Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection

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

Abstract

The status quo approach to training object detectors requires expensive bounding box annotations. Our framework takes a markedly different direction: we transfer tracked object boxes from weakly-labeled videos to weakly-labeled images to automatically generate pseudo ground-truth boxes*, which replace manually annotated bounding boxes. We first mine discriminative regions in the weakly-labeled image collection that frequently/rarely appear in the positive/negative images. We then match those regions to videos and retrieve the corresponding tracked object boxes. Finally, we design a hough transform algorithm to vote for the best box to serve as the pseudo GT for each image, and use them to train an object detector. Together, these lead to state-of-the-art weakly-supervised detection results on the PASCAL 2007 and 2010 datasets.*

1. Introduction

Object detection is a fundamental problem in computer vision. While tremendous advances have been made in recent years, existing state-of-the-art methods [9, 30, 12, 11] are trained in a strongly-supervised fashion, in which the system learns an object category's appearance properties and precise localization information from images annotated with bounding boxes. However, such carefully labeled exemplars are expensive to obtain in the large numbers that are needed to fully represent a category's variability, and methods trained in this manner can suffer from unintentional biases or errors imparted by annotators that hinder the system's ability to generalize to new, unseen data [35].

To address these issues, researchers have proposed to train object detectors with relatively inexpensive *weak supervision*, in which each training image is only weakly-labeled with an image-level tag (e.g., "car", "no car") that states an object's presence/absence but not its location [39, 10, 26, 32, 33, 3]. These methods typically mine discriminative visual patterns in the training data that frequently occur in the images that contain the object and rarely in the images that do not. However, due to scene clutter, intra-class appearance variation, and occlusion, the discriminative patterns often do not tightly fit the object-of-interest; they either correspond to a small part of the object such as a car's wheel instead of the entire car, or include the surrounding context such as a car with portions of the surrounding road (Fig. 1 bottom, green boxes). Consequently, the detector that is trained using these patterns performs substantially worse than strongly-supervised algorithms.

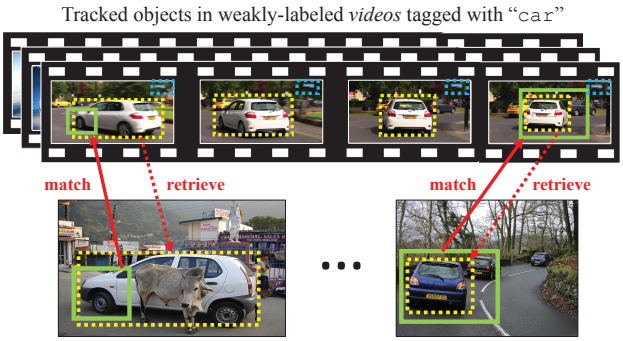

Figure 1. **Main idea.** (**top**) Automatically tracked objects (yellow and blue boxes) in weakly-labeled videos *without any human initialization*. (**bottom**) Discriminative visual regions (green boxes) mined in weakly-labeled training images. For each discriminative region, we find its best matching region across all videos, and retrieve its overlapping tracked object box (yellow dotted box) back to the image. The retrieved boxes are used as *pseudo ground-truth* to train an object detector. Our approach improves object localization by expanding the initial visual region beyond a small object part (**bottom-left**) or removing the surrounding context (**bottom-right**). In practice, we combine the retrieved boxes from multiple visual regions in an image to produce its best box.

Main idea. So, how can we create accurate object detectors that do not require expensive bounding box annotations? Our key idea is to use motion cues from videos as a *substitute for strong human supervision*. Given a weakly-labeled image collection and videos retrieved using the same weak-label (e.g., "car"), we first automatically track and localize candidate objects in the videos, and then transfer their relevant tracked object boxes to the images. We transfer the object boxes by mining discriminative visual regions in the image collection, and then matching them to regions in the videos. See Fig. 1.

Since temporal contiguity and motion signals are lever-

aged to localize and track the objects in video, their transferred boxes can provide precise object localizations in the weakly-labeled images. Specifically, they can expand the initial discovered region to provide a fuller coverage of the object, or decrease the spatial extent of the initial discovered region to remove the surrounding context (Fig. 1 bottom, yellow boxes). We then use the transferred boxes to generate *pseudo ground-truth* bounding boxes on the weakly-labeled images to train an object detector, replacing standard human-annotated bounding boxes. To account for noise in the discovered discriminative visual regions, video tracking, and image-to-video matches, we retrieve a large set of object boxes and combine them with a hough transform algorithm to produce the best boxes.

What is the advantage of transferring object boxes to images instead of directly learning from videos? In general, images provide more diverse *intra-category* appearance information than videos, especially given the same amount of data (e.g., a 1000-frame video with a single object instance vs. 1000 images with $\sim$1000 different object instances), and are often of higher quality since frames from real-world (e.g., YouTube) videos typically suffer from motion blur and compression artifacts. Importantly, in this way, our framework opens up the possibility to leverage the huge *static* imagery available online, much of which is already weakly-labeled.

Contributions. In contrast to existing strongly-supervised object detection systems that require expensive bounding box annotations, or weakly-supervised systems that rely solely on appearance-based grouping cues within the image dataset, we instead transfer tracked object boxes from videos to images to serve as *pseudo ground-truth* to train an object detector. This eliminates the need for expensive bounding box annotations, and compared to existing weakly-supervised algorithms, our approach provides more complete and tight localizations of the discovered objects in the training data. Using videos from the YouTube-Objects dataset [28], we demonstrate that this leads to state-of-the-art weakly-supervised object detection results on the PASCAL VOC 2007 and 2010 datasets.

