Adapting Object Detectors from Images to Weakly Labeled Videos

Omit Chanda[1]
omitum@cs.umanitoba.ca
Eu Wern Teh[1]
umteht@cs.umanitoba.ca
Mrigank Rochan[1]
mrochan@cs.umanitoba.ca
Zhenyu Guo[2]
zhenyu.guo@sengled.com
Yang Wang[1]
ywang@cs.umanitoba.ca

[1] Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
[2] Sengled, Canada

Abstract

Due to the domain shift between images and videos, standard object detectors trained on images usually do not perform well on videos. At the same time, it is difficult to directly train object detectors from video data due to the lack of labeled video datasets. In this paper, we consider the problem of localizing objects in weakly labeled videos. A video is weakly labeled if we know the presence/absence of an object in a video (or each frame), but we do not know the exact spatial location. In addition to weakly labeled videos, we assume access to a set of fully labeled images. We incorporate domain adaptation in our framework and adapt the information from the labeled images (source domain) to the weakly labeled videos (target domain). Our experimental results on standard benchmark datasets demonstrate the effectiveness of our proposed approach. Our work can be used for collecting large-scale video datasets for object detection.

1 Introduction

We consider the problem of localizing objects in weakly labeled videos. A weakly labeled video only has the object label at the video level without the exact location of the object. For example, if a video is labeled as “car”, we assume that each frame of the video contains the object “car”. Our goal is to localize the car in each frame. In addition to the weakly labeled videos (which we call *target domain*), we assume that we also have access to a set of fully labeled images where the object bounding boxes are annotated (which we call *source domain*). However, there is a domain shift between the images in the source and target domains. Our goal is to use both sources of data to improve the object localization in the target video domain. See Fig. 1 for an illustration of our problem setup.

There have been significant advances in image understanding (e.g. object detection) in recent years. Much of the progress is enabled by the availability of large-scale annotated datasets, such as PASCAL [5], ImageNet [3], MS COCO [11]. Compared with the great success in image understanding, the progress in video understanding is relatively slow. One possible reason is the lack of annotated data in the video domain. In order to detect certain objects in videos, people usually take the off-the-shelf object detectors trained on the image dataset and apply the detector on each frame in the video. However, the state-of-the-art object detectors are often trained on datasets such as PASCAL or ImageNet, where images in those datasets are often obtained from online image websites (e.g. Flickr) and tend to have certain characteristics (e.g. object centric, high resolution). In contrast, the image characteristics of video frames tend to be different from the images in standard image datasets used in computer vision. For example, video frames tend to have lower resolutions and they might contain motion blurs. This is especially true in many real-world applications, e.g. surveillance videos. Object detectors trained on standard image datasets may not generalize well on those videos due to the domain shift between images and videos. Of course, one simple solution is to train the object detector directly in the video domain. However, this is not yet feasible due to the lack of large-scale annotated video datasets.

Our work is motivated by the following observation. Although it is difficult to collect labeled video data where object bounding boxes are annotated, it is relatively easy to collect weakly labeled videos, where the object label is provided at the video or frame level. If we can successfully localize objects from such weakly labeled videos, we will have an inexpensive way of collecting large-scale datasets in the video domain.

Our work is closely related to weakly supervised object localization. The novelty of our work is that we combine weakly supervised object localization and domain adaptation. Our proposed method can take advantage of both the weakly labeled videos in the target domain and the fully labeled images in the source domain. Instead of naively fusing the images from these two different domains, our model uses domain adaptation to account for the domain shift between the images from these two domains.

The main contribution of this work is that we incorporate domain adaptation in weakly supervised object localization. In this paper, we consider the fully labeled images as the source domain and the weakly labeled videos as the target domain. Our proposed method can exploit both sources of data and account for the domain shift between these two domains. Our method can potentially be used in many real-world settings. For example, if we have access to weakly labeled surveillance videos collected in a particular area, we can use our method as a way of collecting training data for learning object detectors specifically tailored to these surveillance videos.

2 Related Work

In this section, we review two lines of previous work that are closely related to this paper: weakly supervised object localization and domain adaptation in object recognition.

