Noise-Aware Fully Webly Supervised Object Detection

Yunhang Shen[1], Rongrong Ji[1]*, Zhiwei Chen[1], Xiaopeng Hong[2],
Feng Zheng[3], Jianzhuang Liu[4], Mingliang Xu[5], Qi Tian[4]
[1]Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University, [2]Xi'an Jiaotong University
[3]Department of Computer Science and Engineering, Southern University of Science and Technology
[4]Noah's Ark Lab, Huawei Technologies, [5]Zhengzhou University
shenyunhang01@gmail.com, rrj@xmu.edu.cn, zhiweichen@stu.xmu.edu.cn
hongxiaopeng@mail.xjtu.edu.cn, zhengf@sustech.edu.cn, liu.jianzhuang@huawei.com
iexumingliang@zzu.edu.cn, tian.qi1@huawei.com

Abstract

We investigate the emerging task of learning object detectors with sole image-level labels on the web without requiring any other supervision like precise annotations or additional images from well-annotated benchmark datasets. Such a task, termed as fully webly supervised object detection, is extremely challenging, since image-level labels on the web are always noisy, leading to poor performance of the learned detectors. In this work, we propose an end-to-end framework to jointly learn webly supervised detectors and reduce the negative impact of noisy labels. Such noise is heterogeneous, which is further categorized into two types, namely background noise and foreground noise. Regarding the background noise, we propose a residual learning structure incorporated with weakly supervised detection, which decomposes background noise and models clean data. To explicitly learn the residual feature between clean data and noisy labels, we further propose a spatially-sensitive entropy criterion, which exploits the conditional distribution of detection results to estimate the confidence of background categories being noise. Regarding the foreground noise, a bagging-mixup learning is introduced, which suppresses foreground noisy signals from incorrectly labelled images, whilst maintaining the diversity of training data. We evaluate the proposed approach on popular benchmark datasets by training detectors on web images, which are retrieved by the corresponding category tags from photo-sharing sites. Extensive experiments show that our method achieves significant improvements over the state-of-the-art methods [1].

*Corresponding author.

[1]Code and dataset are available at: https://github.com/shenyunhang/NA-fWebSOD.

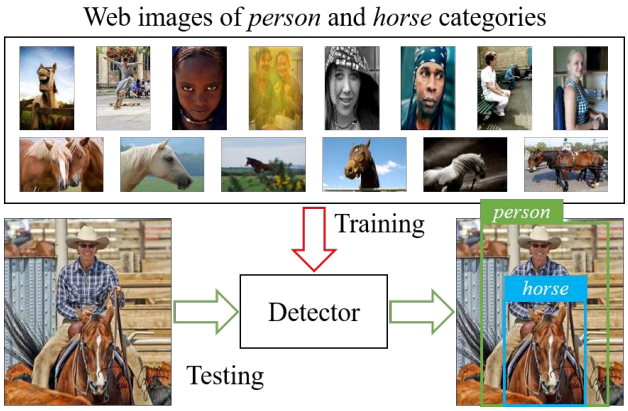

Figure 1: The overall flowchart of fully webly supervised object detection.

1. Introduction

Most object detection methods [18, 40, 33, 39, 14, 13, 15] rely on strong supervision, *i.e.*, ground-truth bounding boxes, from well-annotated datasets [12, 32] for training. Methods such as Mask R-CNN [22] even leverage fine-grained pixel-level masks for supervision. Clearly, collecting annotations of bounding boxes or pixel-level masks is labor-expensive, which results in a serious limitation of existing methods in terms of both the category diversity and label quantity. It is infeasible to learn an object detector to effectively handle numerous object categories using such a setting. One way to reduce the requirement of strong supervision is the Weakly Supervised Object Detection (**WSOD**), which only relies on manual image-level annotations for training [37, 6, 50, 56, 45, 44]. However, for applications needing very large-scale image sets and categories, image-level annotations still require enormous human effort. In contrast, along with the popularity of photo-sharing sites

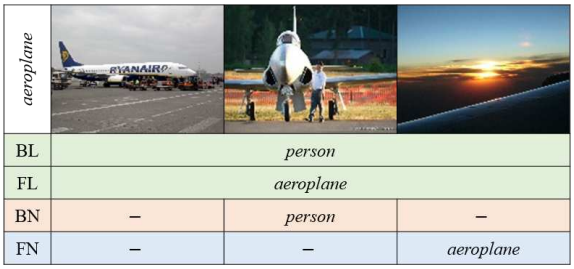

Figure 2: This figure depicts several web images retrieved by query *aeroplane*. The corresponding background labels (BL), foreground labels (FL), background noise (BN) and foreground noise (FN) are enumerated when target categories only have *aeroplane* and *person*.

like Flickr, there has been an explosion of images with noisy tags available on the web. It is thus desirable to learn object detectors from such large-scale web resources with noisy image-level labels, which is referred to as Webly Supervised Object Detection (**WebSOD**). In this paper, we focus on fully WebSOD (**fWebSOD**), *i.e*., the most extreme case of WebSOD where only web images are available and no well-annotated benchmark is involved during training, as shown in Fig. 1. Compared to WSOD and WebSOD, fWebSOD is more feasible to learn diverse and numerous object categories in real-world scenarios without any other form of knowledge, *e.g*., precious annotations or additional images from well-annotated benchmark datasets.

Although fWebSOD is challenging without any existing work in the literature, several attempts of WebSOD have been made [11, 7, 53]. Prevailing methods for the WebSOD task directly learn a detector from web labels with a simple-to-complex strategy [7, 52] or using additional data, *e.g*., Google books ngrams corpora [11] and PASCAL VOC images [53]. Such methods have one main drawback: They do not explicitly reduce the negative impact of image-level label noise in web data, which introduces potential risks of degenerating the performance of a learned detector significantly. Moreover, these methods usually follow a two/multi-stage scheme during training or testing. Little work explores an end-to-end pipeline for fWebSOD task.

In this paper, we address the above drawbacks by well handling image-level noisy labels in an end-to-end fashion. We categorize the heterogeneous noise into two types, namely background noise and foreground noise. We define several relative concepts here. (i) *Background labels* and *foreground labels* are referred to as the background and foreground parts of the image-level labels, respectively. (ii) *Background noise*, *i.e*., *missing labels*, denotes those background labels that fail to describe foreground categories existing in the image. For example, the instances of categories *person* and *areoplane* coexist in the second image of Fig. 2, but the category *person* is not labelled, which is defined as the background noise. (iii) *Foreground noise* denotes those foreground labels where no instance of the foreground category appears in the image. For example, the last image of Fig. 2 does not contain any aeroplane of the target category *areoplane*.