2. Related Work

Weakly-supervised object detection. While recent state-of-the-art *strongly-supervised* methods [12, 30, 15, 11] using deep convolutional neural networks (CNN) [18, 17] have shown great object detection accuracy, they require thousands of expensive bounding-box annotated images.

To alleviate expensive annotation costs, weakly-supervised methods [39, 10, 26, 32, 33, 3] train models on images labeled only with object presence/absence labels, without any location information of the object. Early efforts [39, 10] focused on simple datasets with a single prominent object in each image (e.g., Caltech-101). Since then, a number of methods [7, 26, 32, 33, 34, 4, 25] learn detectors on more realistic and challenging datasets (e.g., PASCAL VOC [27]). The main idea is to identify discriminative regions that frequently appear in positive images and rarely in negative ones. However, their central weakness is that due to large intra-category appearance variations, occlusion, and background clutter, they often mislocalize the objects in the training images, which results in sub-optimal detectors. We address this challenge by matching the discriminative regions to videos to retrieve automatically-tracked object boxes back to the images. This results in better localization on the weakly-labeled training set, which leads to more accurate object detectors.

Learning with videos. Video offers something that static images cannot: it provides motion information, a strong cue for grouping objects (the "law of common fate" in Gestalt psychology). Existing methods learn part-based animal models [29], learn detectors from images while using video patches for regularization [20], or augment training data from videos for single-image action recognition [2]. While some work consider learning object category models directly from (noisy) internet videos [28, 14, 23], we are exploring a rather different problem: we use video data to simulate human annotations, but ultimately use *image data* to train our models. Critically, this allows our framework to potentially take advantage of the huge static image data available on the Web, which existing video-only learning methods cannot.

Finally, recent work uses videos for semi-supervised object detection with bounding box annotations as initialization [21], or trains a CNN for feature learning using tracking as supervision and fine-tuning the learned representation with bounding box annotations for detection [38]. In contrast, we do not require *any* bounding box annotations.

3. Approach

We are given a weakly-labeled image collection $S_I=\{I_1,\ldots,I_N\}$, in which images that contain the object-of-interest (e.g., "car") are labeled as positive and the remaining images are labeled as negative. We are also given a weakly-labeled video collection $S_V=\{V_1,\ldots,V_M\}$ whose videos contain the positive object-of-interest, but where and when in each video it appears is unknown.

There are three main steps to our approach: (1) identifying discriminative visual regions in S_I that are likely to contain the object-of-interest; (2) matching the discriminative regions to tracked objects in videos in S_V and retrieving the tracked objects' boxes back to the images in S_I; and (3) training a detector using the images in S_I with the retrieved object boxes as supervision.

3.1. Mining discriminative positive visual regions

We first mine discriminative visual regions in the image collection S_I that frequently appear in the positive images

Figure 2. Example positive regions in the top-4 automatically mined discriminative clusters for aeroplane, bird, boat, car, cat, cow, dog, horse, motorbike, and train. While the discovered regions are relevant to the positively-labeled object category, most of them do not localize the object well, capturing only an object-part (e.g., cat, cluster 1) or including the surrounding context (e.g., aeroplane, cluster 2).

and rarely in the negative ones; these regions will likely correspond to the object-of-interest or a part of it. For this, we follow a similar approach to [31, 8, 33, 34].

For each image in S_I, we generate $\sim$2000 object proposals (rectangular regions) using selective search [36], and describe each proposal with a *pool5* activation feature using AlexNet [17] pre-trained for ImageNet classification. For each region, we find its best matching (nearest neighbor) region in each image in S_I (regardless of image label) using cosine similarity. Each region and its k closest nearest neighbors form a cluster. We then rank the clusters in descending order of the number of cluster instances that are from the positive images. Since we create clusters for every region in every image, many will be redundant. We therefore greedily remove near-duplicate clusters that contain many near-identical regions to any higher-ranked cluster, as measured by spatial overlap of more than 25% IOU between 10% of their cluster members. Finally, for each remaining cluster, we discard any negative regions.

Let $\mathcal{P}$ be the set of all positive regions in the top-C ranked clusters. While $\mathcal{P}$ contains many diverse and discriminative regions of the object-of-interest (see Fig. 2), most of the regions will not tightly *localize* the object for three main reasons: (1) the most discriminative regions usually correspond to object-parts, which tend to have less appearance variation than the full-object (e.g., face vs. full-body of a cat), (2) co-occurring "background" objects are often included in the region (e.g., airplane with sky), and (3) most of the initial object proposals are noisy and do not tightly fit any object to begin with. Thus, the regions in $\mathcal{P}$ will be sub-optimal for training an object detector, since they are not well-localized; this is the central weakness of all existing weakly-supervised methods. We next explain how to use videos labeled with the same weak-label (e.g., "car") to improve the localization.

3.2. Transferring tracked object boxes

For now, assume that we have a (noisy) object track in each video in S_V, which fits a bounding box around the positive object in each frame that it appears. In Sec. 3.4, we explain how to obtain these tracks.

