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

Abstract

We propose a novel way of using videos to obtain high precision object proposals *for weakly-supervised object detection. Existing weakly-supervised detection approaches use off-the-shelf proposal methods like edge boxes or selective search to obtain candidate boxes. These methods provide high recall but at the expense of thousands of noisy proposals. Thus, the entire burden of finding the few relevant object regions is left to the ensuing object mining step. To mitigate this issue, we focus instead on improving the* precision *of the initial candidate object proposals. Since we cannot rely on localization annotations, we turn to* video *and leverage motion cues to automatically estimate the extent of objects to train a Weakly-supervised Region Proposal Network (W-RPN). We use the W-RPN to generate high precision object proposals, which are in turn used to re-rank high recall proposals like edge boxes or selective search according to their spatial overlap. Our W-RPN proposals lead to significant improvement in performance for state-of-the-art weakly-supervised object detection approaches on PASCAL VOC 2007 and 2012.*

1. Introduction

Object detection has seen tremendous progress in recent years [11, 23, 16, 22]. 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 [33, 6, 32, 27, 17, 25, 1, 28, 34]. 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.

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

Figure 1. Given a weakly-labeled image (a), standard weakly-supervised object detection methods start by generating hundreds to thousands of object proposals (b). The ensuing object localizer must then perform the extremely difficult task of mining the one or two relevant object regions out of all the noisy proposals. We instead generate a few high-precision proposals (c), using a weakly-supervised region proposal network (W-RPN) trained without any bounding box annotations.

Most existing methods start off by using an off-the-shelf object proposal method like selective search [31] or edge boxes [37] 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 create such an algorithm in the weakly-supervised setting *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., [23]—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., computed using edge boxes [37] or selective search [31]), 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.

1.1. 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 using weakly-labeled videos. To also ensure high recall, we use spatial overlap with our proposals to rank the (noisy) high recall proposals of methods like edge boxes or selective search. 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, 28] and network architectures. Code and models are available at https://github.com/kkanshul/w-rpn

2. Related work

Weakly-Supervised object detection. In contrast to fully-supervised methods [11, 23, 16, 22], weakly-supervised methods [33, 6, 26, 27, 5, 25, 17, 1, 28, 34] alleviate the need for expensive bounding box annotations, which make them more scalable. However, existing weakly-supervised 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. We demonstrate that using these proposals to rank the proposals of existing methods [37, 31] can lead to significant improvements in the performance of weakly-supervised object detectors. While most weakly-supervised detection algorithms learn using images, some also leverage videos, similar to our approach. In particular, [20] learns a static image object detector using weakly-labeled videos but is limited by the domain gap between images and videos. To overcome this, [25] instead 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. In our work, we propose to instead train a weakly-supervised region proposals network (W-RPN) using videos, which (in theory) can generate candidate boxes even for loosely-similar image instances. In practice, we find that our W-RPN leads to significant improvement in detection performance over [25].

Learning object proposals. Object proposal methods aim to generate candidate object regions for an ensuing detector or segmentation model; see a great survey by [13]. Early models based on hand-crafted features [2, 7, 4] as well as recent CNN based models [23, 15, 19] require bounding box or segmentation annotated data. Weakly-supervised object detection methods typically use selective search [31] or edge boxes [37], 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 [29] 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. Furthermore, it optimizes for recall rather than precision ($\sim$2000 proposals generated per image).

