2010 and 31.3% on VOC 2012, outperform the state-of-the-art methods.

4.5 Model analysis

To understand our model better, we also carry out several controlled experiments to examine how each component affects the final performance. Experimental results are summarized in Table 4.

Data augmentation is crucial A good WSD algorithm should improve when more training data is supplied. Here we enrich the training data through multi-scale image resizing and

Table 1 Quantitative comparison in terms of correct localization (CorLoc) on VOC 20 07 trainval set

	Bilen [4]	Cinbis [11]	Wang [45]	Ren [34]	Baseline	Ours
Plane	66.4	65.3	80.1	79.2	80.4	**82.9**
Bike	59.3	55	63.9	56.9	69.0	66.7
Bird	42.7	52.4	51.5	46	50.5	**53.8**
Boat	20.4	**48.3**	14.9	12.2	28.7	20.7
Bottle	21.3	18.2	21	15.7	32.1	**32.8**
Bus	63.4	66.4	55.7	58.4	68.5	**75.1**
Car	74.3	**77.8**	74.2	71.4	72.7	73.9
Cat	**59.6**	35.6	43.5	48.6	35.5	39.0
Chair	21.1	**26.5**	26.2	7.2	19.6	18.2
Cow	58.2	67	53.4	**69.9**	65.8	66.4
Table	14	46.9	16.3	16.7	49.8	**52.9**
Dog	38.5	48.4	**56.7**	47.4	37.7	42.8
Horse	49.5	**70.5**	58.3	44.2	66.7	61.2
Motor	60	69.1	69.5	75.5	81.5	**84.7**
Person	19.8	**35.2**	14.1	41.2	13.6	19.4
Plant	39.2	35.2	38.3	39.6	46.9	**49.5**
Sheep	41.7	**69.6**	58.8	47.4	64.9	66.0
Sofa	30.1	43.4	**47.2**	32.3	31.7	41.7
Train	50.2	64.6	49.1	49.8	62.7	**65.0**
tv	44.1	43.7	60.9	18.6	65.9	**71.0**
mAP	43.7	52.0	48.5	43.9	52.2	**54.2**

The entries with the best results are bold-faced

Table 2 Quantitative comparison in terms of average precision (AP) on VOC 2007 test set

	Bilen [4]	Cinbis [11]	Wang [45]	Ren [34]	Song [41]	Baseline	Ours
Plane	46.2	39.3	48.8	41.3	36.3	49.0	**52.9**
Bike	46.9	43.0	41.0	39.7	47.6	**52.2**	51.1
Bird	24.1	28.8	23.6	22.1	23.3	**30.1**	28.6
Boat	**16.4**	20.4	12.1	9.5	12.3	6.3	13.9
Bottle	12.2	8.0	11.1	3.9	11.1	14.4	**15.1**
Bus	42.2	45.5	42.7	41.0	36	53.1	**53.7**
Car	47.1	47.9	40.9	45.0	46.6	49.1	**50.4**
Cat	**35.2**	22.1	35.5	19.1	25.4	12.2	14.7
Chair	7.8	8.4	11.1	1.0	0.7	**12.6**	5.7
Cow	28.3	33.5	34.7	34.0	23.5	35.3	**36.4**
Table	12.7	23.6	18.4	16.0	12.5	32.3	**36.3**
Dog	21.5	29.2	**35.3**	21.3	23.5	17.1	28.9
Horse	30.1	38.5	34.8	32.5	27.9	37.1	**38.8**
Mbike	42.4	47.9	51.3	43.4	40.9	53.4	**54.3**
Person	7.8	20.3	17.2	**21.9**	14.8	3.2	4.5
Plant	**20.0**	**20.0**	17.4	19.7	19.2	19.4	19.7
Sheep	26.8	**35.8**	26.8	21.5	24.2	30.4	32.0
Sofa	20.8	30.8	**32.8**	22.3	17.1	7.6	26.0
Train	35.8	41.0	35.1	36.0	37.7	45.7	**46.9**
tv	29.6	20.1	45.6	18.0	11.6	47.5	**48.2**
mAP	27.7	30.2	30.9	25.4	24.6	30.6	**32.9**

The entries with the best results are bold-faced

random horizontal flipping. Enlarging the training set improves mAP on VOC2007 from 30.6% to 32.9%.