For each positive image region in $\mathcal{P}$, we search for its n best matching video regions across all videos in S_V and return their corresponding tracked object boxes to improve the localization of the object in its image. There is an important detail we must address to make this practical: matching with *fc7* features (of AlexNet [17]) can be prohibitively expensive, since each candidate video region (e.g., selective search proposal) would need to be warped to 227x227 and propagated through the deep network, and there can be $\sim$2000 such candidate regions in every frame, and millions of frames. Instead, we perform matching with *conv5* features, which allows us to forward-propagate an entire video frame just once through the network since convolutional layers do not require fixed-size inputs. To compute the *conv5* feature maps, we use deep pyramid [13], which creates an image pyramid with 7 levels (where the scale factor between levels is $2^{-1/2}$) and computes a *conv5* feature map for each level (for the 1st level, the input frame is resized such that its largest dimension is 1713 pixels). We then match each positive image region to each frame in each video densely across location and scale in a sliding-window fashion in *conv5* feature space, using cosine similarity. Note that this restricts matching between regions with similar aspect ratios, which can also help reduce false positive matches.

Given a positive image region's n best matching video regions, we return each of their frame's tracked object bounding box (if it has any spatial overlap with the matched video region) back to the positive region's image, while preserving relative translation and scale differences. Specifi-

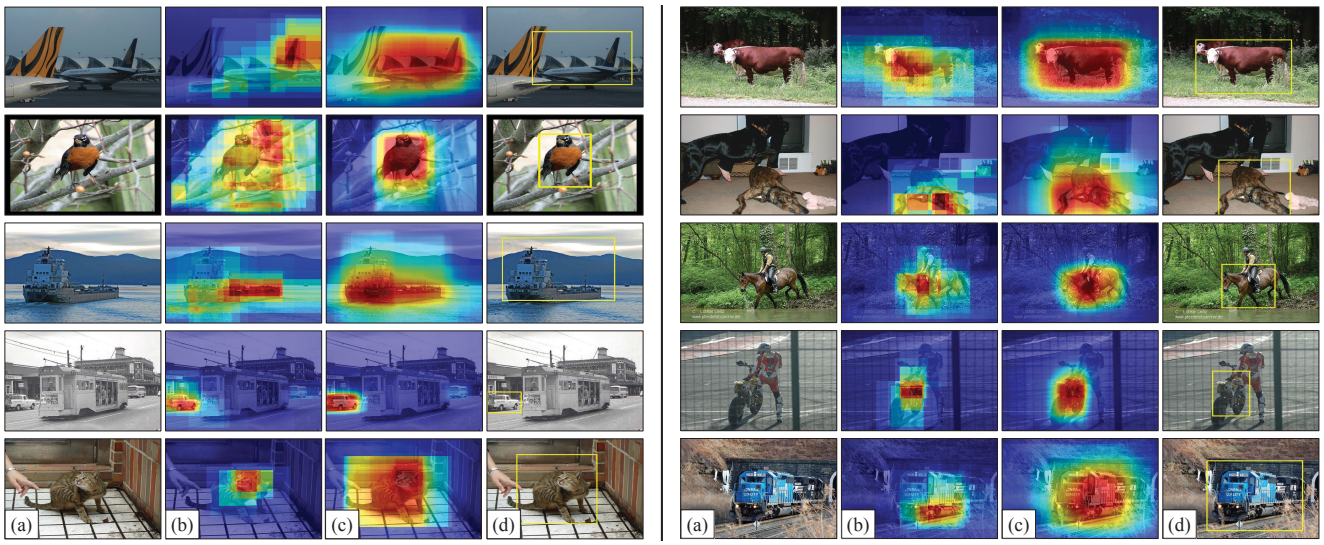

Figure 3. (a) Weakly-labeled positive image for aeroplane, bird, boat, car, cat, cow, dog, horse, motorbike, and train. (b) Heatmap showing the distribution of the initial discriminative positive regions found in the image. (c) Heatmap showing the distribution of the transferred video object boxes in the image. (d) Our automatically discovered pseudo ground-truth box. Notice how the initial discriminative regions focus more on object-parts, whereas the transferred boxes focus more on the full object. This leads to better localization of the object in the weakly-labeled positive image. **Best viewed on pdf.** *Results for all images can be found in the supplementary material.*

(a) Positive image region (b) Best matching video region and tracked object box (c) Retrieved box conditioned on matching region

Figure 4. We match a positive image region (a) to all video frames in a sliding-window fashion, and for the best matching vido region (green box) (b), we retrieve its overlapping tracked object box (yellow dotted box) back to the image (c).

cally, we can parameterize any region with its top-left and bottom-right coordinate values: $[x_{min}, y_{min}, x_{max}, y_{max}]$. Denote a positive image region as $\boldsymbol{r}$, its matched video region as $\boldsymbol{v}$, and the corresponding overlapping tracked region as $\boldsymbol{t}$. Then, the returned bounding box $\boldsymbol{r'}$ is: $\boldsymbol{r'} = \boldsymbol{r} + (\boldsymbol{t} - \boldsymbol{v})$. See Fig. 4. We repeat this for all n best matching video regions, and for each positive region in $\mathcal{P}$.