Weakly supervised object localization [1, 2, 10, 14, 15, 19, 20, 25, 26, 27] has been an area of active research in recent years. The goal is to develop methods to localize objects of interest in images without requiring detailed annotation on the training data. Many of these methods use some form of multiple instance learning. Bilen *et al.* [2] use latent SVM by treating bounding boxes as latent variables. Bilen *et al.* [1] propose an end-to-end architecture that combines object classification and detection in a single network. Teh *et al.* [26] introduce an attention-based network to select a subset of object proposal and use the features from the selected proposal to classify an image. Rochan *et al.* [15] use word vector

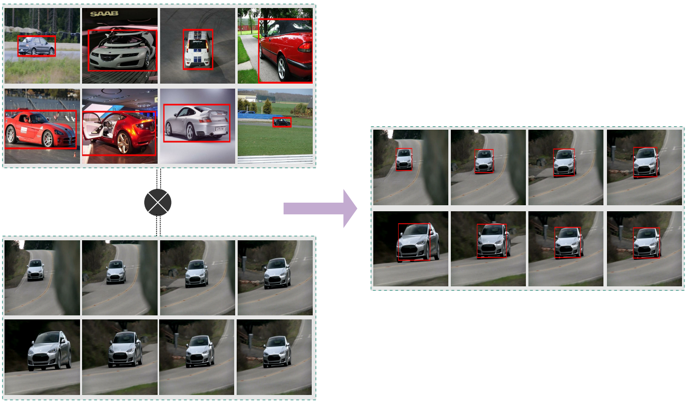

Figure 1: Our goal is to localize objects in weakly labeled videos by taking advantage of fully labeled images from a slightly different domain. For a particular object category (e.g. “car”), we have a collection of fully labeled images (top left) where the object bounding boxes are annotated. On the other hand, we have weakly labeled video data (bottom left) where we know about the presence of the object (e.g. “car”), but we do not know the exact location of the object in each frame of the video. Our goal is to use both sources of data to localize the object in the videos (right) while taking into account the domain shift between images and videos.

to determine the degree of influence of external object detector on its weakly supervised localization network. Singh *et al.* [18] use tracked objects in videos to simulate strong human supervision for weakly-supervised object detection.

Domain adaptation is another line of research closely related to our work. The goal of domain adaptation is to transform the feature representations and models learned in one domain (called "source domain") to work well in a different domain (called "target domain"). Sun *et al.* [22] propose an SVM based domain adaptation technique using a second order statistics called Correlation Alignment (CORAL) to minimize the discrepancy between source and target domains. Sharma *et al.* [17] introduce an incremental approach based on multiple instance learning (MIL) to minimize the discrepancy between different domains. Hoffman *et al.* [6] propose a domain adaptation technique for semantic segmentation problem where they have applied global and categorical alignment for transferring information from one domain to another. Su and Maji [21] propose cross quality distillation (CQD) technique to train recognition model. There is also work on domain adaptation between images and videos. Tang et al. [23] propose a domain adaptation technique for object detection from video data, where their system is trained on labeled image data and unlabeled video data. Kalogeiton et al. [9] analyze how the domain adaptation between images and videos affect the performance of object detection.

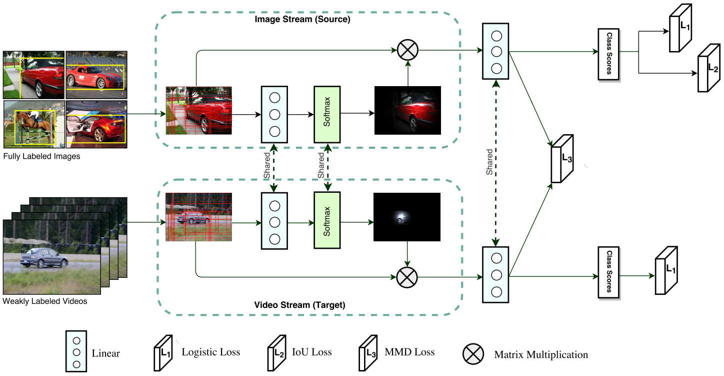

Figure 2: Overview of our proposed framework. Our model can be seen as a two-stream architecture that operates on images (source domain) and video frames (target domain). In the image stream, since we have the ground-truth object bounding box annotations, we define loss functions on the whole image classification and object bounding box localization. In the video stream, we define a loss function only on the whole image classification, since the videos are weakly labeled and we do not have ground-truth bounding box annotation of the object. Our model also has a loss function that accounts for the domain shift of images and video frames. It encourages the feature representations learned from these two domains to be similar.

