You reap what you sow: Using Videos to Generate High Precision Object Proposals for Weakly-supervised Object Detection

Krishna Kumar Singh and Yong Jae Lee
University of California, Davis

1. Introduction

Object detection has seen tremendous progress in recent years [4]. We now have detectors that can accurately detect objects in the presence of severe clutter, scale changes, viewpoint/pose changes, occlusion, etc. However, existing state-of-the-art algorithms require expensive and error-prone bounding box annotations for training, which severely limits the number of categories that they can be trained to recognize.

To tackle this issue, researchers have proposed to use only *weak-supervision* in which image-level object presence labels (like ‘dog’ or ‘no dog’) are provided rather than bounding box annotations [6, 1, 7]. In this setting, object detection is often formulated as multiple instance learning and solved using non-convex optimization in which the predicted object localizations on the training set and model learning are iteratively updated. Most existing methods start off by using an off-the-shelf object proposal method like selective search [9] or edge boxes [10] to generate thousands of candidate object proposals (Fig. 1 (b)). They then perform the extremely difficult data mining task of localizing the few relevant object regions among the thousands of noisy proposals in each image (i.e., akin to *finding a needle in a haystack*). Since there is no supervisory signal other than the image class label, this process often results in inaccurate initial object bounding box guesses which either correspond to only an object part or include background. Ultimately, these errors lead to inaccurate object detectors.

Rather than creating yet another approach that tries to mine through the thousands of noisy candidate object proposals to find the few relevant regions, we instead propose to take a step back and improve the *initialization step*: specifically, to generate a much smaller yet reliable initial candidate object proposal set such that we can turn an extremely difficult data mining problem into a more manageable one (Fig. 1 (c)). In principle, this sounds straightforward: create a new object proposals method that produces higher precision compared to existing methods. However, the key challenge is to perform this in the weakly-supervised setting *without any bounding box annotations*.

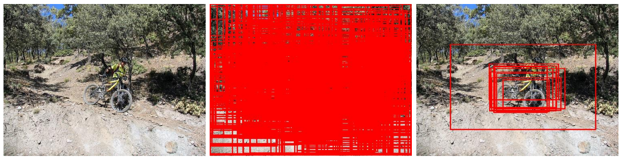
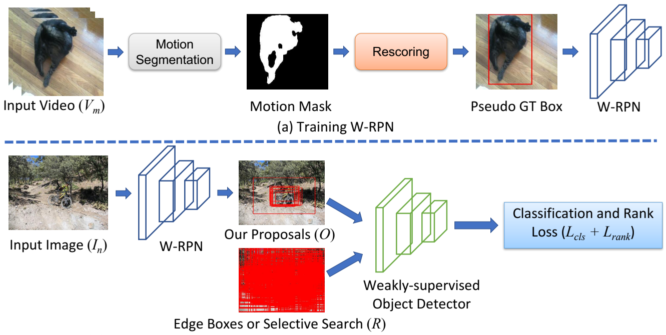

(a) Weakly-labeled image (b) Edge box proposals (c) Our proposals

Figure 1. Given a weakly-labeled image (a), weakly-supervised object detection methods start by generating hundreds to thousands of object proposals (b). We instead generate a few high-precision proposals (c), using a weakly-supervised region proposal network (W-RPN) trained without any bounding box annotations.

To address this challenge, we turn to weakly-labeled *video*, as motion-based segmentation can often provide accurate delineations of objects without any localization annotation. Furthermore, even when trained in the fully-supervised setting, today’s object proposal approaches generate hundreds of object proposals to ensure high recall e.g., [4]—which is fine when ground-truth bounding box annotations are provided—but would still be too many for our weakly-supervised object localization setting. Thus, instead of optimizing for recall, we instead optimize for precision; i.e., we aim to generate $\sim$10 candidate proposals per image while maximizing the chance that the relevant object regions are present in them, which will make the job of the ensuing mining step much easier. But in order to detect all objects, the proposals also need to have high recall, which with $\sim$10 proposals would not be attainable. We therefore use our proposals to rank existing high recall object proposals (e.g., edge boxes [10] or selective search [9]), based on their spatial overlap. To train the weakly-supervised object detector, we formulate a principled end-to-end learning objective that combines: (1) mining class-relevant object regions and (2) ranking of object proposals.