Each positively-labeled image in S_I (that has at least one positive region) now has a set of retrieved bounding boxes, up to n from each positive region in the image. Some will tightly fit the object-of-interest, while others will be noisy due to incorrect matches/tracks. We thus use the hough transform to vote for the best box in each image. Specifically, we create a 4-dimensional hough space in which each box casts a vote for its $[x_{min}, y_{min}, x_{max}, y_{max}]$ coordinates. We select high density regions in the continuous hough space with mean-shift clustering [5], which helps the voting be robust to noise and quantization errors [19]. The total vote for box coordinate $\boldsymbol{l}$ is a weighted sum of the votes in its spatial vicinity:

$$vote(\boldsymbol{l}) = \sum_i vote(\boldsymbol{r'}_i) \cdot K\Big(\frac{\boldsymbol{l} - \boldsymbol{r'}_i}{b}\Big), \qquad (1)$$

where the kernel K is a radially symmetric, non-negative function centered at zero and integrating to one, b is the mean-shift kernel bandwidth, i indexes over the positive regions in the image, and $vote(\boldsymbol{r'}_i)$=1, $\forall i$. We select $\boldsymbol{l}$ with the highest vote as the final box for the image. If the highest vote is less than a threshold $\theta = 20$, then there is not enough evidence to trust the box so we discard it. See Fig. 3 (c-d) for example distributions of the transferred bounding boxes and final selected bounding box. We repeat this hough voting process for each positively-labeled image in S_I.

3.3. Training an object detector

We can now treat the final selected boxes as pseudo ground-truth (GT)—as a *substitute for manually annotated boxes*—to train an object detector, with any algorithm developed for the strongly-supervised setting. We use the state-of-the-art Regions with CNN (R-CNN) system [12]. Briefly, R-CNN computes CNN features over selective search [36] proposals, trains a one-vs-all linear SVM (with GT boxes as positives and proposals that have less than 0.3 intersection-over-union overlap (IOU) with any GT box as negatives) to classify each region, and then performs bounding box regression to refine the object's detected location.

There are three considerations to make when adapting R-CNN to our work: (1) each positively-labeled image has at most one pseudo GT box, which means that negative regions from the same image must be carefully selected since the image could have multiple positive instances (e.g., multiple cars in a street scene) but our pseudo GT may only be

covering one of them; (2) some positively-labeled images may have no pseudo GT box (i.e., if there were not enough votes), which means that we would not be making full use of all the positive images; and (3) some pseudo GT boxes may be inaccurate even after hough voting due to noise in the matching or tracking. These can all lead to a sub-optimal detector if not handled carefully.

To address the first issue, we train an R-CNN model with the pseudo GT boxes as positives, and any selective search proposal that has an IOU less than 0.3 *and* greater than 0.1 with a pseudo GT box as negatives. In this way, we minimize the chance of mistakenly labeling a different positive instance in the image as negative, but at the same time, select mis-localized regions (that have some overlap with a pseudo GT) as *hard-negatives*. We treat all selective search proposals in any negatively-labeled image in S_I as negative.

To address the second and third issues, we perform a latent SVM (LSVM) update [9] given the initial R-CNN model from above to update the pseudo GT boxes. For images that do not have a pseudo GT box, we fire the R-CNN model and take its highest-scoring detection in the image as the pseudo GT box. For images that already have a pseudo GT box, we take the highest-scoring detection that has at least 0.5 IOU with it, which prevents the updated box from changing too much from the initial box. We then re-train the R-CNN model with the updated pseudo GT boxes.

Finally, we also fine-tune the R-CNN model to update not only the classifier but also the features using our pseudo GT boxes, which results in an even greater boost in detection accuracy (as shown in Sec. 4.3). Fine-tuning CNN features has not previously been demonstrated in the weakly-supervised detection setting, likely due to existing methods producing too many false detections in the training data. Our discovered pseudo GT boxes are often quite accurate, making our approach amenable for fine-tuning.

3.4. Unsupervised video object tracking

Our framework requires an accurate unsupervised video object tracker, since its tracked object boxes will be used to generate the pseudo GT boxes on the weakly-labeled images. For this, we use the unsupervised tracking method of [40], which creates a diverse and representative set of spatial-temporal object proposals in an unannotated video. Each spatial-temporal proposal is a sequence of boxes fitting an object over multiple frames in time.[1]

Briefly, the method begins by leveraging appearance and motion objectness to score a set of static object proposals in each frame, and then groups high-scoring proposals across frames that are similar in appearance and frequently appear throughout the video. Each group is then ranked according to the average objectness score of its instances.

[1]We also tried the video segmentation method of [24]. However, it fails to produce good segmentations when an object is not moving. Ultimately, transferring its object boxes resulted in a slightly worse detector.

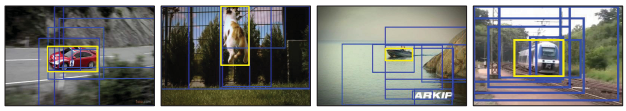

Figure 5. Examples showing the spatio-temporal boxes generated with [40] (blue), and our automatically selected box (yellow).

For each group, the method trains a discriminative tracking model with the group's instances as positives and all non-overlapping regions in their frames as negatives, and tracks the object in each instance's adjacent frames. The model is then retrained with the newly tracked instances as positives, and the process iterates until all frames are covered. The output is a set of ranked spatio-temporal tracks that fit a box around the objects in each frame that they appear. The method also has a pixel-segmentation refinement step, but we skip it for speed. See [40] for details.