To handle the background noise, we decompose such noise by modelling clean data with residual learning. To this end, the reliable parts of background labels need to be identified from the massive noisy data explicitly. We observe that the distribution of accurate detection results for background categories is spatially scattered and numerically uniform. Motivated by this observation, for the background labels, we resort to producing spatially scattered proposals with uniform and moderate scores, whilst punishing the detection results where only a small minority of clustered proposals produce high scores. To handle the foreground noise, we collect multiple images of the same foreground label to synthesize a set of new training samples, as inspired by Multiple Instance Classification (MIC) [1] where any instances with positive labels will move positive labels to the corresponding bags. Such multi-instance-bagging mechanism enables to suppress the influence of foreground noise from incorrect labels, and to simultaneously maintain the diversity of training samples.

In particular, we propose an end-to-end learning framework to jointly learn fully webly supervised detectors and reduce the negative impact of image-level noisy labels. Given a set of target categories, we query photo sharing sites like Flickr to retrieve the corresponding web images automatically. To tackle the background noise, we design a residual learning structure incorporated with weakly supervised detection. A novel spatially-sensitive entropy criterion is further proposed to estimate the spatial and numerical entropy of detection results in the bounding-box search space. The criterion estimates the confidence of background labels being noise. To handle the foreground noise, a bagging-mixup learning strategy is introduced to collect multiple images of the same foreground label to synthesize a set of new training samples, each of which is a convex combination of all images in the bag. Extensive experiments show that the proposed framework achieves significant improvements over state-of-the-art methods [11, 7, 53] on PASCAL VOC and MS COCO. In summary, the main contributions of this paper are as follows:

- We propose a residual learning structure incorporated with weakly supervised detection in an end-to-end framework, which learns fully webly supervised detectors and reduces the negative impact of noisy labels by decomposing noise and modelling clean data.

- A spatially-sensitive entropy criterion and a bagging-mixup learning are further proposed to explicitly estimate the confidence of background labels being noisy and suppress the influence of foreground noise from

incorrect labels, respectively.

- Our models trained on only web data, which has about $4,000$ images each category, achieves significant improvements over the state-of-the-art methods on popular benchmark, *i.e*., PASCAL VOC and MS COCO.

2. Related Work

Weakly supervised object detection. WSOD refers to learning an object detection model with only image-level annotations that only indicates the presence of an object category. Recent approaches combine convolutional neural networks (CNNs) and Multiple Instance Classification (MIC) [1] into a unified framework [6, 28, 10, 51, 42, 43]. The learning stage of MIC alternates between selecting positive samples and training an appearance model. There are some methods focusing on proposal-free paradigms by taking advantage of deep feature maps [4, 3, 62, 59] and class activation maps [61, 20, 59]. Some works also use additional annotations and data to improve the performance, *e.g*., object size estimation [47], instance count annotation [16], video motion cue [49] and human verification [38]. Knowledge transfer for progressive cross-domain adaptation is also exploited, *e.g*., data domain adaption [46] and task domain adaption [25]. Instead of optimizing the MIC, some methods optimize the objective function of instance-level localization. For example, the works in [30, 27, 16, 50] mine the high-confidence proposals, which are then treated as positive samples to train a fully supervised model. Many efforts [60, 17] have been made to mine high-quality bounding boxes. To further improve the robustness, some works [50, 31, 54, 58] combine weakly supervised MIC models and fully supervised detectors.

Webly supervised learning. Webly supervised learning has been widely studied in the past decade, which is typically used in image classification [5, 34, 35, 36, 21, 63], object detection [11, 7, 53], and semantic segmentation [55, 41, 24]. The domain adaptation approach by Bergamo *et al*. [5] is proposed to combine manually annotated examples and web data to learn image classifiers. Mahajan *et al*. [34] showed that training large-scale hashtag prediction leads to improvements in image classification and object detection tasks. To cope with the label noise, Niu *et al*. [35] proposed to join variational autoencoder and classification network to leverage the image-level information. Guo *et al*. [21] leveraged curriculum learning by measuring the complexity of data using distribution density for image classification. Niu *et al*. [36] combined webly supervised learning and zero-shot learning to learn zero-shot fine-grained classifiers. Zhuang *et al*. [63] proposed to input multiple web images to CNNs and pool parts of the neuron activations as the final representation for classification. Wei *et al*. [55] utilized easy web data to assist semantic segmentation with image-level labels. Shen *et al*. [41] proposed to utilize complementary information of web and target data to generate training masks for semantic segmentation. Hong *et al*. [24] used classifiers to identify relevant spatio-temporal volumes in web video and generated object masks for segmentation.

Webly supervised object detection. There are a few attempts in the literature for WebSOD. Divvala *et al*. [11] trained deformable part models from the web data with Google n-grams corpora to expand the categories. Chen *et al*. [7] proposed a two-stage approach to learn a detector from web data, which initiates CNNs with simple Google images and fine-tunes them on more complex Flickr images. Tao *et al*. [53] focused on knowledge transfer from web data to target data with adversarial domain adaptation. Different from the work in [11, 7], we reduce the negative impact of noisy image-level labels in web data by handling both background noise and foreground noise in an end-to-end manner. In contrast to [53] where the target dataset is used, we aim at fWebSOD that trains detectors using only the web images without any image from human-annotated datasets, *e.g*., PASCAL VOC [12] or MS COCO [32]. Compared to WSOD and WebSOD, fWebSOD does not rely on any other form of knowledge, *e.g*., manual annotations or additional images, and is able to handle diverse and numerous object categories in real-world scenarios.

3. The Proposed Method

Given a set of N_c categories, we retrieve web images by using the category labels as the query keywords and construct the training data $D = \{I_i, \mathbf{t}_i\}_{i=1}^{N_D}$, where I_i is a crawled web image and $\mathbf{t}_i \in \Re^{N_c}$ is the corresponding one-hot label vector. We employ the basic WSDDN [6] as the base model in our framework. We first extract the features $\phi = \{\phi_i\}_{i=1}^{N_b}$ of N_b object proposals $\{b_i\}_{i=1}^{N_b}$ of an image from the backbone by a spatial pyramid pooling layer [23]. The pooled features are transformed by two fully-connected (FC) layers, which output the proposal features $\phi^{\text{fc}} = \{\phi_i^{\text{fc}}\}_{i=1}^{N_b}$. Then the proposal features are forked into two streams, *i.e*., a classification stream and a detection stream, producing two score matrices $X^c, X^d \in \Re^{N_b \times N_c}$ by two FC layers, respectively. Both the score matrices are normalized by the softmax function $\sigma(\cdot)$ over categories and proposals, respectively:

$$\sigma(X^c)_{ij} = \frac{e^{X^c_{ij}}}{\sum_{k=1}^{N_c} e^{X^c_{ik}}},\ \sigma(X^d)_{ij} = \frac{e^{X^d_{ij}}}{\sum_{r=1}^{N_b} e^{X^d_{rj}}}. \quad (1)$$