3 Our Approach

Our approach is based on the attention network in [26]. The attention network is proposed for localizing objects in weakly supervised images, where the objects are only annotated at the whole image level. In our work, we extend the attention network to leverage the fully annotated data in the source domain (images) and the weakly supervised data in the target domain (videos). Our model also performs domain adaptation to account for the domain shift between images and videos.

Figure 2 shows the architecture of our proposed method. During training, our model can be seen as a two-stream architecture, where one stream operates on the fully labeled images in the source domain and the other stream operates on the weakly labeled video frames in the target domain. In the first stream, we define a loss function that encourages the model to produce correct image-level object label and object bounding boxes that are consistent with the annotations in the source domain. In the second stream, since we do not have ground-truth object bounding boxes, we define the loss function only on the whole image classification on the video frames. These two streams share their model parameters. We also incorporate a domain adaptation loss to account for the domain shift of these two different domains.

3.1 Object Proposals and Attentions

Given an image from either the source or target domain, the first step of our approach is to generate a shortlist of object proposals in this image. We use the edge boxes algorithm [28] for generating the object proposals. Let K be the number of object proposals generated on the image $\mathbf{x}$. We represent each proposal $\mathbf{x}_i$ $(i = 1, 2, ..., K)$ as a 4096-dimensional CNN feature vector [7].

We use a technique similar to [26] to generate an image level features from the features of object proposals in an image. For each object proposal $\mathbf{x}_i$, we compute an attention score a_i indicating the probability that this object proposal contains the object of interest. We achieve this by applying a linear mapping on $\mathbf{x}_i$ followed by a softmax operation. Let $\mathbf{w}_a$ denotes a vector of parameters for the linear mapping, the attention score a_i is calculated as follows:

$$g_i = \mathbf{w}_a^\top \mathbf{x}_i \tag{1a}$$

$$a_i = \frac{\exp(g_i)}{\sum_{j=1}^{K} \exp(g_j)}, \quad i = 1, 2, ..., K \tag{1b}$$

Next, we use the attention scores to combine the object proposals to get an image-level feature vector $\mathbf{z}$ as:

$$\mathbf{z} = \sum_{i=1}^{K} a_i \mathbf{x}_i \tag{2}$$

3.2 Image Stream

In the image stream, our network receives fully supervised image data, where the objects in an image are annotated with their bounding boxes. We can also easily obtain the whole image classification label from the object bounding box annotations. For example, if an image has at least one bounding box of "car", we consider "car" to be a positive class for the whole image classification.

In this stream, we assume that we have a set of N images $\{\mathbf{x}_s^{(n)}, y^{(n)}\}_{n=1}^{N}$, where $\mathbf{x}_s^{(n)}$ denotes the n-th image in the source domain and $y^{(n)} \in \{-1, 1\}$ indicates the corresponding image-level class label (i.e. presence/absence of the object of interest in the image). For an image $\mathbf{x}_s$, we use the image-level features $\mathbf{z}_s$ (Eq. 2) to generate a classification score. Similar to [26], we classify the whole image by a linear classifier with parameters $\mathbf{w}_{s,c}$:

$$f(\mathbf{x}_s; \{\mathbf{w}_a, \mathbf{w}_{s,c}\}) = \mathbf{w}_{s,c}^\top \mathbf{z}_s \tag{3}$$

where $f(\mathbf{x}_s; \{\mathbf{w}_a, \mathbf{w}_{s,c}\})$ is the score of classifying $\mathbf{z}_s$ to be a positive class.

We use the logistic loss for the whole image classification on the images in the source domain:

$$\ell_{class}(\{\mathbf{w}_a, \mathbf{w}_{s,c}\}) = \frac{1}{N} \sum_{n=1}^{N} \log\left(1 + \exp\left(-y^{(n)} f\left(\mathbf{x}_s^{(n)}; \{\mathbf{w}_a, \mathbf{w}_{s,c}\}\right)\right)\right) \tag{4}$$

Similar to [26], we use the attention score a_i of each object proposal to localize the object of interest. If the attention score a_i is high, we consider the object proposal more likely to contain the object. For the object localization prediction, we define a loss that measures how well the localized bounding box matches the ground-truth bounding box. This loss is defined as follows. Given a predicted bounding box B_p and a ground-truth bounding box B_{gt}, we compute the intersection-over-union (IoU) as $\frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$. If this IoU measurement is greater than 0.5, we consider this predicted bounding box to be correct and assign it a value of 1. Otherwise we assign a value of 0. We can then assign 0 or 1 to each of the object proposals in an image to indicate whether this proposal has enough overlap with the ground-truth bounding box. We use $v_i^{(n)}$ to denote this 0/1 value of the i-th object proposal in the n-th training image. We use $a_i^{(n)}$ to denote the attention score of this object proposal. Our localization loss is defined as:

$$\ell_{localize} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K}(v_i^{(n)} - a_i^{(n)})^2 \tag{5}$$

This loss function will encourage our model to produce attention scores that are consistent with $v_i^{(n)}$.