For each video in S_V, we take the 9 highest-ranked tracks generated by [40]. Not all of these tracks will correspond to the object-of-interest. We therefore use our mined positive regions in $\mathcal{P}$ to try to select the relevant one in each frame. Specifically, given frame f, we match each positive region $\boldsymbol{r}_i$ to it in a sliding-window fashion in *conv5* feature space (as in Sec. 3.2), and record its best matching box $\boldsymbol{v}_i^f$ in the frame. We score a tracked box $\boldsymbol{t}_j^f$ in frame f as: $score(\boldsymbol{t}_j^f) = \sum_i \text{IOU}(\boldsymbol{v}_i^f, \boldsymbol{t}_j^f) \times sim(\boldsymbol{r}_i, \boldsymbol{v}_i^f)$, where i indexes the positive regions in $\mathcal{P}$, j is the index of a tracked video box, and sim is cosine similarity. We choose the tracked box with the highest score, and discard the rest. Our selection criterion favors choosing a box in each video frame that has high-overlap with many good matches from discriminative positive regions. See Fig. 5 for examples. The selected video boxes are provided as input to the video-matching module described in Sec. 3.2.

4. Experiments

We analyze: (1) localization accuracy of our discovered pseudo GT boxes on the weakly-labeled training images, (2) detection performance of our trained models on the test images, (3) ablation studies analyzing the different components of our approach, and (4) our selection criterion for choosing the relevant object track in each video frame.

Datasets. We use videos from YouTube-Objects [28] and images from PASCAL VOC 2007 and 2010. We evaluate on their **10 shared classes** (treating each as a positive in turn): aeroplane, bird, boat, car, cat, cow, dog, horse, motorbike, train. YouTube-Objects contains 9-24 videos per class; each video is 30-180 sec; 570K total frames. We only use each video's weak category-label (i.e., we do not know in which frames or regions the object appears). Each video is divided into shots with similar color [28]; we generate object tracks for each shot using [40]. VOC 2007 is used by all existing state-of-the-art weakly-supervised detection algorithms; VOC 2010 is used by [4]. For VOC 2007 and 2010, we use

VOC 2007 train+val	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	mean CorLoc
Initial pseudo GT (with all images)	48.8	33.9	13.3	57.3	46.5	32.2	44.4	40.8	48.2	43.7	40.9
Initial pseudo GT (excluding missed images)	58.8	49.6	17.7	64.7	60.4	44.8	52.8	55.3	54.3	53.0	51.1
Updated pseudo GT (with all images)	58.8	49.6	15.4	64.9	59.0	43.2	51.2	57.5	63.1	54.4	51.7

Table 1. Localization accuracy in terms of CorLoc on the VOC 2007 train+val set. We evaluate our initial and updated pseduo GT boxes. The final boxes (third row) provide very good localizations in the training data, which leads to accurate training of object detectors.

the train+val (5011 imgs) and train set (4998 imgs), respectively, to discover the pseudo GT boxes. For both datasets, we report detection results on the test set using average precision. In contrast to existing weakly-supervised methods (except [33, 34]), we do not discard instances labeled as *pose, difficult, truncated*, and restrict the supervision to the image-level object presence/absence labels to mimic a more realistic (difficult) weakly-supervised scenario.

Implementation details. For mining discriminative regions, we take k=(# positive images)/2 nearest neighbors, and top C=200 clusters. When matching a positive region to video, we adjust its box to have roughly 48 *conv5* cells using a sizing heuristic [22], and compute matches in every 8th frame for speed. For the mean-shift bandwidth b, we train separate detection models for b=$[100, 250, 500, 1000]$ and validate detection accuracy over our automatically selected object tracks on YouTube-Objects (i.e., we treat them as noisy GT); even though the discovered tracks can be noisy, we find they produce sufficiently good results for cross-validation. To compute deep features, we use AlexNet pre-trained on ILSVRC 2012 classification, using Caffe [17, 16]. We *do not* use the R-CNN network fine-tuned on PASCAL data [12].

To fine-tune our detector, we take our discovered pseudo GT boxes over all 10 categories to fine-tune the CNN (AlexNet pre-trained on ILSVRC2012 classification) by replacing its 1000-way classification layer with a randomly-initialized 11-way classification layer (10 categories plus background). We treat all selective search proposals with 0.6$\geq$IOU with a pseudo GT box as positives for that box's category, and all proposals with 0.1$\leq$IOU$\leq$0.3 with a pseudo GT box as negatives. All proposals from images not belonging to any of the 10 categories are also treated as negatives. We start SGD at a learning rate of 0.001 and decrease by $\times\frac{1}{10}$ after 20,000 iterations. In each SGD iteration, 32 positives (over all classes) and 96 negatives are uniformly sampled to construct a mini-batch. We perform 40,000 SGD iterations.

4.1. Pseudo ground-truth localization accuracy

We first analyze the localization accuracy of our discovered pseudo GT boxes on the VOC 2007 train+val dataset. We use the correct localization (CorLoc) measure [7], which is the fraction of positive training images in which the predicted object box has an intersection-over-union overlap (IOU) greater than 50% with any ground-truth box. As mentioned in [4], CorLoc is not consistently measured across previous studies, due to changes in the training sets (for example, we do not exclude the images annotated as *pose, difficult, truncated*). Thus, we only use it to analyze our own pseudo GT boxes, and use detection accuracy to compare against the state-of-the-art.

Table 1 shows the results. Our initial pseudo GT boxes produce an average CorLoc score of 40.9% across all categories (first row). However, we initially miss discovering a pseudo GT box in 12% of the images, which pulls down the average. (Recall we only keep the most confident box in each image that has at least $\theta = 20$ votes.) If we only consider the images in which a pseudo GT is initially found, then our average increases to 51.1% (second row). By detecting the missed pseudo GT boxes and updating the existing ones using the R-CNN model trained with the initial pseudo GT boxes (via an LSVM update), our final CorLoc average improves to 51.7% (third row). For the boat category, our low performance is due to boats often occurring with water; since water seldom appears in other categories, many water regions are mistakenly found to be discriminative, which leads to inaccurate localizations of the boat. (See supp. material for a further detailed breakdown of the error cases per class.) For the remaining categories, our pseudo GT boxes localize the objects well, and we will see in Sec. 4.3 that they lead to robust object detectors.