Then the Hadamard product of the two streams outputs the detection score matrix: $X^s = \sigma(X^c) \odot \sigma(X^d)$. To acquire image-level classification scores, a sum pooling is further applied: $\mathbf{y}_k = \sum_{r=1}^{N_b} X^s_{rk}$, where X^s_{rk} is the score of the r-th proposal and the k-th category in X^s. Then we obtain

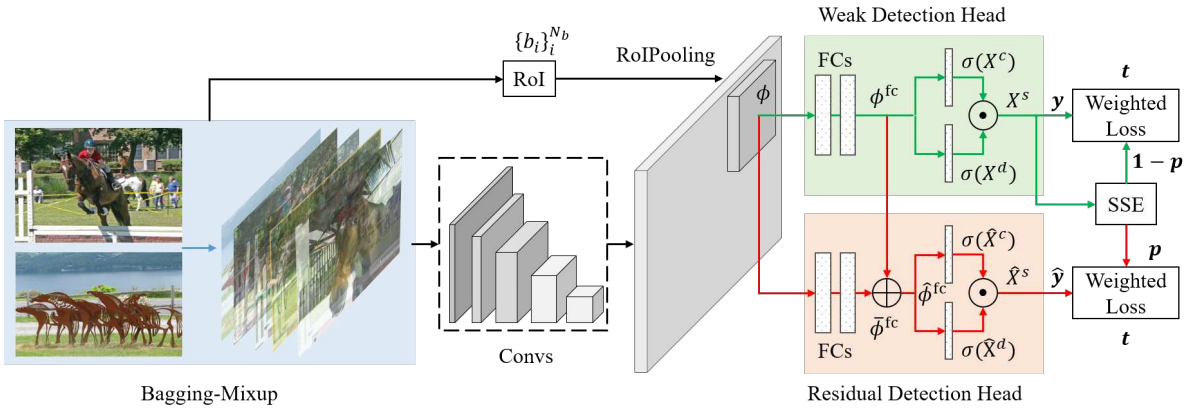

Figure 3: Overview of the proposed framework. Our method consists of three components: First, a bagging-mixup (BM) learning strategy constructs a set of new training images with the same foreground label to suppress the negative influence of foreground noise from incorrect labels. Second, the residual detection (RD) head and weak detection (WD) head are responsible for decomposing background noise and modelling clean data, respectively. Third, the proposed spatially-sensitive entropy (SSE) criterion is utilized to estimate the confidence of trusting image-level background labels.

a baseline cross-entropy loss function $\mathcal{L}^{\text{baseline}}$:

$$\mathcal{L}^{\text{baseline}} = \sum_{k=1}^{N_c} \Big\{ \mathbf{t}_k \log \mathbf{y}_k + (1-\mathbf{t}_k) \log(1-\mathbf{y}_k) \Big\}. \quad (2)$$

Our baseline approach is to learn a detector directly on the web data D. However, as shown in the subsequent experiments, the performance of such a detector drops dramatically compared to the detector trained on manual-annotated image-level labels. One main reason is that web data is noisy. To conquer this issue, we present an end-to-end learning framework to reduce the negative impact of image-level noisy labels in web data from two aspects, *i.e.*, background noise and foreground noise, as illustrated in Fig. 3.

3.1. Noise Decomposition

To reduce the negative impact of background noise, we propose a residual feature learning structure incorporated with weakly supervised detection to decompose background noise and model clean data. We leverage multi-task learning to learn two detection heads, *i.e.*, weak detection head and residual detection head, which share the backbone.

The weak detection (WD) head has the pooled features ϕ as its input and outputs proposal features ϕ^{fc} and detection scores X^s, which is similar to our baseline approach. The loss function of the WD head for category k is:

$$\mathcal{L}_k^{\text{WD}} = \mathbf{t}_k \log \mathbf{y}_k + (1-\mathbf{t}_k) \log(1-\mathbf{y}_k). \quad (3)$$

The proposed residual detection (RD) head is targeted for learning the residual features between reliable and unreliable parts in the massive noisy data. Specifically, the pooled features ϕ are mapped to residual features $\bar{\phi}^{\text{fc}} = \{\bar{\phi}_i^{\text{fc}}\}_{i=1}^{N_b}$ by two FC layers. We sum the residual features $\bar{\phi}^{\text{fc}}$ and the proposal features ϕ^{fc} from the WD head to get noise features $\hat{\phi}^{\text{fc}} = \{\hat{\phi}_i^{\text{fc}}\}_{i=1}^{N_b}$, where $\hat{\phi}_i^{\text{fc}} = \bar{\phi}_i^{\text{fc}} + \phi_i^{\text{fc}}$. Similar to the WD head, $\hat{\phi}^{\text{fc}}$ is fed into a classification stream and a detection stream, followed by the softmax operation and the sum pooling, which produces image-level classification scores $\hat{\mathbf{y}}_k = \sum_{r=1}^{N_b} \hat{X}_{rk}^s$. Given a category k, the loss function of the RD head is:

$$\mathcal{L}_k^{\text{RD}} = \mathbf{t}_k \log \hat{\mathbf{y}}_k + (1-\mathbf{t}_k) \log(1-\hat{\mathbf{y}}_k). \quad (4)$$

Finally, we obtain the overall loss function, which is the sum of all category-specific weighed sums of $\mathcal{L}_k^{\text{WD}}$ and $\mathcal{L}_k^{\text{RD}}$:

$$\mathcal{L} = \sum_{k=1}^{N_c} \big\{ (1-\mathbf{p}_k)\mathcal{L}_k^{\text{WD}} + \mathbf{p}_k \mathcal{L}_k^{\text{RD}} \big\}, \quad (5)$$

where $\mathbf{p}_k \in [0,1]$ is the estimated confidence of the k-th background label being a noisy label in an image.

From the perspective of learning the relation between clean data and noisy labels, the RD head works as a decomposition term, which helps the WD head to utilize the reliable information among the massive noisy data, whilst avoiding big influence by the unreliable information. When label of category k has low confidence of being noise, *i.e.*, $\mathbf{p}_k$ is low, the RD head is suppressed and the WD head models reliable information for category k. When $\mathbf{p}_k$ is high, the RD head leverages the proposal features ϕ^{fc} from the WD head and produces noise features $\hat{\phi}^{\text{fc}}$ to predict the unreliable label of category k, which imposes the RD head to decompose the noise by learning the residual features $\bar{\phi}^{\text{fc}}$. Thus, the residual learning structure jointly decomposes background noise and models clean data based on the confidence $\mathbf{p}$, which controls the gradient flow through the network. The confidence $\mathbf{p}$ can be seen as an information gate. We visualize the proposal scores and pixel gradient maps of the WD and RD heads in Fig. 4. WD and RD have high responses to ground-truth labels and foreground labels

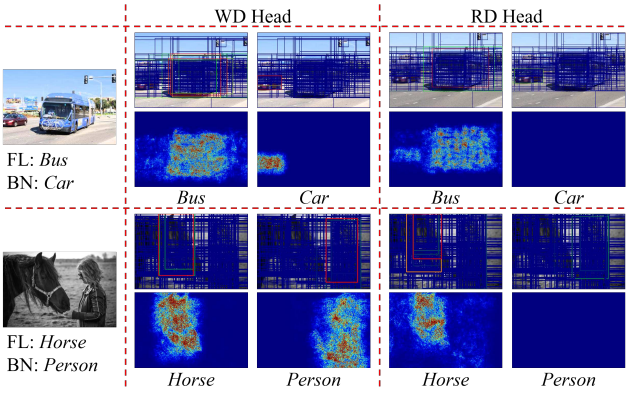

Figure 4: Proposal scores and the corresponding pixel gradient maps of the WD and RD heads for two web images.