Contributions. We have three main contributions: 1) Unlike existing weakly-supervised object detection approaches, we focus on improving the initial object proposal step to generate a few high precision candidate regions us-

To appear at 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition


Input Video (V_m)
Motion Segmentation
Motion Mask
Rescoring
Pseudo GT Box
W-RPN
Input Image (I_n)
W-RPN
Our Proposals (O)
Weakly-supervised Object Detector
Classification and Rank Loss ($L_{cls} + L_{rank}$)
Edge Boxes or Selective Search (R)

(a) Training W-RPN

(b) Using W-RPN proposals to train weakly-supervised object detection network

Figure 2. (a) Framework for training a weakly-supervised region proposal network (W-RPN) using videos. (b) Once trained, the W-RPN can be used to generate high precision proposals (O) for an input weakly-labeled image (I_n).

ing videos. For high recall, we use spatial overlap with our proposals to rank the (noisy) high recall proposals of methods [9, 10]. 2) We formulate the above two objectives with a principled learning objective that can be optimized end-to-end. 3) Our proposals lead to significant improvement in the performance of state-of-the-art weakly-supervised detection methods on the PASCAL VOC datasets. Our approach generalizes to different weakly-supervised approaches [1, 7] and network architectures.

2. Related work

Weakly-Supervised object detection. In contrast to fully-supervised methods [4], weakly-supervised methods [6, 1, 7] alleviate the need for expensive bounding box annotations, which make them more scalable. However,these methods often suffer from the common drawback of localizing only the most discriminative object part or including co-occurring background regions. This is largely due to these methods solving the very difficult task of mining a small number of true object regions from thousands of noisy proposals per image. In our work, we focus our efforts on finding a few but highly-precise object proposals. While most weakly-supervised detection algorithms learn using images, some also leverage videos, similar to our approach. In particular, [6] transfers tracked boxes from weakly-labeled videos to weakly-labeled images as pseudo ground-truth to train the detector directly on images. However, the transferring is done through non-parametric nearest neighbor matching, which is slow and requires highly-similar video instances for each image instance.

Learning object proposals. Object proposal methods aim to generate candidate object regions for an ensuing detector model. Weakly-supervised object detection methods typically use selective search [9] or edge boxes [10], since these methods do not require bounding box annotations. These proposals have high recall but are noisy in nature. We show that our proposals—which also do not require bounding box annotations but are optimized for precision—can help a weakly-supervised detector down-weight the noisy proposals while focusing on the most relevant ones. Recent work [8] also proposes to learn a weakly-supervised region proposal network. But unlike our approach, it only relies on images and does not make use of videos.

3. Approach

We are given an image dataset $I = \{I_1, \dots, I_N\}$, in which each image is weakly-labeled with object presence labels (e.g., image contains a "dog"). We are also given a video collection $V = \{V_1, \dots, V_M\}$. In some of these videos, we have video-level labels (analogous to image-level labels) and in others, we have no labels whatsoever. There are two main steps to our approach: (1) learning a weakly-supervised region proposal network (W-RPN) on video collection V to generate a few high-precision proposals in the training images in I; (3) using those proposals to bias the selection of relevant object regions when training a weakly-supervised object detector. Fig. 2 shows our entire framework.

3.1. Learning a W-RPN using videos

Motion segmentation in videos. It is well-known that in videos, motion cues can be used to segment objects without any supervision. Thus, to train our W-RPN without any bounding box annotations, we first generate motion segments in each video, and then treat the resulting segmentations as pseudo ground-truth (Fig. 2 (a)). We start by generating unsupervised motion segments in a video. We adapt the unsupervised variant of the Non-Local Consensus Voting (NLC) video segmentation approach by [3]. Given the motion segmentations produced by NLC, we then train a deep convolutional motion segmentation network. Although the motion segmentation network produces good segmentations in frames in which the objects are salient in terms of motion, it does not perform well on frames that are either very blurry or noisy due to compression artifacts. Thus, to automatically choose the good frames on which to train our W-RPN, we perform rescoring based on image classifier and outlier scoring. Please refer [5] for details.

Training W-RPN and proposal generation. We train our W-RPN on the final selected frames. For each frame, we threshold the motion prediction mask to produce a binary segmentation image, and fit a tight bounding box to the largest connected component. We then treat each box as pseudo ground-truth to train the model. Our network architecture is identical to that of the RPN in Faster-RCNN [4]: it produces a binary foreground/background classification score for each candidate proposal and performs bounding box regression. We next use our trained W-RPN to generate high precision candidate object regions in the weakly-labeled image set I.