The final loss function in the image stream is the combination of Eq. 4 and Eq. 5, i.e.:

$$\mathcal{L}_1(\{\mathbf{w}_a, \mathbf{w}_{s,c}\}) = \ell_{class} + \ell_{localize} \tag{6}$$

3.3 Video Stream

In the video stream, our network receives video frames (target domain) that are weakly labeled, where only the classification label of a frame is available. Let us assume that there are M video frames $\{\mathbf{x}_t^{(m)}, y^{(m)}\}_{m=1}^{M}$. Given a video frame $\mathbf{x}_t$, we use the frame-level feature $\mathbf{z}_t$ (Eq. 2) to generate a classification score. We then classify the whole frame by a linear classifier with parameters $\mathbf{w}_{t,c}$:

$$f(\mathbf{x}_t; \{\mathbf{w}_a, \mathbf{w}_{t,c}\}) = \mathbf{w}_{t,c}^{\top}\mathbf{z}_t \tag{7}$$

where $f(\mathbf{x}_t; \{\mathbf{w}_a, \mathbf{w}_{t,c}\})$ is the score of classifying $\mathbf{z}_t$ to be a positive instance of the object class.

Similar to the image stream, we use the logistic loss to measure the classification performance on the video frames in the target domain.

$$\mathcal{L}_2(\{\mathbf{w}_a, \mathbf{w}_{t,c}\}) = \frac{1}{M}\sum_{m=1}^{M}\log\left(1 + \exp\left(-y^{(m)} f\left(\mathbf{x}^{(m)}; \{\mathbf{w}_a, \mathbf{w}_{t,c}\}\right)\right)\right) \tag{8}$$

Unlike the image stream, we do not define the loss in terms of object localization in the video stream, since we do not have the ground-truth object bounding boxes in the target domain.

3.4 Domain Adaptation

The images in the source and target domains may have very different characteristics. For example, frames in a video tend to be low resolution and have motion blur. In this section, we incorporate a domain adaptation loss in our model to account for this domain shift between source and target domains.

We use the multiple kernel maximum mean discrepancies (MK-MMD)[12] to define the domain adaptation loss. MK-MMD has been proved to be a very effective domain adaptation technique. Let image-level features $\mathbf{z}_s$ (Eq. 2) and frame-level features $\mathbf{z}_t$ (Eq. 2) be the feature vector from source domain and target domain, respectively. We use p and q to denote the distribution of the data from the source and target domains. The MK-MMD loss is defined as:

$$\mathcal{L}_3 = d_k^2(p,q) \triangleq \|E_p[\mathbf{z}_s] - E_q[\mathbf{z}_t]\|_{\mathcal{H}_k}^2 \tag{9}$$

In Eq. 9, $E_p[\mathbf{z}_s]$ and $E_p[\mathbf{z}_t]$ denote the mean feature vectors of $\mathbf{z}_s$ and $\mathbf{z}_t$ from source and target domains respectively. In practice, $E_p[\mathbf{z}_s]$ and $E_p[\mathbf{z}_t]$ can be computed as the average of training data in the source domain and target domain. $\mathcal{H}_k$ denotes producing a kernel Hilbert space associated with certain kernel k. In this work, we use the following kernel:

$$k(x_s, x_t) = \langle \mathbf{z}_s, \mathbf{z}_t \rangle \tag{10}$$

MK-MMD finds a feature representation by minimizing the discrepancy between different domain examples and makes the features more similar between domains. Here in this work we use Maximum Mean discrepancy loss $\mathcal{L}_3$ to minimize the discrepancy between $\mathbf{z}_s$ and $\mathbf{z}_t$.