3. Approach

We are given an image dataset $I = \{I_1, \ldots, I_N\}$, in which each image is weakly-labeled with object presence labels (e.g., image contains a "dog"). We are also given a video collection $V = \{V_1, \ldots, V_M\}$. 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 three main steps to our approach: (1) learning a weakly-supervised region proposal network (W-RPN) on video collection V; (2) using the trained W-RPN 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 de-

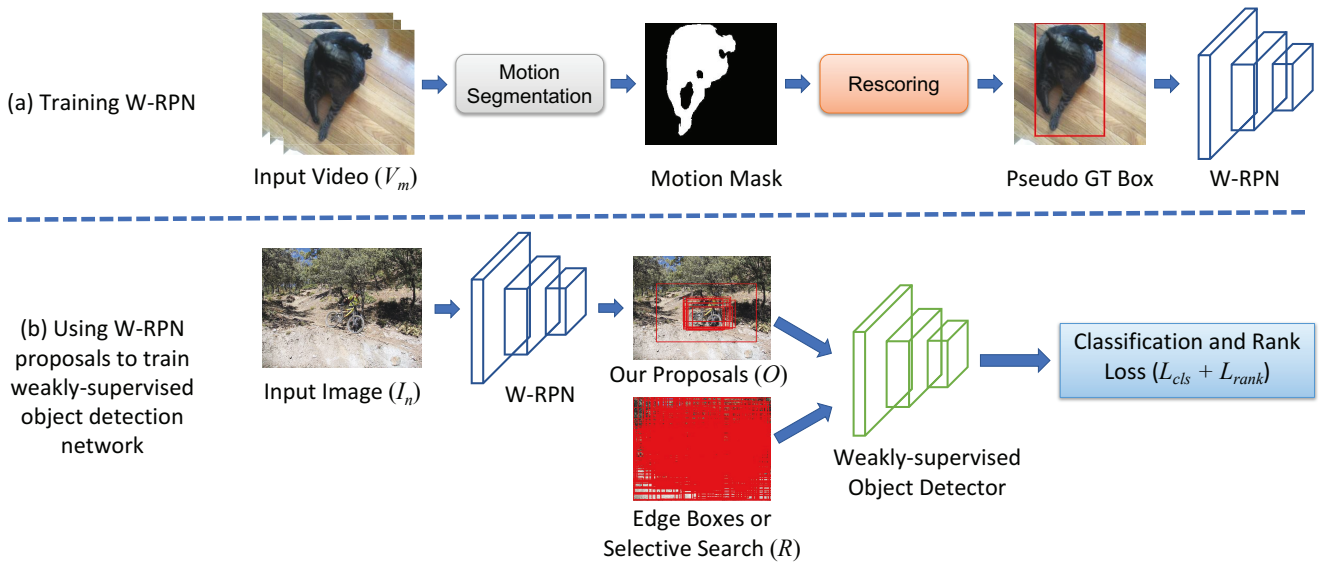

Figure 2. (a) Framework for training a weakly-supervised region proposal network (W-RPN) using videos. Given frames from a video (V_m), we first compute segments using motion cues, and then rescore them according to an automatic measure of segment quality. High scoring motion segments are used to compute pseudo ground-truth boxes to train the W-RPN. (b) Once trained, the W-RPN can be used to generate high precision proposals (O) for an input weakly-labeled image (I_n). These high precision proposals are used to rank the high recall proposals of edge boxes or selective search (R) according to their spatial overlap during the training of a weakly-supervised object detector.

tector. Fig. 2 shows our entire framework.

3.1. Learning a W-RPN using videos

The first step is to train a weakly-supervised region proposal network (W-RPN) without any bounding box annotations. For this, we make use of motion cues in video to automatically identify the extent of objects.

3.1.1 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)). A similar idea has been explored previously for self-supervised feature learning [18] and for transferring boxes from videos to images for weakly-supervised detection [25]. However, to our knowledge, this idea has not been explored for learning object proposals in the weakly-supervised setting.

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 [18][1]. Briefly, NLC computes a per-pixel motion saliency in which pixels that move differently from their surroundings are considered salient. The per-pixel saliency scores are averaged over superpixels, and then propagated to other frames via a nearest neighbor voting scheme where the neighbors are determined by appearance and spatial position features. We refer the reader to [10] for more details. We apply NLC on the YFCC100m video dataset [30] and on the YouTube-Bounding Boxes video dataset [21]. For the latter, we apply it only on videos that have the same weak-labels as our weakly-labeled images in I.

