Frame-to-Frame Aggregation of Active Regions in Web Videos for Weakly Supervised Semantic Segmentation

Jungbeom Lee[1] Eunji Kim[1] Sungmin Lee[1] Jangho Lee[1] Sungroh Yoon[1,2,*]
[1] Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea
[2] ASRI, INMC, ISRC, and Institute of Engineering Research, Seoul National University
{jbeom.lee93, kce407, simonlee0810, ubuntu, sryoon}@snu.ac.kr

Abstract

When a deep neural network is trained on data with only image-level labeling, the regions activated in each image tend to identify only a small region of the target object. We propose a method of using videos automatically harvested from the web to identify a larger region of the target object by using temporal information, which is not present in the static image. The temporal variations in a video allow different regions of the target object to be activated. We obtain an activated region in each frame of a video, and then aggregate the regions from successive frames into a single image, using a warping technique based on optical flow. The resulting localization maps cover more of the target object, and can then be used as proxy ground-truth to train a segmentation network. This simple approach outperforms existing methods under the same level of supervision, and even approaches relying on extra annotations. Based on VGG-16 and ResNet 101 backbones, our method achieves the mIoU of 65.0 and 67.4, respectively, on PASCAL VOC 2012 test images, which represents a new state-of-the-art.

1. Introduction

Semantic segmentation is one of the most important tasks in computer vision and one that has made tremendous progress with fully annotated pixel-level labels [52, 3]. However, real applications of semantic image segmentation require a large variety of object classes and a great deal of labeled data for each class, and labeling pixel-level annotations is laborious. This problem can be addressed by weakly supervised methods that use more easily obtainable annotations such as scribbles, bounding boxes, or image-level tags.

Weakly supervised semantic segmentation methods have evolved rapidly. Scribble supervision [39] can now achieve 97% of the performance of manually supervised semantic segmentation with the same backbone network, and even the weakest image-level supervision [24] can produce 90% of that performance. However, the rate at which segmentation techniques relying on weak annotations are improving is declining rapidly. For instance, under image-level supervision, a 7.9% improvement on PASCAL VOC 2012 validation images [7] was achieved between 2016 and 2017 [23, 2], 1.8% between 2017 and 2018 [46], but only 0.8% since then [24].

Most methods of weakly supervised segmentation use localization maps activated by a classifier (e.g. a CAM [53]) as a proxy ground truth to train their segmentation network. However, because such localization maps are coarse and therefore activate only part of a target object, they provide a poor version of proxy ground truth. Some researchers have

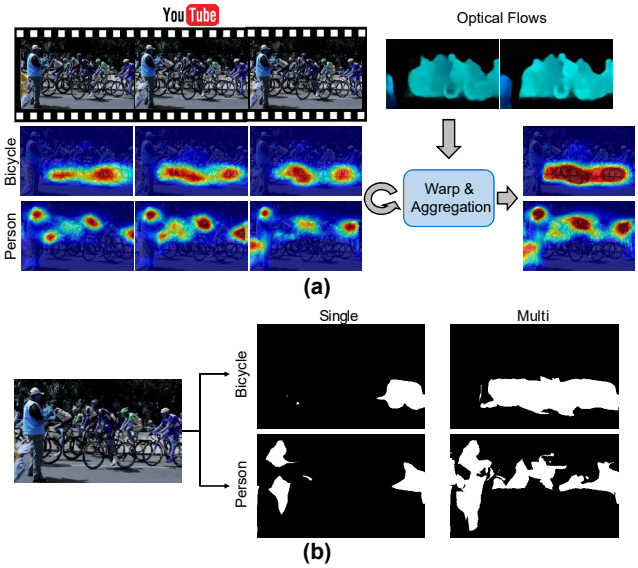

Figure 1: (a) Our method discovers activated regions from each frame and aggregates them into a single frame using a warping technique based on optical flow. (b) Comparison of generated proxy ground truth masks obtained from a single frame (1st row) and aggregated for multiple frames (2nd row). The latter covers larger region of the target objects because it embodies information from several frames.

*Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

tried to address this problem by erasing discriminative part of the target object [44, 15] or introducing different sizes of receptive fields [46, 24, 25], but these developments encounter limitations because of the finite information that can be obtained from the image-level annotations, resulting in a coarse proxy ground truth. Of course, stronger methods of supervision, such as instance-level saliency [16], scribbles [40, 41], or bounding boxes [4] can be used, but the extra supervision required makes it harder to expand object classes or move to a new domain.

One way of reducing the effort required for strong supervision is to use so-called webly supervised segmentation, which makes use of the vast amount of image and video data on the web, which can be used in combination with other existing weakly annotated data. However, the data obtained from the web may be of low quality or may not depict any objects corresponding to the search term. Attempts have been made to improve the quality of such data by filtering method using knowledge learned from images with accurate image-level annotations (e.g. the PASCAL VOC 2012 dataset [7]), or by using existing image segmentation techniques such as GrabCut [32]. Despite labeling issues, some of the sophisticated webly supervised methods have shown remarkable performance [35, 12, 14, 21]. However, additional usage of web images brings the same level of coarseness as existing weakly annotated data, so the performance improvements using web images are merely due to the effects of data augmentation. On the other hand, the temporal variations in video allow a classifier to activate different regions of the target object, so that video offers the possibility of obtaining better pixel-level annotations than static images.

We propose a method of collecting regions activated in different frames, using a warping technique based on optical flow. Optical flow provides pixel-level displacements between two successive frames, making it possible to deduce which parts of the frames correspond. Warping the activated areas in the first frame into the second frame determines which areas of the second frame should be activated. These warped areas can be aggregated with the parts of the second frame which are activated on their own merits. We repeat this step so that the areas activated in several frames become available in a single frame. Figure 1(a) shows how this works on an example, and Figure 1(b) demonstrates the resulting increase in the area of the activated regions, and how this corresponds more clearly to the ground truth of the original image. In addition, the availability of multiple frames in a video allows more effective filtering to refine inaccurate labels (see Section 3.2).

Existing methods of webly supervised segmentation depend on off-the-shelf segmentation techniques [12, 35] such as GrabCut [32], complicated optimization of energy function [12, 42], or heuristic constraints [35] to generate a proxy ground truth of the data obtained from the web. None of these are needed by our technique.

The main contributions of this paper can be summarized as follows:

- We propose simple data filtering and incremental warping techniques which allow web videos to be used as an additional data source in weakly supervised semantic image segmentation.
- We empirically demonstrate that our method of processing web video data improves the performance of several methods of weakly supervised segmentation.
- Our technique significantly outperforms other state-of-the-art methods on the Pascal VOC 2012 benchmark in both weakly supervised and webly supervised settings.

2. Related Work

Over the past four years, improvements of over 20% in the PASCAL VOC 2012 benchmark have been achieved by fully supervised semantic segmentation [27, 48]. However, because of the difficulty of obtaining pixel-level annotations of all the types of image and classes of object encountered in real applications, it is difficult to apply semantic segmentation widely. Weakly supervised semantic segmentation methods have been proposed to address this problem, and they have achieved promising performance (Section 2.1). Webly supervised semantic segmentation methods using images or videos obtained by web crawlers have been introduced as a way of closing the gap between the performance of weakly and fully supervised methods (Section 2.2).

2.1. Weakly supervised semantic segmentation

The goal of weakly supervised semantic segmentation is to train a image segmentation network with relatively imprecise delineations of the objects to be recognized. Weak supervision can take the form of scribbles [41, 40], bounding boxes [4], or image-level tags, which is the approach on which will now focus.