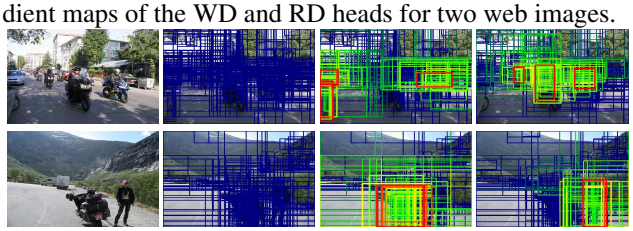

Figure 5: The distributions of detection results. The first column shows input images. The last three columns explain the ideal detection results of background categories and foreground categories, *i.e.*, categories *motorbike* and *person*, respectively. The figure is drawn with the jet color scale, where the red rectangles correspond to high scores and the blue ones to low scores.

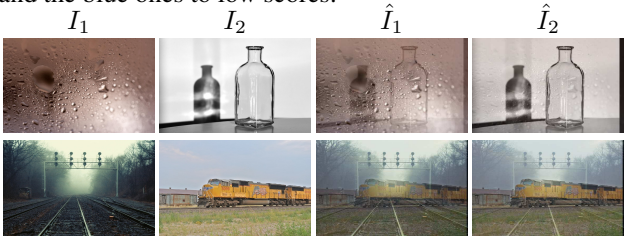

Figure 6: This figure depicts bagging-mixup learning for categories *bottle* (top) and *train* (bottom).

(FL), respectively. The two results combined from WD and RD can decompose BN from FL. Instead of predicting the confidence $\mathbf{p}$ by the model, we estimate it explicitly in online fashion, which is detailed in the next subsection.

3.2. Spatially-Sensitive Entropy Criterion

We observe that the distribution of accurate detection results for background categories is spatially scattered and numerically uniform, whilst the detection results that are spatially clustered and numerically nonuniform may contain instances of target categories, as illustrated in Fig. 5. Motivated by this observation, for the background labels, we resort to producing spatially scattered proposals with low scores, whilst punishing the detection results where a small minority of clustered proposals produce high scores.

We utilize the Shannon entropy as a sparsity indicator to describe the conditional distribution of detection results, which estimates the confidence $\mathbf{p}$ of background labels being noise. It is noted that the detection results consist of both confidence scores and bounding boxes. Suppose we have a result with only two bounding boxes $\{b_A, b_B\}$ and the corresponding detection scores $\{\frac{1}{2}, \frac{1}{2}\}$ for a category. If b_A and b_B have no overlap, we can estimate the entropy as $\ln 2$. However, if b_A and b_B have a large intersection-over-union (IoU), *e.g.*, $\text{IoU}(b_A, b_B) \geq 0.9$, one would expect the entropy to be lower. In the latter case, b_A and b_B are near points in the bounding-box searching space, and the detection result is sparser than the former case. Therefore, it is difficult to accurately estimate the sparsity of detection results without the spatial information of the bounding boxes. To handle this, we propose a Spatially-Sensitive Entropy (SSE) criterion to estimate the sparsity by introducing spatial information in this subsection.

We compute the Shannon entropy of detection scores as:

$$E_{rk} = -X^s_{rk} \ln X^s_{rk}, \tag{6}$$

where $E \in \Re^{N_b \times N_c}$ and $X^s_{rk} \in [0, 1]$. We also compute the Jaccard index matrix $J \in \Re^{N_b \times N_b}$ as $J_{ij} = \text{IoU}(b_i, b_j)$, where b_i and b_j denote the i-th and the j-th proposals, respectively. We obtain an entropy regularizer as:

$$G = E \oslash (JE), \tag{7}$$

where $\oslash$ is the Hadamard division and $G \in \Re^{N_b \times N_c}$. The denominator term JE sums up all the entropies of individual detection scores weighted by their spatial information, *i.e.*, the IoU between two proposals. Then, the original entropy E is divided by the weighted sum of entropies, which is in the range of $[0, 1]$. Our intuition is that the entropy decreases according to the IoU between each pair of proposals. If the detection bounding boxes have no overlap with each other, then $JE = E$, and G is an all-one matrix.

Then, the refined entropy after considering the spatial information among proposals is computed as:

$$\hat{E} = G \odot E, \tag{8}$$

where $\odot$ is the Hadamard product. The confidence of the background label k being noise in Eq. 5 is computed as:

$$\mathbf{p}_k = \begin{cases} 1 - \frac{\sum_r^{N_b} \hat{E}_{rk}}{\mathbf{z}_k} & \text{if } \mathbf{t}_k = 0 \\ 0 & \text{if } \mathbf{t}_k = 1 \end{cases}, \tag{9}$$

where $\mathbf{p}, \mathbf{z} \in \Re^{N_c}$ and $\mathbf{z}_k = -\mathbf{y}_k \ln \frac{\mathbf{y}_k}{N_b}$. We use $\mathbf{z}_k$ to denote the maximum entropy of detection results given image-level prediction $\mathbf{y}_k$ for the k-th category and N_b bounding boxes. Therefore, $\frac{\sum_r^{N_b} \hat{E}_{rk}}{\mathbf{z}_k}$ is in the range of $[0, 1]$ in Eq. 9.

To further verify the above analysis, we compute the SSE criterion $\mathbf{p}$ of 200 web images randomly sampled from

Table 1: The datasets in the experiments.

Category	Dataset	#Images	
		Training	Testing
VOC	PASCAL VOC 2007 [12]	-	4,952
	PASCAL VOC 2012 [12]	-	1,0991
	Flickr-Clean [55]	41,625	-
	Flickr-VOC	88,064	-
COCO	MS COCO [32]	-	5,000
	Flickr-COCO	335,324	-

Flickr-VOC after model training and normalize them in the range of $[0, 1]$. The average $\hat{E}$ of foreground and background are 0.07 and 0.78, respectively. For BN, *i.e.*, foreground missing, $\mathbf{p}$ has an average of 0.93. The Pearson correlation between SSE and BN is as high as 0.91.