Given the motion segmentations produced by NLC, we then train a deep convolutional motion segmentation network, similar to [18]. However, in [18], the network is based on the AlexNet architecture, and the final output layer is a fully-connected layer that produces a motion prediction for a fixed grid size (56×56). We instead use the Pyramid Scene Parsing Network [35] which has a fully-convolutional output layer rather than a fully-connected one. This allows us to take in arbitrary resolution video frames, which we find to produce more reliable motion segmentations.

3.1.2 Rescoring of motion segments

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 the following operations.

[1]The difference between this and the original NLC [10] is that [10] relies on an edge detector trained on labeled edge images, whereas the implementation of [18] that we use is completely unsupervised.

First, we train an image classifier on our weakly-labeled images I using their corresponding weak object labels, and then fire the classifier on each video frame. We take the frames that produce the highest classification scores (in our experiments, we take the top 1000 frames per category-of-interest) for the corresponding object classifier. This not only chooses frames that have relevant objects that we ultimately care about, but it also tends to choose *image-like* frames, which can be beneficial when we apply our W-RPN on images since these video frames will have a smaller domain difference to images and thus will generalize better.

Second, for each selected frame, we score how well the object-of-interest is segmented. Since we do not have ground-truth, we cannot know for sure how good the segmentation actually is. However, we can assume that a frame in which the object "stood-out" in terms of motion with respect to its surrounding background would likely have resulted in a more reliable motion segmentation. To this end, we sort the frames according to the outlier score proposed in [12]. Specifically, for each frame with per-pixel motion prediction scores, we treat as an outlier any pixel p whose motion prediction score m_p is either $> 1.25 \times Q1$ or $< 1.25 \times Q3$, where $Q1$ and $Q3$ denote the first and third quantile motion score, respectively. Then the outlier score m_f for frame f is computed by summing the outlier motion scores and normalizing by the sum of all pixel motion scores in the frame: $m_f = \sum_{p \in Outlier}(m_p) / \sum_p (m_p)$. If the motion prediction scores are uniformly distributed, the frame outlier score will be low, whereas if the motion prediction scores have a peaky distribution, the frame outlier score will be high. We choose the top 500 frames per category-of-interest with the highest scores. Figure 3 shows some video frames and corresponding thresholded motion masks with high and low outlier scores.

3.1.3 Training W-RPN

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 [23]: it produces a binary foreground/background classification score for each candidate proposal and performs bounding box regression. To further refine our proposals, we also pass them through an additional bounding-box regressor module. (We found this extra refinement step to produce tighter boxes.) We apply NMS for the initial generated boxes, but do not perform NMS for the final refined outputs. This is mainly because in the weakly-supervised setting, our pseudo ground-truth boxes cannot be completely trusted; thus, we do not want to suppress a box just because there is another box nearby with a higher score since the lower scoring box could in fact be a better proposal for the object. Although this produces some redundant boxes (as shown in Fig. 1 (c)), overall, we find that this leads to higher precision.

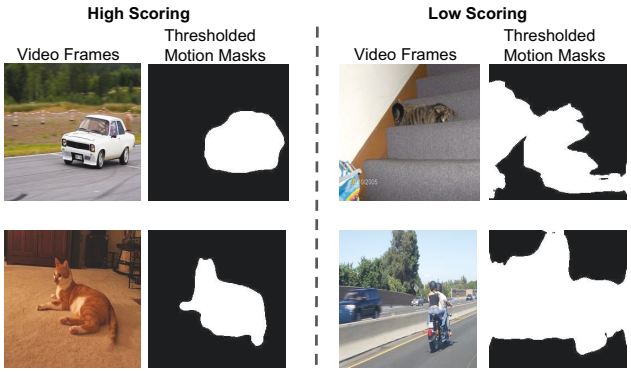

Figure 3. Visualization of thresholded motion masks for video frames with high vs. low outlier scores. The frames on the left whose motion masks have higher outlier scores have more accurate motion masks compared to the frames on the right with lower outlier scores. We use the outlier score to decide the reliable frames for training the W-RPN.