Attention net helps The attention net consistently improves object detection over the baseline in different experimental settings. Without data augmentation, attention net improves detection performance by 2.6%. With data augmentation, attention net improves detection performance by 2.4%.

Atrous is better and faster As described in Section 3, we used the atrous version of AlexNet on attention net. If we do not apply 'atrous algorithm' on Conv6 and not subsample parameters on Conv6 and Conv7, the result is slightly worse (1.0%) while the speed is about 52% slower.

Table 3 Quantitative comparison in terms of detection average precision (AP) on the PASCAL VOC 2010 and VOC 2012

MODEL	mAP	
	VOC 2010	VOC 2012
Oquab [30]	--	11.7
Ren [34]	--	23.8
Cinbis [11]	27.4	--
Bency [2]	--	26.5
Baseline	29.8	30.4
Ours	**31.0**	**31.3**

The entries with the best results are bold-faced

Table 4 Effects of various design choices and components on model performance

Data augmentation		√			√	√
Use attention net			√	√	√	√
Use atrous				√		√
VOC2007 test mAP	28.0	30.5	30.3	30.6	31.9	**32.9**

The entries with the best results are bold-faced

4.6 Visualization analysis

4.6.1 Visualize attention maps

In Fig. 6, we visualize the attention maps learned by the proposed attention net. These specialized maps are expected to be strongly activated at discriminative positions of objects. For example, on a cat, more activations are put on the head and the main body. Also, more attention is put on a bike's wheels and handles. These are consistent with humans' prior knowledge because such parts are most discriminative.

4.6.2 Sample detection results

Visual detection results on the VOC 2007 test set are shown in Fig. 7. We observe that our method could localize objects subject to heavy occlusion and complex background. Even small objects can be accurately discovered. However, several bad cases still exist. In particular, some visual-similar objects are mis-classified, e.g., windows are mis-classified as monitors, and a table is identified as a chair. We believe this kind of mistakes can be corrected with more training data and deeper backbone architecture like VGG16 [39] and ResNet [19].

5 Conclusion

In this paper, we propose a decoupled attention-based deep model to optimize region-level object representation for weakly supervised detection task. The decoupled architecture enables us to distill background clutters and unrelated visual clues within an object bounding box, and to provide discriminative feature representation for object localization. In addition, the proposed object representation model can be seamlessly integrated into a state-of-the-art weakly

Fig. 6 Visualization of attention maps. Red colors indicate large activations while blue colors represent small activations. In the examples, most of the large activations come from discriminative object parts