Our final model is learned by optimizing the combination of the loss functions defined in Eqs. 6 8 9:

$$\mathcal{L}(\mathbf{x}_s, \mathbf{x}_t; \mathbf{w}_a, \mathbf{w}_{s,c}, \mathbf{w}_{t,c}) = \mathcal{L}_1(\mathbf{w}_a, \mathbf{w}_{s,c}) + \mathcal{L}_2(\mathbf{w}_a, \mathbf{w}_{t,c}) + \mathcal{L}_3 \tag{11}$$

3.5 Temporal Smoothing

If we are simultaneously localizing the object in all frames of a video, we can use temporal smoothing to further improve the result of localization. The intuition is that the appearance and spatial location of the bounding box found in each frame should be consistent across the video. We use an idea similar to [16] for the temporal smoothing.

We consider the frames in a video as an undirected chain graph where a node in the graph represents a frame with certain number of object proposals and an edge represents the temporal relationship between two adjacent frames. Suppose there are T frames $X_1, X_2, ..., X_T$ in a particular video. We represent the bounding box that contains object of interest in each of the T frames as $P_1, P_2, ..., P_T$, where $P_i \in \{1, 2, ..., K\}$ denotes one of the K object proposals that is selected to contain the object in the i-th frame. We solve the following optimization problem to enforce temporal consistency across frames of a video:

$$\max_{P_1, P_2, ..., P_T} \sum_i \varphi(P_i, X_i) + \sum_{i,i+1} \psi(P_i, P_{i+1}) \tag{12}$$

where $\varphi(P_i, X_i)$ is a unary potential indicating how likely a particular object proposal can contain the object of interest. We set $\varphi(P_i, X_i)$ as the attention score of the object proposal. The pairwise potential $\psi(P_i, P_{i+1})$ applies the constraint on the bounding boxes between two adjacent frames. Following [16], we define the temporal relationship $C_{temporal}(P_i, P_j)$ between two bounding boxes of adjacent frames P_i and P_j in a video as follows,

$$C_{temporal}(P_i, P_j) = \alpha \left(\|f_c(P_i) - f_c(P_j)\|_2^2 + \|f_a(P_i) - f_a(P_j)\|_2^2 \right) \tag{13}$$

Here $f_c(P_i)$ denotes the center of the bounding box and $f_a(P_i)$ denotes the area of the bounding box. The intuition behind this temporal formulation is that the object position and size do not change significantly between two adjacent frames. Using the temporal relationship $C_{temporal}(P_i, P_j)$, we define the pairwise potential $\psi(P_i, P_j)$ between two adjacent frames of a video as follows:

$$\psi(P_i, P_j) = exp\left(-\left(C_{temporal}(P_i, P_j) \right)^2 \right) \tag{14}$$

4 Experiments

In this section, we present the experimental results of our proposed method. We first describe the datasets and experiment setup (Sec. 4.1). Then we present the results on two different datasets (Sec. 4.2).

4.1 Experiment Setup

Dataset: To evaluate our model, we need a dataset of images with bounding boxes as the source domain, and a dataset of weakly labeled videos as the target domain. Given these datasets in two domains, our goal is to localize the object in the target domain.

We use the images from the PASCAL VOC2007 dataset [5] as the fully supervised source domain. This dataset contains 20 object classes. The bounding boxes of these objects are provided. We then consider two different video datasets as the target domain and present our experimental results. The object categories of these video datasets are a subset of the 20 object categories in PASCAL VOC2007. Each video in the target domain is weakly labeled. For a given object category (e.g. "car"), if a video is labeled as positive for this object category, we assume that this object appears in every frame of the video. Our goal is to localize this object in each frame. We build our model for each object category separately. For example, when we learn a model for "car", we consider the frames of car videos as positive instances, and frames of other videos as negative instances.

Performance metric: We measure how well our proposed model localizes the object in the videos in the target domain using the CorLoc measurement defined in [4]. For each frame in a video, we measure the intersection-over-union (IoU) between the localized object bounding box and the ground-truth object bounding box. If the IoU is greater than 0.5, we consider this frame to be correctly localized. The CorLoc is computed as the percentage of the video frames that are correctly localized.

Baselines: We compare our model with several baselines. (1) *Video only*: This baseline ignores the images in the source domain and directly localizes the objects in the target (video) domain using the weakly supervised learning. This baseline is equivalent to the attention network in [26]; (2) *Image only*: This baseline ignores the videos in the target domain during learning and only uses the fully labeled images in the source domain. This is equivalent to learning a standard object detector (fast RCNN) using the fully labeled images in the source domain, then directly applying the object detector for object localization on the videos in the target domain. Since no bounding box information of target domain is available in this setting, we cannot fine-tune the "image-only" fully supervised model on the video dataset; (3) *Video + image*: This baseline is equivalent to ours, except that it does not perform domain adaptation. In other words, it learns the model parameters by optimizing $\mathcal{L}_1 + \mathcal{L}_2$.