3.2. Generating high precision proposals on weakly-labeled images

We next use our trained W-RPN to generate high precision candidate object regions in the weakly-labeled image set I.

Since we care more about precision than recall, we purposely take only a handful of confident object proposals in each image. To maximize the chance that those proposals are fired on the relevant objects, we take the image classifier from above, and produce a class activation map (CAM) [36] for the corresponding class for each image. CAM produces high-scores for regions that contributed to the final classification; i.e., regions that are relevant to the class. We mask out very low scoring regions, and fire the W-RPN on the masked image. To account for any noise in the CAM predictions, we also fire the network on the original image. Our final candidate list of proposals for an image is the top-k (where $k \approx 10$) proposals computed on the original image and the masked image.

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

Now that we have a set of high precision object proposals in each image in I, the final step is to train a weakly-supervised detector. 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] as many recent state-of-the-art

approaches use it as initialization.[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 [37] or selective search [31], 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.

[2]To demonstrate the generalizability of our approach, we also use our proposals with OICR [28], a recent state-of-the-art weakly-supervised detection method based on WSDDN.

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 [3] 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:

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

where λ balances the terms.

Ultimately, the rank loss L_{rank} influences which regions in R should get high class probability while minimizing the classification loss L_{cls}. As a result, WSDDN learns to provide higher class probabilities to proposals r_i in R which have high overlap with our proposals in O; i.e., those that are more likely to contain the whole object. Due to this, during testing, we only need the R candidate regions and our proposals O are no longer needed, as the network will have learned to incorporate the class-specific priorities.

4. Results

We quantitatively measure three different aspects of our W-RPN proposals: 1) how precisely they cover the relevant objects, 2) improvement in weakly-supervised object detection performance by using them, and 3) generalizability across different network architectures and approaches. We also show qualitative detection results.

4.1. Datasets

We evaluate on the PASCAL VOC 2007 [8] dataset, which is the most widely used dataset for weakly-supervised object detection. It consists of 20 object categories. For training, we use the *trainval* set (5011 images) and evaluate on the *test* set (4952 images). We also compute results on PASCAL VOC 2012 [9]. We again train using the *trainval* set (11,540 images) and evaluate on the *test* set

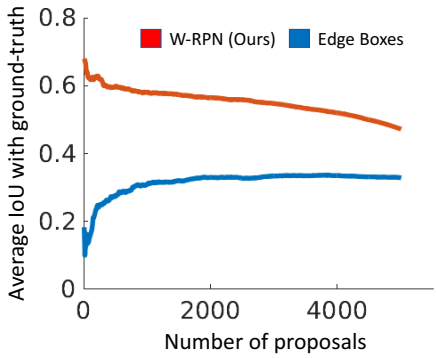

Figure 4. Average IoU of our top-n W-RPN proposals (red) and edge box proposals (blue) with the ground-truth boxes. Our top proposals localize the objects much more accurately.

(10,991 images). For evaluation, a predicted box is correct if it has more than 50% IoU with the ground-truth box. We compute Average Precision (AP) on the *test* set, and CorLoc on the *trainval* set which measures the percentage of training images of a class for which the most confident detected box has at least 50% overlap with at least one of the ground-truth instances.

4.2. Implementation details

We randomly choose 500 video clips per class from the YouTube-Bounding Boxes video dataset [21] to train the motion segmentation network. There are only 14 overlapping classes between YouTube-Bounding Boxes and PASCAL VOC 2007, so for the remaining 6 PASCAL classes, we download videos (500 for each class) from YouTube. We train the image classifier on the PASCAL dataset to choose good video frames for training the W-RPN. To ensure diversity, we add constraints that the selected frames are at least 3 seconds apart from one another and at max 5 frames per video are chosen. For motion prediction, we use Pyramid Scene Parsing Network [35] with Resnet34 and dilated convolution for the encoder, and pyramid pooling with bilinear upsampling for the decoder.