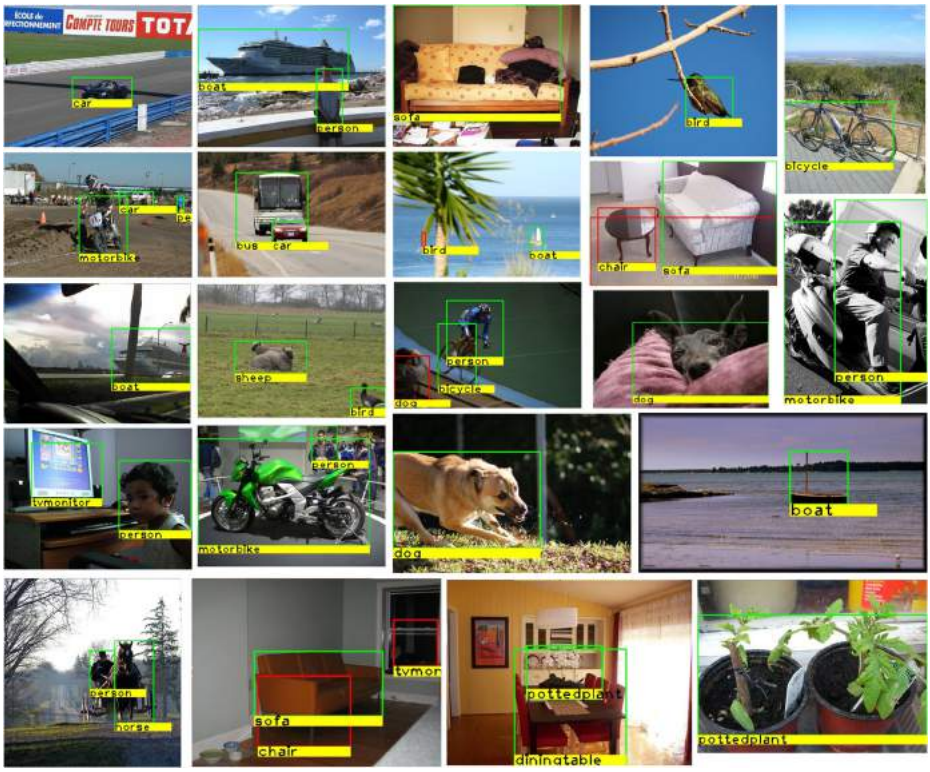


Fig. 7 Detection examples on VOC 2007 test set. Green boxes indicate correct detections; red boxes represent false detections

supervised detection framework, and the entire model can be trained end-to-end. Extensive evaluation results on PASCAL VOC 2007, VOC 2010 and VOC 2012 datasets demonstrate that our approach effectively improves weakly supervised object detection.

Acknowledgements This work is supported by Chinese National Natural Science Foundation under Grants 61471049, 61372169 and 61532018.

References

1. Ba J, Mnih V, Kavukcuoglu K (2015) Multiple object recognition with visual attention. International Conference on Learning Representations, In, pp 1–10
2. Bency AJ, Kwon H, Lee H, Karthikeyan S, Manjunath BS (2016) Weakly supervised localization using deep feature maps. European Conference on Computer Vision
3. Bilen H, Vedaldi A (2016) Weakly supervised deep detection networks. IEEE Conference on Computer Vision and Pattern Recognition
4. Bilen H, Pedersoli M, Tuytelaars T (2015) Weakly supervised object detection with convex clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp 1081–1089
5. Chang X, Yang Y (2016) Semi-supervised feature analysis by mining correlations among multiple tasks. IEEE Trans Neural Netw Learn Syst. doi:10.1109/TNNLS.2016.2582746