We use stochastic gradient descent to optimize the loss function with a momentum of 0.5 and learning rate of 0.001. We fix the mini-batch size in our experiments as 50. In each mini-batch we forward 50 images from the image domain and 50 frames from the video domain into the two-stream network. We define domain adaptation loss on each mini-batch in our training. We use a single NVIDIA Tesla K40 GPU in our experiments.

4.2 Results

We show results on two video datasets. We use each of the video datasets as the target domain and the PASCAL VOC2007 dataset as the source domain.

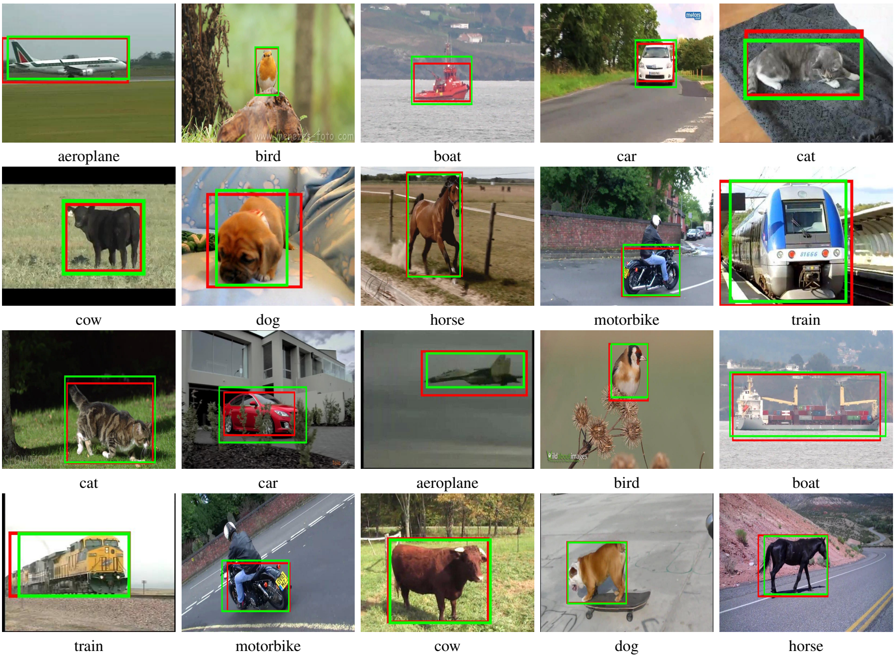

Figure 3: Qualitative examples of object localization of our network on the YouTube-Objects dataset. Ground-truth bounding boxes marked in Green whereas predicted bounding boxes marked in Red.

method	aero	bird	boat	car	cat	cow	dog	horse	bike	train	avg
Joulin et al.(video)[8]	25.12	31.18	27.78	38.46	**41.18**	28.38	33.91	35.62	23.08	25	30.97
Papazoglou et al.[13]	**65.4**	**67.30**	38.9	65.2	46.3	40.2	**65.3**	48.4	39	25	50.1
proposal only[15]	51.69	54.84	32.54	**85.71**	14.53	75.68	55.65	53.42	51.69	39.29	51.50
proposal + transfer[15]	56.04	30.11	39.68	**85.71**	24.79	**87.83**	55.65	60.27	61.8	51.79	55.37
video only [26]	55.07	62.37	43.65	84.62	28.21	66.22	58.26	53.42	62.92	39.29	55.40
image only	15.5	9.6	14.3	26.4	11.11	25.7	16.5	11	28.1	17.9	17.6
image + video	58.94	61.29	47.62	85.71	34.19	68.92	63.48	61.64	67.42	55.36	60.46
ours	60.39	62.37	**48.41**	**85.71**	34.19	71.62	**66.09**	**63.01**	**70.79**	**57.14**	**61.97**

Table 1: CorLoc results on the YouTube-Objects dataset.

YouTube-Objects Dataset: This dataset [14] is collected from videos of 10 different object classes. The ground-truth bounding box is provided for one frame per video. It consists of these frames with ground-truth bounding boxes (i.e. one frame per video).