We threshold the motion prediction mask at 0.05 motion score to create a binary mask. For W-RPN training, any proposal with IoU greater than 0.7 with a pseudo ground-truth box is considered positive, and any proposal with IoU less than 0.3 is considered negative. We use VGG16 [24] as the base architecture for the W-RPN. For L_{rank}, we only consider regions in R whose maximum IoU with our proposals (O) is greater than 0.7 or below 0.3 but greater than 0. This avoids providing any priority to uncertain proposals in R which have neither too high nor too low overlap with our proposals in O. We also do not want to assign priority to a region which does not overlap with our proposals at all because O has very few proposals and it is possible that we could have missed some objects. For all experiments,

(a) Weakly-labeled image (b) Edge boxes proposals (c) Edge boxes top-10 proposals (d) Our top-10 proposals

Figure 5. 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).

we set $k = 10$; i.e., top-10 proposals (O) for every image in I. For training WSDDN [1] and OICR [28], we use the publicly available codes and follow their paper protocols.

4.3. Precision of W-RPN proposals

We first measure how well our W-RPN proposals (O) wholly contain the object-of-interest compared to edge box [37] proposals (R). For both our approach and edge boxes, we take the highest scoring proposal in each training image, and then sort the resulting proposals in descending order of their score. For each proposal, we then compute its IOU with the highest-overlapping ground-truth box. Fig. 4 shows the average IoU for the top-n proposals (where n varies from 1 to number of training images). We can see that our W-RPN proposals (red) have a much higher average IoU with the ground-truth boxes compared to the edge box proposals (blue). This indicates that our top proposals are much more precise and more likely to fully contain the object-of-interest. For 18 out of 20 classes (including both moving and stationary objects), our proposals get much better average ground-truth IoU compared to edge boxes; see

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 [25]	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. Detection mAP of WSDDN [1] with and without our W-RPN proposals on PASCAL VOC 2007 test set. Using our proposals with WSDDN results in a significant boost in detection performance (second row vs. third row). We also compare to 'Track and Transfer' [25] (first row), which like our approach also makes use of weakly-labeled videos to improve detection in weakly-labeled images. For most of the 10 classes that 'Track and Transfer' reported results on, our approach outperforms it by a big margin.

supplementary material for per-class IoU curves. The reason why our W-RPN can work well even for stationary objects is that in the videos that the W-RPN is trained on, as long as the background is changing with respect to the foreground stationary object e.g., due to camera motion, it can be segmented out.

In Fig. 5, we visualize the top-10 proposals of our W-RPN vs. edge boxes. In most 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, and can significantly improve weakly-supervised object detection performance as we show next.

4.4. Quantitative object detection results

We next evaluate the impact of our W-RPN proposals for weakly-supervised object detection. As described in the approach, we build upon the WSDDN [1] detector as it is the basis of many recent weakly-supervised approaches [14, 28, 34]. 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' [25] 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 [31], and transfers tracked boxes from videos to images using non-parametric nearest neighbor matching. This is not only slow, but also less likely to generalize to novel instances as the direct matching requires a very similar video instance for each image instance.

	VOC 2007		VOC 2012	
Method	CorLoc	mAP	CorLoc	mAP
OICR [28]	60.6	41.2	62.1	37.9
OICR + LP [29]	63.8	45.3	64.9	40.8
OICR + W-RPN (Ours)	**66.5**	**46.9**	**67.5**	**43.2**

Table 2. OICR [28] performance with and without our W-RPN proposals. Using our W-RPN proposals, OICR gets a significant boost on PASCAL VOC 2007 and 2012. We also outperform OICR + LP [29] which also learns proposals in the weakly-supervised setting but does not make use of videos.