3.2. Training a weakly-supervised object detection and ranking network

Our proposals can be incorporated into any existing weakly-supervised approach, but in this work, we build upon the Weakly-Supervised Deep Detection Network (WSDDN) [1][2]. Fig. 2 (b) depicts how we use our proposals for training a weakly-supervised object detector. WSDDN takes p proposals of a training image as input and outputs the probability for each of them to belong to C classes. By minimizing a binary log loss summed over each class, it learns to detect objects while being trained for the image classification task:

$$L_{cls}(I_n) = -\sum_{j=1}^{C} c_j \log(s_j) + (1 - c_j)\log(1 - s_j), \quad (1)$$

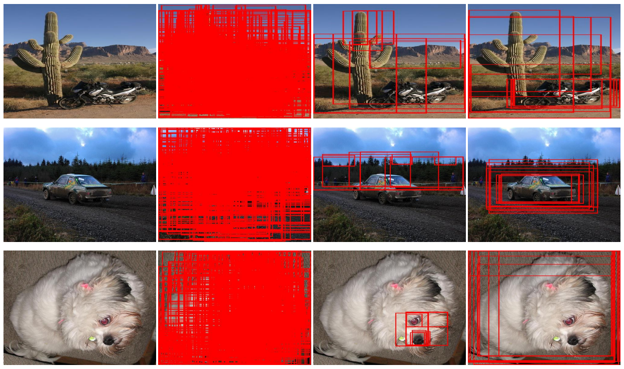
where s_j is the score for class j obtained by summing the class probabilities across all proposals in image I_n and c_j is a binary label whose value is 1 if I_n contains class j.

Compared to the standard setting of having thousands of object proposals, in our setting, we only have a few (k) high-precision proposals for each image. Although this makes the job of WSDDN much easier, it will miss a lot of objects in the dataset since the proposals have low recall. To alleviate this issue, instead of using our proposals directly, we use them to rank the region candidates of an existing proposal approach that has high recall.

Concretely, let $R = \{r_1, r_2, ..., r_p\}$ be the p candidate regions generated using a high recall proposals method like edge boxes [10] or selective search [9], and $O = \{o_1, o_2, ..., o_k\}$ be our k proposals for image I_n. We would like to modify the WSDDN objective such that it not only selects relevant object regions in R that belong to a particular class c_j, but also enforces that those selected regions have high spatial overlap to relevant object proposals in O. To this end, we first compute a class-specific priority score for each region r_i and label c_j pair:

$$P(r_i, c_j) = c_j \cdot \text{IoU}(r_i, o_{i^*}) \cdot s_j(o_{i^*}|W_O), \quad (2)$$

where IoU denotes spatial intersection-over-union, and $o_{i^*} = \arg\max_{o_k \in O} \text{IoU}(r_i, o_k)$ is the highest overlapping proposal in O for r_i. $s_j(o_{i^*}|W_O)$ is the score for class c_j for proposal o_{i^*} which is obtained by first training WSDDN using only our proposals in O. Since an image can be highly-cluttered and contain multiple objects, not every proposal in O will be relevant to class c_j. Thus, this class score modulates the priority so that only those regions in R that have high overlap to *class-relevant* proposals in O end up receiving high priority. Finally, multiplying the priority by c_j ensures that only the classes present in the image produce a non-zero priority score. The priority scores are normalized for every present class to sum to 1.


(a) Weakly-labeled image

(b) Edge boxes proposals

(c) Edge boxes top-10 proposals

(d) Our top-10 proposals

Figure 3. Visualization of our and edge boxes top-10 proposals. Red boxes denote the proposal. Out proposals (d) localize the object tightly more often than the highest scoring edge box proposals (c). Our proposals are used to rank the edge boxes proposals (b).

Using the class-specific priority scores, we then formulate the following rank loss:

$$L_{rank}(I_n) = -\sum_{j=1}^{C} \sum_{r_i \in R} P(r_i, c_j) \cdot \log(s_j(r_i|W_R)), \quad (3)$$

where $s_j(r_i|W_R)$ is the score for class c_j for region r_i, computed by re-training WSDDN using only the regions in R. This ranking loss is inspired by [2] and enforces that the above class specific priority order for the regions in R is maintained. Specifically, this loss function takes two lists of the scores and minimizes the cross-entropy between them. In our case, for regions r_i in R, we have two lists of scores in the form of class specific priority $P(r_i, c_j)$ and class score $s_j(r_i|W_R)$. Hence, this loss will enforce the class scores of the r_i regions to follow the ordering of the class-specific priority scores. As a result, over training, any r_i with a high class-specific priority score will likely end up getting a high class score. Similarly, any r_i with a low class-specific priority score will likely get a lower class score.