3.3. Bagging-Mixup Learning

To reduce the negative impact of foreground noise, we propose a novel bagging-mixup strategy for data augmentation, which is inspired by the multi-instance-bagging mechanism to efficiently handle incorrect labels. In particular, the bagging-mixup strategy applies convex combinations of all images with the same foreground label in the bag to synthesize a set of training images. Therefore, bagging-mixup aims at suppressing the probability of using incorrect labels, whilst maintaining the diversity of training samples.

Bagging-mixup learning consists of three steps. First, we randomly sample N_a web images $\{I_i\}_{i=1}^{N_a}$ with the same label $\mathbf{t}$, *i.e.*, the same foreground label. Second, we random draw blending ratios $\{\lambda_i\}_{i=1}^{N_a}$ from a Dirichlet distribution $\text{Dir}(\alpha_1, \ldots, \alpha_{N_a})$, where $\alpha_1 = N_a\alpha_2 = \cdots = N_a\alpha_{N_a}$. Finally, bagging-mixup constructs multiple synthetic training images with the same label:

$$\hat{I}_i = \quad \lambda_1 I_i + \sum_{m,n}^{\{2,\ldots,N_a\},\{1,\ldots,N_a\}\setminus i} \lambda_m I_n \quad , \qquad (10)$$

where $i \in \{1, 2, \ldots, N_a\}$. The visual comparisons of the original and synthetic images are illustrated in Fig. 6. Then the synthetic images $\{\hat{I}_i\}_{i=1}^{N_a}$ are fed to the model as training samples with label $\mathbf{t}$. We do not extract object proposals from the synthetic images, which is infeasible in terms of efficiency during training. Instead, we translate proposal coordinates of the original images to the synthetic images.

The proposed bagging-mixup learning is distinct from the mixup [57] in the following two aspects. First, mixup is agnostic to the category, as it randomly samples data among all categories. Bagging-mixup is category-specific by sampling images with the same label, which is designed to be robust to foreground noise. Second, mixup only exploits partial information of each image pair to generate a single image. Bagging-mixup constructs multiple synthetic images, each of which is a convex combination of all images in the bag with the weights sampled from a Dirichlet distribution, which also maintains the diversity of training data.

4. Experimental Evaluation

4.1. Training Datasets

Flickr-VOC and Flickr-COCO. We construct two new datasets called Flickr-VOC and Flickr-COCO to train the detectors. The categories from PASCAL VOC [12] and MS COCO [32] are employed as queries to retrieve images from the Flickr photo-sharing website. No other query criteria, *e.g.*, date of capture, photographer's name, *etc*, are specified. For each category, we crawl images in about the first $4,000$ search results returned by the Flickr API. In total, $83,905$ and $335,327$ images are collected without any post-processing for Flickr-VOC and Flickr-COCO, respectively.

Flickr-Clean [55]. Flickr-Clean [55] is constructed from Flickr with PASCAL VOC [12] categories, which has $41,625$ images in total. Different from our Flickr-VOC, Flickr-Clean [55] is post-processed by a salient object detector (DRFI [26]) and saliency-cut segmentation [8] to remove noisy data and to keep only simple images. In other words, Flickr-Clean is filtered from the original web data and contains human annotations from [26, 8]. Therefore, our crawled Flickr-VOC is more challenging and closer to the real-world web dataset.

4.2. Testing Datasets

PASCAL VOC 2007 and 2012 [12]. When training on Flickr-VOC and Flickr-Clean [55], we evaluate the detectors on the test sets of PASCAL VOC 2007 and 2012 [12], which have $4,092$ and $10,991$ test images over 20 categories, respectively. In our evaluation, we ensure that none of the PASCAL VOC images (including the trainval and test sets) exists in our training set.

MS COCO [32]. When training on Flickr-COCO, we evaluate the detectors on MS COCO [32], which is among the most challenging datasets for object detection. It consists of 80 object categories. Our experiments involve $5,000$ images of MS COCO validation (minival) for testing. More detailed statistics about these datasets are given in Tab. 1.

4.3. Evaluation Protocol

For VOC categories, Average Precision (AP) and mean Average Precision (mAP) are used as the evaluation metrics. We follow the standard PASCAL VOC protocol to report the mAP at 50% Intersection-over-Union (IoU) of the detected boxes with the ground-truth. For COCO categories, we also report the standard COCO metrics, including AP at different IoU thresholds and scales.

4.4. Implementation Details

The proposed approach is implemented on 4 GPUs. We report our performance on three backbone networks, *i.e.*, VGG-CNN-F [29] (VGG-F), VGG-CNN-M-1024 (VGG-

Table 2: Comparison to the baselines for object detection on the VOC 2007 test set in terms of AP (%).

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
Training on PASCAL VOC 2007 trainval images with image-level annotations																					
WSDDN VGG-F [6]	42.9	**56.0**	32.0	17.6	10.2	61.8	50.2	29.0	3.8	**36.2**	18.5	31.1	**45.8**	54.5	10.2	**15.4**	36.3	45.2	50.1	43.8	34.5
WSDDN VGG-M [6]	43.6	50.4	32.2	**26.0**	9.8	58.5	**50.4**	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	37.8	**46.9**	53.4	**47.9**	34.9
WSDDN VGG16 [6]	39.4	50.1	31.5	16.3	12.6	**64.5**	42.8	42.6	10.1	35.7	**24.9**	38.2	34.4	**55.6**	9.4	14.7	30.2	40.7	**54.7**	46.9	34.8
Training on Flickr-VOC																					
WSDDN VGG-F	32.4	36.7	31.1	10.7	12.8	48.0	40.2	39.7	10.5	21.4	10.4	24.7	30.4	44.9	12.1	10.2	35.3	30.2	35.3	1.8	25.9
WSDDN VGG-M	6.6	24.3	32.3	10.8	13.8	37.3	37.5	41.5	7.6	24.4	5.5	29.6	30.0	47.9	10.4	9.7	35.1	13.9	41.4	20.7	25.5
WSDDN VGG16	35.8	39.5	35.8	9.6	10.0	51.5	39.5	41.3	7.1	22.4	7.4	31.0	33.4	47.3	13.0	9.2	32.7	27.5	44.6	14.2	27.6
Our VGG-F	45.4	38.1	38.9	20.1	13.8	60.8	42.9	55.2	**16.1**	29.2	9.4	33.3	30.9	52.9	14.5	14.9	37.8	28.8	49.2	26.8	32.9
Our VGG-M	45.7	38.5	36.9	20.6	**16.9**	55.2	38.8	57.5	14.8	25.0	10.6	38.7	39.3	51.8	16.3	13.6	38.0	34.6	46.3	26.1	33.3
Our VGG16	**45.9**	39.6	**39.8**	21.1	14.4	60.9	39.9	**61.5**	15.6	32.5	14.1	**44.8**	45.2	51.7	**18.0**	13.8	**38.9**	32.1	47.2	23.5	**35.1**

Table 3: Comparison to the SOTAs for object detection on the VOC 2007 test set in terms of AP (%).