Most methods of image-level annotation are based on a class activation map (CAM) [53]. However, it is widely known that a CAM only identifies small discriminative parts of a target object [49, 17, 44]. Thus, the localization maps obtained by CAM are insufficiently complete to be used as a proxy ground truth to train a segmentation network. Several techniques have been proposed to expand these activated regions to the whole target object. Erasing methods [22, 44, 15] prevent a classifier from focusing solely on the discriminative parts of objects by removing those regions. Other methods construct CAMs that embody the multi-scale context of the target object, by means of different sizes of receptive fields. MDC [46] computes CAMs

from features which have different receptive fields, realized by several convolutional blocks dilated at different rates. Pyramid Grad-CAM [25] collects features from each of several densely connected layers and merges the resulting localization maps. FickleNet [24] selects features stochastically by using modified dropout technique: this not only prevents the classifier from concentrating solely on the discriminative part, but also uses a different receptive field for every inference.

Region growing methods try to expand regions of the target object, starting from the initial CAM as a seed. AffinityNet [1] and CIAN [8] consider pixel-level semantic affinities, which identify relationship of pixels, and grow regions based on those affinities. SEC [23] and DSRG [17] progressively refine initial localization maps during the training of their segmentation network by means of resulting segmentation maps during training time, which are refined using a conditional random field (CRF).

2.2. Webly supervised semantic segmentation

Along with the growth of weakly supervised semantic segmentation, some researchers have attempted to improve the performance of weakly supervised methods using additional images or videos obtained from the web. WebS-i2 [21] collects two types of web data: images showing objects of the target class against a white backgrounds and images containing common backgrounds without any objects of interest class. It then trains the segmentation network by means of an iterative refinement process on realistic images whose weak annotations are accurately labelled. Web-Crawl [12] selects videos from YouTube by examining thumbnails, and segments them by spatio-temporal graph-based optimization. Bootstrap-Web [35] finds images which are expected to be easy to segment using heuristic constraints such as the sizes of objects, and exchanges knowledge between two networks; one is trained by filtered easy-to-segment web images and the other is trained by realistic hard-to-segment images. All these methods improve segmentation performance to some extent, but they are largely dependent on complicated optimization methods, on heuristic constraints, or on off-the-shelf segmentation methods such as GrabCut [32]

3. Proposed Method

Our goal is to train a segmentation network with weakly annotated image data $\mathcal{I}$ and web video data $\mathcal{V}$. While the image-level labels of each image in $\mathcal{I}$ have been manually annotated, $\mathcal{V}$ is annotated with noisy video-level tags because it has been collected from the web by searching with the name of each object class as the search term. Our training procedure has the following steps: A deep neural network is trained on $\mathcal{I}$ to identify classes of object (Section 3.1). After it has been trained, the network processes the videos in $\mathcal{V}$, and the classification results can be used to filter out the irrelevant frames or update the labels of the videos $\mathcal{V}$ (Section 3.2). The proxy ground truths for selected sequences of frames are then generated using an incremental warping method (Section 3.3). Finally, the proxy ground truth is used to train a segmentation network (Section 3.4). The overall procedure is shown as Algorithm 1.

Algorithm 1: Overall Procedure

Input: Image dataset $\mathcal{I}$, web video dataset $\mathcal{V}$
1 Train a classifier using $\mathcal{I}$ — Sec. 3.1
2 $\hat{\mathcal{V}} \leftarrow$ Data filtering for $\mathcal{V}$ — Sec. 3.2
3 **Generating Proxy Ground Truth:** — Sec. 3.3
4 $\quad \mathcal{M}_{\mathcal{I}}, \mathcal{M}_{\hat{\mathcal{V}}} \leftarrow$ Masks from CAMs on $\mathcal{I}$ and $\hat{\mathcal{V}}$ — Sec. 3.3.1
5 $\quad \hat{\mathcal{M}}_{\hat{\mathcal{V}}} \leftarrow$ Union by incremental warping $\mathcal{M}_{\hat{\mathcal{V}}}$ — Sec. 3.3.2
6 **Training Segmentation Network:** — Sec. 3.4
7 $\quad L_{\mathcal{I}} \leftarrow$ Compute segmentation loss using $(\mathcal{I}, \mathcal{M}_{\mathcal{I}})$
8 $\quad L_{\hat{\mathcal{V}}} \leftarrow$ Compute segmentation loss using $(\hat{\mathcal{V}}, \hat{\mathcal{M}}_{\hat{\mathcal{V}}})$
9 $\quad$ Update segmentation network by $L_{\mathcal{I}} + L_{\hat{\mathcal{V}}}$