4.2. Pseudo ground-truth visualization

We next visualize our discovered pseudo GT on the VOC 2007 train+val set. In each image pair in Fig. 6, we display a heatmap of the transferred video object boxes and the final selected pseudo GT box. Our method accurately localizes the object-of-interest in many images, even in difficult cases where the object is in an atypical pose (1st dog), partially-occluded (2nd car), or in a highly-cluttered scene (2nd cat). The last column shows some failure cases. The most prominent failure case is when there are multiple instances of the same object category that are spatially close to each other. This is due to a sub-optimal mean-shift bandwidth parameter b, which is used in the voting of the pseudo GT box. Although we automatically select b via cross-validation on the video tracks (see implementation details), it is fixed *per-category*. Using an adaptive bandwidth [6] to automatically find an optimal value *per-image* may help to alleviate such errors. Importantly, these errors occur in only a few images. See the supp. material for results on all images.

Overall, the qualitative results demonstrate that by transferring object boxes from automatically tracked objects in video, we can accurately discover the objects' full spatial extent in the weakly-labeled image collection.

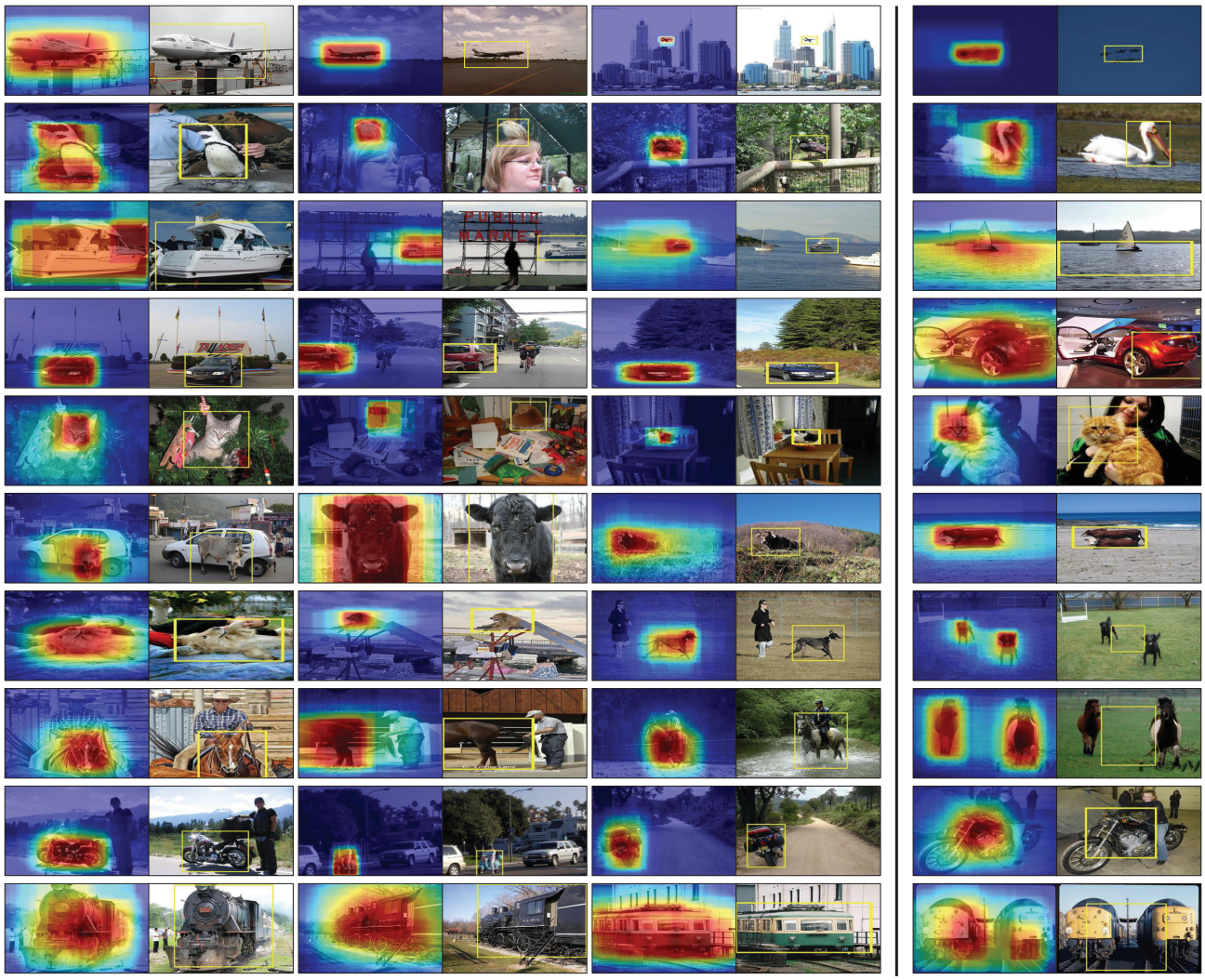

Figure 6. Qualitative results on the VOC 2007 train+val set. In each image pair, the first image shows a heatmap of the transferred video object boxes and the second image shows the final selected pseudo ground-truth box. Our approach accurately discovers the spatial extent of the object-of-interest in most of the images. The last column shows mis-localized examples. Our approach can fail when there are multiple instances of the same object category in the image (e.g., aeroplane, dog, horse, train) or when the object's appearance is very different from that found in videos (e.g., car). **Best viewed on pdf.**

4.3. Weakly-supervised detection accuracy

We next compute detection accuracy using the R-CNN model trained using our pseudo GT boxes. We compare with state-of-the-art weakly-supervised detection methods [33, 34, 1, 37, 4] that use the same AlexNet CNN features pre-trained on ILSVRC 2012. Note that our approach and the previous methods all use the same PASCAL VOC training images to train the detectors. Our use of videos is only to get better pseudo GT boxes on the training images.