Our final loss is the combination of the classification loss and rank loss where λ balances the terms:

$$L_{final}(I_n) = L_{cls}(I_n) + \lambda \cdot L_{rank}(I_n), \quad (4)$$

4. Results

Precision of W-RPN proposals In Fig. 3, we visualize the top-10 proposals of our W-RPN vs. edge boxes. In most

[2] To demonstrate the generalizability of our approach, we also use our proposals with OICR [7]

Method	aero	bike	bird	boat	bottl	bus	car	cat	chair	cow	table	dog	horse	mbk	per	plan	she	sofa	train	tv	mean
Track and Transfer [6]	53.9	-	**37.7**	13.7	-	-	**56.6**	51.3	-	24.0	-	38.5	47.9	47.0	-	-	-	-	48.4	-	-
WSDDN [1]	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
WSDDN + W-RPN (Ours)	**55.9**	**52.6**	27.4	**20.7**	7.8	63.6	54.8	**55.7**	4.9	**37.6**	**35.6**	**59.4**	**52.0**	54.8	**19.6**	12.9	**31.9**	**44.2**	**57.4**	39.2	**39.4**

Table 1. Using our proposals with WSDDN [1] results in a significant boost in detection performance (second row vs. third row) on PASCAL VOC 2007 test set. We also outperform 'Track and Transfer' [6] (first row), which also makes use of weakly-labeled videos.

	VOC 2007		VOC 2012	
Method	CorLoc	mAP	CorLoc	mAP
OICR [7]	60.6	41.2	62.1	37.9
OICR + LP [8]	63.8	45.3	64.9	40.8
OICR + W-RPN (Ours)	**66.5**	**46.9**	**67.5**	**43.2**

Table 2. Using our W-RPN proposals, OICR [7] gets a significant boost on PASCAL VOC 2007 and 2012. We also outperform OICR + LP [8].

cases, our proposals tightly fit the object-of-interest whereas edge box proposals frequently localize object parts or focus on the background. For example, in the third row, our proposals tightly fit the dog's body whereas edge box proposals focus on the dog's face. Overall, our proposals have much higher precision than edge boxes.

Quantitative object detection results. In Table 1, we show the results of training WSDDN (base model VGG L, which is same as VGG16) using edge boxes only (WSDDN) vs. training WSDDN using our W-RPN proposals to rank edge box proposals (WSDDN + W-RPN). Our proposals give a significant boost of 4.6% in mAP. The boost is especially large for objects with distinct discriminative parts (e.g., the face) like person, cat, horse, and dog. For these objects, with thousands of noisy object proposals, the weakly-supervised detector easily latches onto the most discriminative part. In contrast, our W-RPN proposals down-weight such noisy proposals, which in turn leads to significant improvement in detection performance.

We also compare our results with 'Track and Transfer' [6] which like our approach, also uses weakly-labeled videos to improve weakly-supervised object detection. In Table 1, we show our results for the 10 classes reported by 'Track and Transfer'. Again, we obtain a significant boost of 5.7% mAP for these 10 classes (47.6 vs. 41.9). Unlike our approach, 'Track and Transfer' relies on a noisy object proposal method like selective search [9], and transfers tracked boxes from videos to images using non-parametric nearest neighbor matching.

Generalizability of W-RPN proposals. We evaluate how our proposals perform with a different weakly-supervised detection method. Specifically, we take OICR [7], which to our knowledge is the best performing weakly-supervised detection approach with publicly-available code. As OICR is based on WSDDN, we apply L_{rank} with OICR in the same way as with WSDDN. We use selective search proposals and VGG16 as the base model, following the original OICR paper. Table 2 left shows that for PASCAL VOC 2007, we obtain a significant boost of 5.7% mAP and 5.9% CorLoc using our W-RPN proposals (OICR + W-RPN) over OICR with only selective search proposals (OICR). This shows that our proposals are not tied to a specific method, and can generalize to different weakly-supervised approaches.

In Table 2 right, we also measure our performance on the PASCAL VOC 2012 dataset. We show the performance of OICR with and without our proposals. We can see that our approach provides a significant boost of 5.4% and 5.3% for CorLoc and mAP, respectively. This shows that our approach generalizes across different datasets. We also compare to the approach of [8] (OICR + LP), which also trains a weakly-supervised region proposal network but only relies on images. We significantly outperform OICR + LP which shows that using motion cues in videos can lead to higher-quality proposals for weakly-supervised detection.

References

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

[2] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, 2007.

[3] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *CVPR*, 2017.

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

[5] K. K. Singh and Y. J. Lee. You reap what you sow: Using videos to generate high precision objectproposals for weakly-supervised object detection. In *CVPR*, 2019.

[6] K. K. Singh, F. Xiao, and Y. J. Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *CVPR*, 2016.

[7] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017.

[8] P. Tang, X. Wang, A. Wang, Y. Yan, W. Liu, J. Huang, and A. Yuille. Weakly supervised region proposal network and object detection. In *ECCV*, 2018.

[9] J. R. Uijlings, K. E. V. D. Sande, T. Gevers, and A. W. Smeulders. Selective Search for Object Recognition. *IJCV*, 2013.

[10] C. L. Zitnick and P. Dollár. Edge Boxes: Locating Object Proposals from Edges. In *ECCV*, 2014.