Table 1 compares the CorLoc results of our method with other baselines. Our proposed method outperforms all the other approaches. Figure 3 shows qualitative examples of object localization on this dataset. Note that since this dataset only contains one frame from each video, we cannot apply the temporal smoothing (described in Sec. 3.5) on this dataset.

YouTube-Objects-Subset Dataset: This dataset is introduced by Tang et al. [24]. It contains a subset of the videos from the YouTube-Objects dataset, but it has more ground-truth annotations.

Table 2 shows the localization results on this dataset. Since we have ground-truth annotation on all the frames in a video on this dataset, we apply the temporal smoothing technique in Sec. 3.5. We find that our proposed model outperforms other alternatives. Moreover, temporal smoothing further improves the localization results of our method.

method	aero	bird	boat	car	cat	cow	dog	horse	bike	train	avg
proposal only[15]	42.23	51.24	29.54	67.76	14.75	50.20	47.02	22.18	16.44	18.84	36.02
proposal + transfer[15]	45.74	55.47	39.51	58.75	26.51	55.00	43.51	33.71	32.76	**25.63**	41.66
video only [26]	49.19	45.52	43.94	69.32	26.43	60.24	56.03	40.39	40.39	19.91	45.10
image only	17.64	30.02	13.39	22.01	12.58	19.64	22.64	15.71	5.93	8.53	16.81
image + video	50.46	54.39	45.09	71.23	30.67	61.11	62.63	44.54	42.20	22.66	48.5
ours	53.42	59.87	45.95	71.06	32.95	62.09	63.95	45.31	42.93	23.18	50.1
ours + temporal	**54.3**	**60.2**	**47.6**	**72.1**	**34.3**	**63.2**	**65.1**	**46.6**	**44.0**	23.3	**51.1**

Table 2: CorLoc results on the YouTube-Objects-Subset dataset.

5 Conclusion

In this work, we have proposed an approach for localizing objects in weakly labeled video. The novelty of our work is that in addition to weakly labeled videos, we also assume access to a set of labeled images. Instead of directly learning an object detector from the labeled images and applying it on the videos, our proposed approach takes into account the domain shift between images and videos. We use domain adaptation to transfer the knowledge from the labeled images (source domain) to the weakly labeled videos (target domain). Our experimental results show that our proposed method outperforms other alternative approaches. One interesting application of our proposed method is that it provides a possible way of building large-scale video dataset for object detection by using the existing labeled image datasets and the vast amount of weakly labeled videos available online.

Acknowledgments

This work was supported by NSERC Discovery and Engage grants. We thank NVIDIA for donating some of the GPUs used in this work.

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[2] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[4] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 2012.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010.

[6] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[7] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[8] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, 2014.

[9] Vicky Kalogeiton, Vittorio Ferrari, and Cordelia Schmid. Analysing domain shift factors between videos and images for object detection. *In IEEE transactions on pattern analysis and machine intelligence*, 2016.

[10] M. Pawan Kumar, Ben Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, 2010.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

[12] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015.

[13] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *IEEE International Conference on Computer Vision*, 2013.

[14] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] Mrigank Rochan and Yang Wang. Weakly supervised localization of novel objects using appearance transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[16] Mrigank Rochan, Shafin Rahman, Neil DB Bruce, and Yang Wang. Weakly supervised object localization and segmentation in videos. *Image and Vision Computing*, 2016.

[17] Pramod Sharma, Chang Huang, and Ram Nevatia. Unsupervised incremental learning for improved object detection in a video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[18] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *International Confernece on Machine Learning*, 2014.

[20] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. Weakly-supervised discovery of visual pattern configurations. In *Advances in Neural Information Processing Systems*, 2014.

[21] Jong-Chyi Su and Subhransu Maji. Cross quality distillation. *arXiv preprint arXiv:1604.00433*, 2016.

[22] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547*, 2015.

[23] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems*, 2012.

[24] Kevin Tang, Rahul Sukthankar, Jay Yagnik, and Li Fei-Fei. Discriminative segment annotation in weakly labeled video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[25] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[26] Eu Wern Teh, Mrigank Rochan, and Yang Wang. Attention networks for weakly supervised object localization. In *British Machine Vision Conference*, 2016.

[27] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, 2014.

[28] C. Lawrence Zitnick and Piotr Dollar. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 2014.