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
Training on PASCAL VOC 2007 trainval images with proposal/image-level annotations																					
FSOD VGG16 [40]	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6	69.9
WSOD VGG16 [43]	64.8	70.7	51.5	25.1	29.0	74.1	69.7	69.6	12.7	69.5	43.9	54.9	39.3	71.3	32.6	29.8	57.0	61.0	66.6	57.4	52.5
Training on web data from Google and Flickr																					
Divvala *et al.* [11]	14.0	36.2	12.5	10.3	9.2	35.0	35.9	8.4	10.0	17.5	6.5	12.9	30.6	27.5	6.0	1.5	18.8	10.3	23.5	16.4	17.1
Chen *et al.* Google [7]	29.5	38.3	15.1	14.0	9.1	44.3	29.3	24.9	6.9	15.8	9.7	22.6	23.5	34.3	9.7	12.7	21.4	15.8	33.4	19.4	21.5
Chen *et al.* Flickr [7]	30.2	**41.3**	21.7	18.3	9.2	44.3	32.2	25.5	9.8	21.5	10.4	26.7	27.3	42.8	12.6	13.3	20.4	20.9	36.2	22.8	24.4
Training on Flickr-Clean and PASCAL VOC 2007 trainval images																					
Tao *et al.* VGG-M [53]	35.6	31.3	18.2	7.7	9.1	40.4	38.4	23.8	9.7	20.1	**33.4**	22.5	30.9	41.4	9.8	10.8	18.7	28.7	27.1	34.7	24.6
Tao *et al.* VGG16 [53]	40.6	30.1	17.8	15.9	6.4	42.9	40.5	31.5	11.4	20.3	27.4	15.7	24.1	43.8	8.9	12.2	17.7	37.3	32.1	31.0	25.4
Training on Flickr-Clean																					
Our VGG-F	43.7	34.5	32.9	12.6	13.7	54.2	45.2	35.0	11.3	26.0	26.9	22.7	25.7	49.2	20.8	9.1	34.7	48.9	46.6	**38.9**	31.6
Our VGG-M	44.3	37.8	32.5	15.0	14.1	55.2	44.5	32.4	10.9	28.0	26.8	17.9	26.2	49.6	20.2	9.7	35.4	49.4	48.9	37.2	31.8
Our VGG16	44.6	36.6	34.3	18.6	13.8	56.7	**47.2**	37.7	11.6	23.3	32.5	29.1	33.3	52.6	**21.5**	8.9	35.5	**52.4**	45.3	38.2	33.7
Training on Flickr-VOC																					
Our VGG-F	45.4	38.1	38.9	20.1	13.8	60.8	42.9	55.2	**16.1**	29.2	9.4	33.3	30.9	**52.9**	14.5	**14.9**	37.8	28.8	**49.2**	26.8	32.9
Our VGG-M	45.7	38.5	36.9	20.6	**16.9**	55.2	38.8	57.5	14.8	25.0	10.6	38.7	39.3	51.8	16.3	13.6	38.0	34.6	46.3	26.1	33.3
Our VGG16	**45.9**	39.6	**39.8**	**21.1**	14.4	**60.9**	39.9	**61.5**	15.6	**32.5**	14.1	**44.8**	**45.2**	51.7	18.0	13.8	**38.9**	32.1	47.2	23.5	**35.1**

M) and deep VGG-VD16 [48] (VGG16), which are initialized with the weights pre-trained on ImageNet [9].

Training. In all experiments, the size of mini-batch, the learning rate, the momentum, the decay weight and the dropout rate are set to 1, 0.001, 0.9, 0.0005 and 0.5, respectively. We freeze all convolutional layers in our backbones during training. To improve the robustness, we randomly adjust the exposure and saturation of the images by up to a factor of 1.5 in the HSV space. And a random crop with 0.9 of the original image size is applied. We use MCG [2] to generate object proposals for all experiments, including our baseline methods. We set the maximum number of region proposals in an image to $2,048$. All models are trained for 200K iterations. We apply Xavier initialization [19] to initialize the new fully-connected layers. The bagging-mixup hyper-parameter α_1 is set to 1.5.

Testing. We use the output of the WD head X^s as the final detection scores. Detection results are post-processed by a NMS module using a threshold of 0.5 IoU.

4.5. Comparison to Baselines

We first compare the performance of WSDDN trained on PASCAL VOC 2007 with human-annotated image-level labels and directly trained on the Flickr data, *i.e.*, Flickr-VOC. As shown in the first and the second parts of Tab. 2, the performance of the detector trained on Flickr-VOC decreases dramatically. Due to the noisy image-level labels in Flickr-VOC, the performances of the three backbones are only 25.9%, 25.5% and 27.6% in terms of mAP, with losses of 10.1%, 9.7% and 9.3% compared to the WSDDN models trained on PASCAL VOC 2007, respectively. Overall, there is a significant gap between the models trained on the PASCAL VOC and web data. We show that the proposed method on Flickr-VOC achieves the performances of 32.9%, 33.3% and 35.1% with improvements of 7.0%, 7.8% and 7.5%, respectively. It also demonstrates that our method outperforms the baselines by a large margin, and reduces the gap between fWebSOD and WSOD.

4.6. Comparison with State of the Arts (SOTAs)

We compare our method with the state of the arts, including [11, 7, 53]. Tab. 3 shows our results on the PASCAL VOC 2007 test in terms of mAP. Our three models on Flickr-VOC reach 32.9%, 33.3% and 35.1% mAP with VGG-F, VGG-M and VGG16 backbones, respectively, which outperform the state-of-the-art algorithms. Although Flickr-Clean has been post-processed to reduce noise, our models trained on Flickr-VOC still achieve better performance. As Flickr-VOC has more than twice the image number of Flickr-Clean, and our method is able to reduce the negative impact of noise. It is worth noting that, compared to the baseline WSDDN [6] approach, our method has the same inference speed. Our single model VGG-F outperforms the state-of-the-art result 25.4% with a gain of 7.5% in terms of mAP. Note that the comparison of data usage between the proposed framework and the previous methods can better reveal the significance of our work. The works in Divvala *et al.* [11], Chen *et al.* [7] and Tao *et al.* [53] all use external manual knowledge, *e.g.*, Google Books n-grams corpora, easy images from the Google search engine and PASCAL VOC images. However, our method only uses web images without any other form of knowledge.

Table 4: Comparison to the SOTAs for object detection on the VOC 2012 test set in terms of AP (%).