6. Chang X, Yu Y, Yang Y, Xing EP (2016) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Anal Mach Intell 39:1617–1632. doi:10.1109/TPAMI.2016.2608901
7. Chang X, Nie F, Wang S, Yang Y, Zhou X, Zhang C (2016) Compound rank-k projections for bilinear analysis. IEEE Trans Neural Netw Learn Syst 27:1502–1513
8. Chang X, Ma Z, Lin M, Yang Y, Hauptmann AG (2017) Feature interaction augmented sparse learning for fast Kinect motion detection. IEEE Trans Image Process 26:3911–3920
9. Chang X, Ma Z, Yang Y, Zeng Z, Hauptmann AG (2017) Bi-level semantic representation analysis for multimedia event detection. IEEE Trans Cybern 47:1180–1197
10. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2015) Semantic image segmentation with deep convolutional nets and fully connected CRFs. International Conference on Learning Representations, In, pp 1–14
11. Cinbis RG, Verbeek J, Schmid C (2017) Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans Pattern Anal Mach Intell 39:189–203. doi:10.1109/TPAMI.2016.2535231
12. Dai J, Li Y, He K, Sun J (2016) R-FCN: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp 379–387
13. Deselaers T, Alexe B, Ferrari V (2012) Weakly supervised localization and learning with generic knowledge. Int J Comput Vis 100:275–293. doi:10.1007/s11263-012-0538-3
14. Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A (2014) The Pascal visual object classes challenge: a retrospective. Int J Comput Vis 111:98–136. doi:10.1007/s11263-014-0733-5
15. Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: the KITTI dataset. Int J Robot Res 32: 1231–1237. doi:10.1177/0278364913491297
16. Gidaris S, Komodakis N (2015) Object detection via a multi-region & semantic segmentation-aware CNN model. IEEE International Conference on Computer Vision
17. Girshick R (2015) Fast R-CNN. IEEE International Conference on Computer Vision
18. Han J, Zhang D, Cheng G, Guo L, Ren J (2015) Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans Geosci Remote Sens 53:3325–3337
19. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp 171–180
20. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia. pp 675–678
21. Jiang W, Zhao Z, Su F (2016) Bayes pooling of visual phrases for object retrieval. Multimed Tools Appl 75: 9095–9119. doi:10.1007/s11042-015-2939-0
22. Karthikeyan S, Ngo T, Eckstein M, Manjunath BS (2015) Eye tracking assisted extraction of attentionally important objects from videos. Proc IEEE Conf Comput Vis Pattern Recognit. doi:10.1109/CVPR.2015.7298944
23. Krizhevsky A, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proceeding NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, Curran Associates Inc., Lake Tahoe, Nevada — December 03–06, 2012, pp. 1097–1105
24. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S (2016) SSD : single shot MultiBox detector. European Conference on Computer Vision
25. Long J, Shelhamer E (2015) Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition
26. Ma Z, Chang X, Yang Y, Sebe N, Hauptmann AG (2017) The many shades of negativity. IEEE Trans Multimedia 19:1558–1568
27. Ma Z, Chang X, Xu Z, Sebe N, Hauptmann AG (2017) Joint attributes and event analysis for multimedia event detection. IEEE Trans Neural Netw Learn Syst. doi:10.1109/TNNLS.2017.2709308
28. Mnih V, Heess N, Graves A, Kavukcuoglu K (2014) Recurrent models of visual attention. Advances in Neural Information Processing Systems, In, pp 2204–2212
29. Oquab M, Bottou L, Laptev I, Sivic J (1717–1724) (2014) learning and transferring mid-level image representations using convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition. pp, In
30. Oquab M, Bottou L, Laptev I, Sivic J (2015) Is object localization for free? - weakly-supervised learning with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition, In, pp 685–694
31. Papadopoulos DP, Clarke ADF, Keller F, Ferrari V (2014) Training object class detectors from eye tracking data. In: European Conference on Computer Vision. pp 1–16