Tables 2 and 3 show results on the VOC 2007 and 2010 test sets, respectively. Our approach produces the best results with a mAP of 41.9% and 40.1%, respectively. The baselines all share the same high-level idea of mining discriminative patterns that frequently/rarely appear in the positive/negative images. In particular, the detection results produced by [33] is similar to what we would get if we were to train a detector directly on our initially-mined discriminative positive regions. Since those regions often correspond to an object-part (e.g., car wheel) or include surrounding context (e.g., car with road) (recall Fig. 2), these methods have difficulty producing good localizations on the training data, which in turn degrades detection performance. While [34] tries to combine pairs of discriminative regions to provide better spatial coverage of the object, it is still limited by the mis-localization error of each individual region. We instead transfer automatically tracked object boxes from weakly-labeled videos to images, which produces more accurate localizations on the training data and leads to higher detection performance. Our low detection accuracy on cow

Table 2. Detection average precision on the VOC 2007 test set. We compare our approach to state-of-the-art weakly-supervised methods.

VOC 2007 test	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	mAP
Song et al., 2014 [33]	27.6	19.7	9.1	39.1	33.6	20.9	27.7	29.4	39.2	35.6	28.2
Song et al., 2014 [34]	36.3	23.3	12.3	46.6	25.4	23.5	23.5	27.9	40.9	37.7	29.7
Bilen et al., 2014 [1]	42.2	23.1	9.2	45.1	24.9	24.0	18.6	31.6	43.6	35.9	29.8
Wang et al., 2014 [37]	48.9	26.1	11.3	40.9	34.7	**34.7**	34.4	35.4	**52.7**	34.8	35.4
Cinbis et al., 2015 [4]	39.3	28.8	**20.4**	47.9	22.1	33.5	29.2	38.5	47.9	41.0	34.9
Ours w/o fine-tune	50.7	36.6	13.4	53.1	50.8	21.6	37.6	44.0	46.1	43.4	39.7
Ours	**53.9**	**37.7**	13.7	**56.6**	**51.3**	24.0	**38.5**	**47.9**	47.0	**48.4**	**41.9**

Table 3. Detection average precision on the VOC 2010 test set.

VOC 2010 test	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	mAP
Cinbis et al., 2015 [4]	44.6	25.5	**14.1**	36.3	23.2	26.1	29.2	36.0	**54.3**	31.2	32.1
Ours w/o fine-tune	50.9	35.8	8.1	40.5	45.9	26.0	36.4	39.0	45.7	39.4	36.8
Ours	**53.5**	**37.5**	8.0	**44.2**	**49.4**	**33.7**	**43.8**	**42.5**	47.6	**40.6**	**40.1**

VOC 2007 test	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	mAP
Inital pseudo GT	43.4	30.5	11.9	50.2	39.6	16.7	31.6	36.7	42.2	40.7	34.4
Updated pseudo GT	48.0	34.2	12.2	51.3	43	21.9	33.4	39.1	43.8	42.2	36.9
Updated pseudo GT + bbox-reg	50.7	36.6	13.4	53.1	50.8	21.6	37.6	44.0	46.1	43.4	39.7
Updated pseudo GT + fine-tune + bbox-reg	**53.9**	**37.7**	**13.7**	**56.6**	**51.3**	**24.0**	**38.5**	**47.9**	**47.0**	**48.4**	**41.9**

Table 4. Detection average precision on the VOC 2007 test set to evaluate the different components of our approach. See text for details.

can be explained by the poor video tracks produced by [40] (see supp. material), which confirms the need for good object tracks.

Overall, our results suggest a scalable application for object detection, since we can greatly reduce human annotation costs and still obtain reliable detection models.

4.4. Ablation studies

In this section, we conduct ablation studies to tease apart the contribution of each component of our algorithm. Table 4 shows the results. The first and second rows show mAP detection accuracy produced by the R-CNN models trained using the initial and updated (via LSVM update) pseudo GT boxes, respectively. The initial R-CNN model produces 34.4% mAP. Retraining the model with the updated pseudo GT boxes leads to 36.9% mAP, which shows that the extra positive instances and corrected instances are helpful. The third row shows bounding box regression results, which further boosts performance to 39.7% mAP. This confirms that our pseudo GT boxes are well-localized, since the trained bounding box regressor [9, 12] is able to adjust the initial detections to better localize the object.

The last row shows fine-tuning results. Training an R-CNN model with our fine-tuned features improves results on *all 10 categories* to 41.9% mAP for VOC 2007. The improvement is not as significant as in the fully-supervised case, which resulted in a $\sim$9% point increase for VOC 2007 (see Table 2 in [12]). Since our pseudo GT boxes are not perfect, any noise seems to have a more prominent effect than in the fully-supervised case, which has perfect GT boxes. Still, this result confirms our discovered pseudo GT boxes are quite accurate, making our approach amenable for fine-tuning.