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
Training on PASCAL VOC 2012 trainval images with proposal/image-level annotations																					
FSOD VGG16 [18]	82.3	76.4	71.0	48.4	45.2	72.1	72.3	87.3	42.2	73.7	50.0	86.8	78.7	78.4	77.4	34.5	70.1	57.1	77.1	58.9	67.0
WSOD VGG16 [43]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.1
Training on Flickr-Clean and PASCAL VOC 2012 trainval images																					
Tao *et al*. VGG-M [53]	**44.3**	29.8	15.6	6.6	6.0	34.4	24.2	25.1	5.7	20.3	**22.3**	24.9	29.1	45.2	7.8	9.4	12.4	21.4	22.6	26.0	21.7
Training on Flickr-Clean																					
Our VGG-F	41.0	15.1	29.7	10.4	13.2	47.6	42.0	36.8	10.2	16.5	13.2	28.2	20.5	39.6	15.0	8.4	28.0	38.6	9.6	38.2	25.1
Our VGG-M	40.7	17.2	28.1	11.6	12.5	48.4	39.7	31.1	5.5	20.8	12.2	27.3	27.9	37.3	**17.7**	7.0	28.1	**40.1**	10.0	36.6	25.0
Our VGG16	39.6	18.1	30.7	8.4	12.0	51.2	**42.5**	42.6	9.4	20.4	16.1	23.5	26.9	41.6	17.1	7.9	29.1	39.0	9.6	**39.9**	26.3
Training on Flickr-VOC																					
Our VGG-F	40.2	33.7	**38.4**	12.5	13.0	52.7	40.4	41.2	13.0	24.7	18.0	31.6	**32.5**	51.7	12.4	11.6	33.3	30.4	40.5	22.1	29.7
Our VGG-M	42.6	36.7	36.9	17.5	**14.8**	53.6	38.3	44.4	**13.7**	28.8	19.2	24.6	26.8	**52.2**	11.3	10.5	**39.1**	26.6	42.9	21.4	30.1
Our VGG16	41.9	**40.6**	**38.4**	**20.5**	9.0	**56.9**	40.3	**50.1**	13.0	**30.8**	17.2	**32.7**	29.9	51.2	17.0	**13.5**	36.4	39.2	**45.3**	29.2	**32.7**

Table 5: Ablation study of our method for object detection on the VOC 2007 test set in terms of AP (%).

Method	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
Baseline	35.8	39.5	35.8	9.6	10.0	51.5	39.5	41.3	7.1	22.4	7.4	31.0	33.4	47.3	13.0	9.2	32.7	27.5	44.6	14.2	27.6
RD	42.2	34.8	34.5	19.4	9.8	53.1	42.0	46.6	10.6	26.0	9.4	29.9	26.2	48.7	13.7	**18.2**	38.8	28.7	40.7	22.7	29.8
RD + EW	44.3	34.9	38.7	16.3	13.0	55.5	40.7	44.3	15.3	23.5	5.2	35.6	35.3	50.7	14.7	11.6	30.1	**33.1**	45.9	24.9	30.7
RD + SSE	45.4	**46.9**	38.3	19.8	12.4	**61.7**	41.5	47.1	13.9	26.1	11.8	39.1	41.6	**52.8**	16.3	13.7	38.4	32.0	45.7	24.6	33.4
RD + SSE-ALL	**48.0**	42.7	**41.6**	20.1	12.5	60.8	**42.1**	48.7	15.4	27.7	**18.3**	38.8	34.9	52.2	16.8	11.6	**40.0**	32.8	39.0	**27.6**	33.6
RD + SSE + BM2	45.9	39.6	39.8	**21.1**	**14.4**	60.9	39.9	61.5	15.6	**32.5**	14.1	**44.8**	45.2	51.7	18.0	13.8	38.9	32.1	47.2	23.5	35.1
RD + SSE + BM3	45.7	39.9	40.9	20.7	14.3	60.7	39.9	**61.7**	**15.8**	32.1	13.9	44.4	**45.4**	51.8	**18.4**	15.2	38.7	32.0	47.7	24.3	**35.2**
RD + SSE + BM4	45.9	39.0	41.4	20.6	14.3	60.5	39.5	60.9	15.3	32.1	17.2	43.7	44.5	52.3	18.0	14.6	38.7	30.9	**49.3**	24.1	35.1

Table 6: Result on the COCO minival set.

Method	Avg. Precision, IoU:			Avg. Precision, Area:		
	0.5:0.95	0.5	0.75	S	M	L
Training on COCO train images with proposal/image-level annotations						
FSOD VGG16 [18]	21.2	41.5	-	-	-	-
WSOD VGG16 [43]	10.5	20.3	9.2	2.2	10.9	18.3
WSDDN VGG16	9.5	19.2	8.2	2.1	10.4	17.2
Training on Flickr-COCO						
WSDDN VGG16	3.1	7.0	2.3	0.4	2.6	6.9
Our VGG16	**5.4**	**10.6**	**4.6**	**0.6**	**5.1**	**10.7**

In Tab. 4, we also evaluate our method on the PASCAL VOC 2012 test. Our models consistently outperform the state-of-the-art methods. In Tab. 6, we evaluate our method on MS COCO. Compared to using well labelled MS COCO, directly training the WSDDN model on Flickr-COCO results in poor performance (19.2% vs. 7.0% $AP_{0.5}$). However, our framework achieves 10.6% $AP_{0.5}$, which outperforms the state-of-the-art method by a large margin.

4.7. Ablation Study

The residual detection (RD) head. To investigate the effect of the RD head, we set $\mathbf{p}_k$ to 0.5 for all categories. Thus, it always combines the gradient flow from both heads in this setting. As shown in Tab. 5, the result of RD is slightly better than the baseline, as it imposes the model to learn residual features without considering noise explicitly. It also demonstrates that the performance gain does not merely come from the additional parameters of RD head.

The spatially-sensitive entropy (SSE) criterion. To further verify the effects of the SSE criterion, we first use the original entropy to compute the confidence weight for each background label. This is implemented by replacing $\hat{E}$ with E in Eq. 9. As shown in Tab. 5, adding the original entropy weight to the RD head ("RD + EW") achieves 0.9% mAP improvement over the baseline method on Flickr-VOC. Replacing the original entropy by the SSE criterion, "RD + SSE", achieves a large gain of 3.6% mAP on PASCAL VOC 2007 test, which demonstrates the contribution of SSE to the overall network. We also apply the SSE criterion for all categories (background and foreground). We find that "RD + SSE-ALL" helps reduce foreground noise.

The bagging-mixup (BM) learning. We further examine the results of BM with different numbers N_a of images in a bag. In Tab. 5, we can see that the result of two images in a bag ("RD + SSE + BM2") is better than that without BM ("RD + SSE"), which suggests that the combinations of multiple images during training do help suppress the negative impact of foreground noise. We also evaluate the effectiveness of increasing N_a in BM. We find that the gain of more than two images in a bag ("RD + SSE + BM3" and "RD + SSE + BM4") is trivial.