32. Redmon J, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition
33. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceeding NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems, MIT Press Cambridge, Montreal, Canada — December 07–12, 2015, pp. 91–99
34. Ren W, Member S, Huang K, Member S (2016) Weakly supervised large scale object localization with multiple instance learning and bag splitting. IEEE Trans Pattern Anal Mach Intell 38:405–416. doi:10.1109/TPAMI.2015.2456908
35. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115:211–252. doi:10.1007/s11263-015-0816-y
36. Sharma S, Kiros R, Salakhutdinov R (2016) Action recognition using visual attention. International Conference on Learning Representations, In, pp 1–11
37. Shi M, Ferrari V (2016) Weakly supervised object localization using size estimates. In: European Conference on Computer Vision
38. Shih KJ, Singh S, Hoiem D (2016) Where to look: focus regions for visual question answering. IEEE, Las Vegas
39. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. pp 1–14
40. Song HO, Girshick R, Jegelka S, Mairal J, Harchaoui Z, Darrell T (2014) On learning to localize objects with minimal supervision. In: Proceeding ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning vol. 32, Beijing, China, 21–26 June, 2014
41. Song HO, Lee YJ, Jegelka S, Darrell T (2014) Weakly-supervised discovery of visual pattern configurations. In: Proceeding NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems, MIT Press Cambridge, Montreal, Canada, 8–13 December, 2014
42. Treue S, Martinez Trujillo JC (1999) Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399:575–579. doi:10.1038/21176
43. Uijlings JRR, Sande KE a., Gevers T, Smeulders a. WM (2013) Selective search for object recognition. Int J Comput Vis 104:154–171
44. Uijlings JRR, Keller F, Ferrari V (2016) We don't need no bounding-boxes: training object class detectors using only human verification. IEEE Conference on Computer Vision and Pattern Recognition
45. Wang C, Huang K, Ren W, Zhang J, Maybank S (2015) Large-scale weakly supervised object localization via latent category learning. IEEE Trans Image Process 24:1371–1385. doi:10.1109/TIP.2015.2396361
46. Xu H, Saenko K (2016) Ask, attend and answer: exploring question-guided spatial attention for visual question answering. European Conference on Computer Vision, In, pp 451–466
47. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. International Conference on Machine learning
48. You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In, IEEE Conference on Computer Vision and Pattern Recognition, p 10
49. Zhang D, Han J, Li C, Wang J, Li X (2016) Detection of co-salient objects by looking deep and wide. Int J Comput Vis 120:215–232. doi:10.1007/s11263-016-0907-4
50. Zhang D, Han J, Han J, Shao L (2016) Cosaliency detection based on Intrasaliency prior transfer and deep Intersaliency mining. IEEE Trans Neural Netw Learn Syst 27:1163–1176. doi:10.1109/TNNLS.2015.2495161
51. Zhang D, Meng D, Zhao L, Han J (2016) Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In: Proceeding IJCAI'16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, New York, USA, 9–15 July, 2016, pp. 3538–3544
52. Zhang D, Meng D, Han J (2017) Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Trans Pattern Anal Mach Intell 39:865–878. doi:10.1109/TPAMI.2016.2567393
53. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. IEEE Conference on Computer Vision and Pattern Recognition
54. Zhu L, Shen J, Jin H, Xie L, Zheng R (2015) Landmark classification with hierarchical multi-modal exemplar feature. IEEE Trans Multimedia 17:981–993. doi:10.1109/TMM.2015.2431496
55. Zhu L, Shen J, Jin H, Zheng R, Xie L (2015) Content-based visual landmark search via multimodal hypergraph learning. IEEE Trans Cybern 45:2756–2769. doi:10.1109/TCYB.2014.2383389
56. Zhu Z, Liang D, Zhang S, Huang X, Baoli Li SH (2016) Traffic-sign detection and classification in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp 2110–2118

57. Zhu L, Shen J, Liu X, Xie L, Nie L (2016) Learning compact visual representation with canonical views for robust mobile landmark search. In: Proceeding IJCAI'16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, New York, USA, 9–15 July 2016, pp. 3959–3965
58. Zhu L, Shen J, Xie L, Cheng Z (2016) Unsupervised topic hypergraph hashing for efficient mobile image retrieval. IEEE Trans Cybern. doi:10.1109/TCYB.2016.2591068
59. Zhu L, Shen J, Xie L, Cheng Z (2017) Unsupervised visual hashing with semantic assistant for content-based image retrieval. IEEE Trans Knowl Data Eng 29:472–486. doi:10.1109/TKDE.2016.2562624

Wenhui Jiang is a PhD Candidate in the multimedia communication and pattern recognition lab, at Beijing University of Posts and Telecommunications. He was a visiting student at Department of Electrical and Computer Engineering, University of California, Santa Barbara from 2015 to 2016. His current research interests include large-scale image retrieval, object detection and deep learning.

Zhicheng Zhao is an associate professor of Beijing University of Posts and Telecommunications. He was a visiting scholar at School of Computer Science, Carnegie Mellon University from 2015 to 2016. His research interests are computer vision, video semantic understanding and retrieval. He has authored and coauthored more than 50 journal and conference papers.

Fei Su is a professor in the multimedia communication and pattern recognition lab, school of information and telecommunication, Beijing University of Posts and Telecommunications. She received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT in 2000. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009, Her current interests include pattern recognition, image and video processing and biometrics. She has authored and co-authored more than 70 journal and conference papers and some textbooks.