4.5. Video track selection accuracy

Finally, we evaluate our selection criterion in choosing the relevant object box among the 9 tracks produced by the unsupervised video tracking algorithm [40]. For this, we compute the IOU between the tracked object boxes and the ground-truth boxes on the YouTube-Objects dataset [28]. Our automatically selected tracks produce a mean IOU of 45.1 over all 10 categories (see the supp. material for per-category results). While this is lower than the upper-bound mean IOU of 61.9 (i.e., the max IOU among the 9 proposals in each frame) they are sufficiently accurate to produce high-quality pseudo GT boxes. Furthermore, since we selectively retrieve a video object box only if it is overlapping with one of the top $n = 20$ matching video regions of a discriminative positive region, and then further aggregate those transferred boxes through hough voting, we can effectively filter out most of the noisy transferred tracks (as was shown in Fig. 6). Overall, we find that [40] produces sufficiently good boxes, and our selection criterion is in many cases able to choose the relevant one. These lead to accurate pseudo GT boxes on the weakly-labeled images.

Conclusions. We introduced a novel weakly-supervised object detection framework that tracks and transfers object boxes from weakly-labeled videos to images to simulate strong human supervision. We demonstrated state-of-the-art-results on PASCAL 2007 and 2010 datasets for the 10 categories of the YouTube-Objects dataset [28].

Our framework assumes that we have a way to track the object-of-interest in videos, so that we can delineate its box and transfer it to images. This is easier if the object is able to move on its own, but could also work for static objects, as long as the camera is moving. We plan to investigate this in the future. Finally, we intentionally trained our detectors using only the weakly-labeled images, in order to make our results comparable to previous weakly-supervised methods. It would be interesting to explore combining the video tracks with our pseudo GT image boxes for training the object detectors.

Acknowledgements. This work was supported in part by an Amazon Web Services Education Research Grant and GPUs donated by NVIDIA.

References

[1] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *BMVC*, 2014. 7, 8

[2] C. Chen and K. Grauman. Watching Unlabeled Video Helps Learn New Human Actions from Very Few Labeled Snapshots. In *CVPR*, 2013. 2

[3] R. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL Training for Weakly Supervised Object Localization. In *CVPR*, 2014. 1, 2

[4] R. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. In *arXiv:1503.00949*, 2015. 2, 5, 6, 7, 8

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, 2002. 4

[6] D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *ICCV*, 2001. 6

[7] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010. 2, 6

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What Makes Paris Look like Paris? *SIGGRAPH*, 31(4), 2012. 3

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *PAMI*, 32(9):1627–1645, 2010. 1, 5, 8

[10] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *CVPR*, 2003. 1, 2

[11] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 1, 2

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*, 2014. 1, 2, 4, 6, 8

[13] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In *CVPR*, 2015. 3

[14] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly Supervised Learning of Object Segmentations from Web-Scale Video. In *ECCV*, 2012. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 2

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014. 6

[17] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 2, 3, 6

[18] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. In *Neural Computation*, 1989. 2

[19] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008. 4

[20] C. Leistner, M. Godec, S. Schulter, A. Sakari, M. Werlberger, and H. Bischof. Improving Classifiers with Unlabeled Weakly-Related Videos. In *CVPR*, 2011. 2

[21] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Computational baby learning. In *arXiv:1411.2861*, 2015. 2

[22] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 6

[23] I. Misra, A. Shrivastava, and M. Hebert. Watch and learn: Semi-supervised learning of object detectors from videos. In *CVPR*, 2015. 2

[24] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *PAMI*, 36(6):1187–1200, 2014. 5

[25] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? Weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015. 2

[26] M. Pandey and S. Lazebnik. Scene Recognition and Weakly Supervised Object Localization with Deformable Part-Based Models. In *ICCV*, 2011. 1, 2

[27] The PASCAL Visual Object Classes Challenge Results. M. Everingham, L. Van Gool, I. Williams, J. Winn, and A. Zisserman. 2

[28] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning Object Class Detectors from Weakly Annotated Video. In *CVPR*, 2012. 2, 5, 8

[29] D. Ramanan, D. Forsyth, and K. Barnard. Building Models of Animals from Video. *PAMI*, 28(8):1319–1334, 2006. 2

[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014. 1, 2

[31] S. Singh, A. Gupta, and A. Efros. Unsupervised Discovery of Mid-level Discriminative Patches. In *ECCV*, 2012. 3

[32] P. Siva, C. Russell, and T. Xiang. In Defence of Negative Mining for Annotating Weakly Labelled Data. In *ECCV*, 2012. 1, 2

[33] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On Learning to Localize Objects with Minimal Supervision. In *ICML*, 2014. 1, 2, 3, 6, 7, 8

[34] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised Discovery of Visual Pattern Configurations. In *NIPS*, 2014. 2, 3, 6, 7, 8

[35] A. Torralba and A. A. Efros. Unbiased Look at Dataset Bias. In *CVPR*, 2011. 1

[36] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective Search for Object Recognition. *IJCV*, 104(2):154–171, 2013. 3, 4

[37] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014. 7, 8

[38] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 2

[39] M. Weber, M. Welling, and P. Perona. Unsupervised Learning of Models for Recognition. In *ECCV*, 2000. 1, 2

[40] F. Xiao and Y. J. Lee. Track and segment: An iterative unsupervised approach for video object proposals. In *CVPR*, 2016. 5, 8