5. Conclusion

In this work, we focus on training object detectors using only web supervision without requiring any other form of knowledge, *e.g.*, manual annotations or additional images. As image-level labels on the web contain heterogeneous noise, we categorize them into two types, namely background noise and foreground noise. To this end, we present an end-to-end learning framework to learn webly supervised detectors and reduce the negative impact of noisy labels. The proposed framework outperforms the baseline methods and sets new state-of-the-art results on the PASCAL VOC and MS COCO in the task of fWebSOD.

6. Acknowledgment

This work is supported by the Nature Science Foundation of China (No.U1705262, No.61772443, No.61572410, No.61802324 and No.61702136), National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

[1] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *AI*, 2013.

[2] P Arbeláez, J Pont-Tuset, J Barron, F Marques, and J Malik. Multiscale Combinatorial Grouping. In *CVPR*, 2014.

[3] Loris Bazzani, Alessandro Bergamo, Dragomir Anguelov, and Lorenzo Torresani. Self-Taught Object Localization with Deep Networks. In *WACV*, 2016.

[4] Archith J. Bency, Heesung Kwon, Hyungtae Lee, S. Karthikeyan, and B. S. Manjunath. Weakly Supervised Localization using Deep Feature Maps. In *ECCV*, 2016.

[5] Alessandro Bergamo. Exploiting weakly-labeled Web images to improve object classification : a domain adaptation approach. *NeurIPS*, 2010.

[6] Hakan Bilen and Andrea Vedaldi. Weakly Supervised Deep Detection Networks. In *CVPR*, 2016.

[7] Xinlei Chen. Webly Supervised Learning of Convolutional Networks. In *ICCV*, 2015.

[8] Ming-ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip H S Torr, and Shi-min Hu. Global Contrast based Salient Region Detection. *TPAMI*, 2015.

[9] J Deng, W Dong, R Socher, L.-J. Li, K Li, and L Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[10] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly Supervised Cascaded Convolutional Networks. In *CVPR*, 2017.

[11] Santosh K. Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *CVPR*, 2014.

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 2010.

[13] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018.

[14] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020.

[15] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019.

[16] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I. Morariu, and Larry S. Davis. C-WSL: Count-guided Weakly Supervised Localization. In *ECCV*, 2018.

[17] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning. In *CVPR*, 2018.

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

[19] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[20] Amogh Gudi, Nicolai van Rosmalen, Marco Loog, and Jan van Gemert. Object-Extent Pooling for Weakly Supervised Single-Shot Localization. In *BMVC*, 2017.

[21] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images. In *ECCV*, 2018.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *ECCV*, 2014.

[24] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly Supervised Semantic Segmentation using Web-Crawled Videos. In *CVPR*, 2017.

[25] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation. In *CVPR*, 2018.

[26] Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Yihong Gong, Nanning Zheng, and Jingdong Wang. Salient Object Detection: A Discriminative Regional Feature Integration Approach. *IJCV*, 2016.

[27] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep Self-Taught Learning for Weakly Supervised Object Localization. In *CVPR*, 2017.

[28] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. In *ECCV*, 2016.

[29] Chatfield Ken, Simonyan Karen, Vedaldi Andrea, and Zisserman Andrew. Return of the Devil in the Details Delving Deep into Convolutional Nets. In *BMVC*, 2014.

[30] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly Supervised Object Localization with Progressive Domain Adaptation. In *CVPR*, 2016.

[31] Yao Li, Linqiao Liu, Chunhua Shen, and Anton van den Hengel. Image Co-localization by Mimicking a Good Detector's Confidence Score Distribution. In *ECCV*, 2016.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016.

[34] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the Limits of Weakly Supervised Pretraining. In *ECCV*, 2018.

[35] Li Niu, Qingtao Tang, Ashok Veeraraghavan, and Ashu Sabharwal. Learning from Noisy Web Data with Category-level Supervision. In *CVPR*, 2018.

[36] Li Niu, Ashok Veeraraghavan, and Ashu Sabharwal. Webly Supervised Learning Meets Zero-shot Learning: A Hybrid Approach for Fine-grained Classification. In *CVPR*, 2018.

[37] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - Weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[38] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In *CVPR*, 2016.

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015.

[41] Tong Shen, Guosheng Lin, Chunhua Shen, and Ian Reid. Bootstrapping the Performance of Webly Supervised Semantic Segmentation. In *CVPR*, 2018.

[42] Yunhang Shen, Rongrong Ji, Changhu Wang, Xi Li, and Xuelong Li. Weakly Supervised Object Detection via Object-Specific Pixel Gradient. *TNNLS*, 2018.

[43] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic Guidance for Weakly Supervised Joint Detection and Segmentation. In *CVPR*, 2019.

[44] Yunhang Shen, Rongrong Ji, Kuiyuan Yang, Cheng Deng, and Changhu Wang. Category-Aware Spatial Constraint for Weakly Supervised Detection. *TIP*, 2019.

[45] Yunhan Shen, Rongrong Ji, Shengchuan Zhang, Wangmeng Zuo, and Yan Wang. Generative Adversarial Learning Towards Fast Weakly Supervised Detection. In *CVPR*, 2018.

[46] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly Supervised Object Localization Using Things and Stuff Transfer. In *ICCV*, 2017.

[47] Miaojing Shi and Vittorio Ferrari. Weakly Supervised Object Localization Using Size Estimates. In *ECCV*, 2016.

[48] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

[49] Krishna Kumar Singh, Fanyi Xiao, and Yong Jae Lee. Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection. In *CVPR*, 2016.

[50] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple Instance Detection Network with Online Instance Classifier Refinement. In *CVPR*, 2017.

[51] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly Supervised Region Proposal Network and Object Detection. In *ECCV*, 2018.

[52] Qingyi Tao, Hao Yang, and Jianfei Cai. Exploiting Web Images for Weakly Supervised Object Detection. *TMM*, 2018.

[53] Qingyi Tao, Hao Yang, and Jianfei Cai. Zero-Annotation Object Detection with Web Knowledge Transfer. In *ECCV*, 2018.

[54] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-Entropy Latent Model for Weakly Supervised Object Detection. In *CVPR*, 2018.

[55] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation. *TPAMI*, 2017.

[56] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. TS2C: Tight Box Mining with Surrounding Segmentation Context for Weakly Supervised Object Detection. In *ECCV*, 2018.

[57] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *ICLR*, 2018.

[58] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag Learning for Weakly Supervised Object Detection. In *CVPR*, 2018.

[59] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial Complementary Learning for Weakly Supervised Object Localization. In *CVPR*, 2018.

[60] Yongqiang Zhang, Yongqiang Li, and Bernard Ghanem. W2F : A Weakly-Supervised to Fully-Supervised Framework for Object Detection. In *CVPR*, 2018.

[61] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016.

[62] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft Proposal Networks for Weakly Supervised Object Localization. In *ICCV*, 2017.

[63] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In *CVPR*, 2017.