In order to evaluate the importance of L_{rank}, we tried simply combining our top-10 proposals with edge box proposals when training WSDDN (instead of re-ranking the edge boxes with L_{rank}). This baseline only gives a minor boost of 0.6% over using only edge box proposals (WSDDN), which shows that re-ranking the edge box proposals with our proposals is important. Lastly, our approach does not require computing L_{rank} during inference; i.e., as mentioned in Sec. 3.3, we do not need to compute our proposals during testing since the detector will have learned to highly score the high recall object proposals that are more likely to contain the whole object. Hence, it does not change the runtime speed of weakly-supervised detection methods like WSDDN and OICR [28].

4.5. Generalizability of W-RPN proposals

We measure the generalizability of our W-RPN proposals across different network architectures, weakly-supervised approaches, and datasets. We first measure the improvement obtained by our proposals for WSDDN with three different base network architectures (S:small, M:medium, and L:large) [1]. Using our proposals with WSDDN, we obtain a significant boost of 3.8%, 4.1%, and 4.6% for the S, M, and L model, respectively. This shows that our proposals can generalize across different network architectures.

Next, we evaluate how our proposals perform with a different weakly-supervised detection method. Specifically, we take OICR [28], which to our knowledge is the best performing weakly-supervised detection approach with

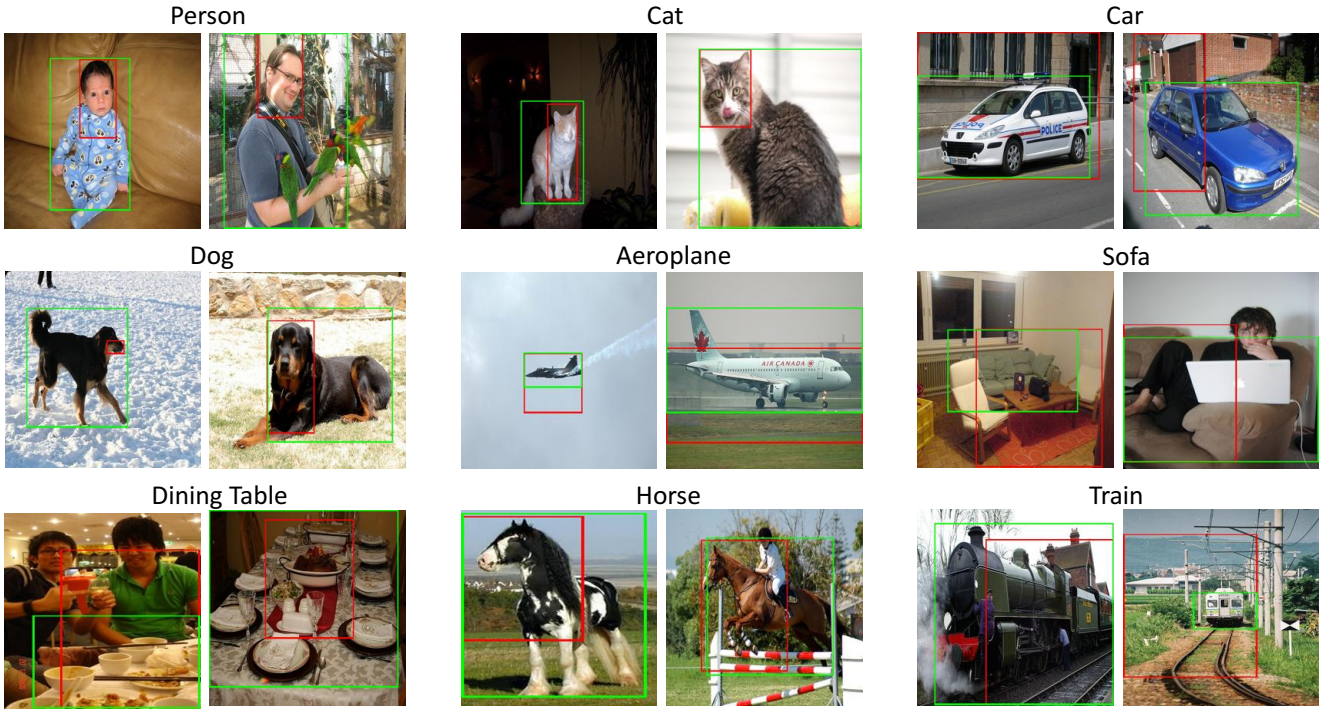

Figure 6. Detection results of WSDDN with (green box) and without (red box) our proposals. Our proposals often lead to better detections as the network learns to detect the whole object rather than focusing only on a discriminative part or co-occurring context. In these examples, by using our proposals, WSDDN is able to detect the full extent of cat, dog, and person whereas it only focuses on their faces without our proposals.

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 [24] as the base model, following the original OICR paper. Table 2 left shows that for PASCAL VOC 2007 [8], 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 [9]. 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 [29] (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.

4.6. Qualitative results

Finally, Fig. 6 shows example detections of WSDDN [1] with (green box) and without (red box) our proposals. Without our proposals, WSDDN tends to focus only on the discriminative part of an object (like the head of a person and cat) rather than the entire object. Also, in the cases of airplane and car, it localizes the co-occurring background. In contrast, our proposals bias the detector to downweight the noisy regions and to instead focus on full object regions.

5. Conclusion

We presented a novel method of using motion information in weakly-labeled videos to learn high precision object proposals. The proposals are used to rank candidate object regions of existing high recall proposal approaches like edge boxes and selective search. Our experiments showed that incorporating our proposals into existing weakly-supervised detection approaches lead to substantial improvement in detection performance. We hope this work will inspire further work on generating good object proposals without strong supervision.

Acknowledgments. This work was supported in part by NSF CAREER IIS-1751206, IIS-1748387, AWS ML Research Award, Google Cloud Platform research credits program, and GPUs donated by NVIDIA.

References

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

[2] Alexe Bogdan, Thomas Deselaers, and Vittorio Ferrari. What is an Object? In *CVPR*, 2010.

[3] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, 2007.

[4] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010.

[5] Ramazan Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. In *arXiv:1503.00949*, 2015.

[6] David J. Crandall and Daniel P. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*, 2006.

[7] Ian Endres and Derek Hoiem. Category Independent Object Proposals. In *ECCV*, 2010.

[8] Mark Everingham, Luc van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[9] Mark Everingham, Luc van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[10] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014.

[11] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[12] Brent A. Griffin and Jason J. Corso. Video object segmentation using supervoxel-based gerrymandering. In *arXiv:1704.05165*, 2017.

[13] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? In *arXiv:1502.05082*, 2015.

[14] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, 2016.

[15] Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. Deepbox: Learning objectness with convolutional networks. In *ICCV*, 2015.

[16] 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.

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

[18] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, 2017.

[19] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NIPS*, 2015.

[20] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning Object Class Detectors from Weakly Annotated Video. In *CVPR*, 2012.

[21] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *arXiv:1702.00824*, 2017.

[22] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *CVPR*, 2017.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] 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.

[26] Parthipan Siva, Chris Russell, and Tao Xiang. In Defence of Negative Mining for Annotating Weakly Labelled Data. In *ECCV*, 2012.

[27] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, 2014.

[28] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017.

[29] 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.

[30] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016.

[31] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective Search for Object Recognition. *IJCV*, 2013.

[32] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014.

[33] Markus Weber, Max Welling, and Pietro Perona. Unsupervised Learning of Models for Recognition. In *ECCV*, 2000.

[34] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *CVPR*, 2018.

[35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[37] C. Lawrence Zitnick and Piotr Dollár. Edge Boxes: Locating Object Proposals from Edges. In *ECCV*, 2014.