3.1. Learning with precise labels

We train a deep convolutional neural network using $\mathcal{I}$, which has precise image-level multi-class labels. In order to obtain class activation maps (CAM) (Section 3.3), we modify the VGG-16 network [36] to be fully convolutional, by removing all the fully connected layers and adding additional convolutional layers so that the number of channels of the final output feature corresponds to the number of classes of interest. We then apply global average pooling (GAP) and a sigmoid function to the feature output by the network, so as to obtain a score for each class. We use these results to update the parameters of the classifier by means of a sigmoid cross-entropy loss function, of a sort widely used for multi-label classification.

3.2. Data filtering for noisy dataset

Images or videos which are expected to depict a certain class of object can be obtained from the web by searching, using the name of that class as the search term. That name is then used to label the images or videos that are acquired. But not all the resulting images or videos will actually show an object of that class, and many of them will contain objects of classes other than that corresponding to the search term. For example, a video obtained by searching for "horse" may show a person riding a horse; but it will just be labeled "horse", which is the search term.

Most webly supervised segmentation methods use the knowledge obtained from precisely annotated image data to eliminate web images or videos which do not depicit any objects corresponding to the search term. For example, Bootstrap-Web [35] uses SEC [23], which is trained by precisely annotated data, to obtain pixel-level class masks from web images. It discards images which have too few or too

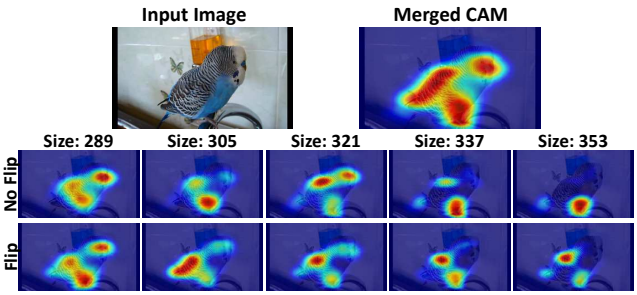

Figure 2: CAMs obtained by flipping and rescaling an input image. Rescaling changes the granularity at which the content of the image is considered. Different regions tend to be activated by flipped image.

many pixels assigned to an object of the search-term class in the corresponding mask. Web-Crawl [12] deals specifically with videos, and rejects a video if it has fewer than 5 frames with classification scores for its search term that achieve a threshold. These methods work on the assumption that there is only a single class of object in an image or video. But automatically collected images or videos can be expected to contain objects of many classes. In PASCAL VOC 2012 data, more than 36% of training images are annotated with more than one class.

We therefore introduce an incremental multi-class filtering and label refining method which considers several successive frames of a video. We can eliminate videos which show no objects of interest and correct inaccurate labels much more effectively than these processes can be performed on single images. When the same class of object is found in consecutive frames of a video, it becomes increasing likely that an object of that class is actually depicted. This means that labels can be assigned with more confidence to a video than to an image, and multi-class labelling becomes more feasible.

A classifier trained in Section 3.1 processes the videos in $\mathcal{V}$, and infers the object classes present in each frame, which are taken to be those with a score larger than a threshold τ. The set $\mathcal{C}$ of labels of these classes is then attached to each frame. We add a sequence of K frames to the filtered video set $\hat{\mathcal{V}}$ if those K frames all show objects from the same set of classes $\mathcal{C}$, and one of those classes corresponds to the search term. We eliminate frames that do not satisfy the above conditions.

3.3. Generating proxy ground truth

We now have a set of images $\mathcal{I}$, with accurate image-level labels, and a set of web videos $\hat{\mathcal{V}}$ with labels which are more accurate than those originally obtained from the search terms. We now describe the creation of a proxy ground truth for the frames in $\hat{\mathcal{V}}$, and its use to train a segmentation network.

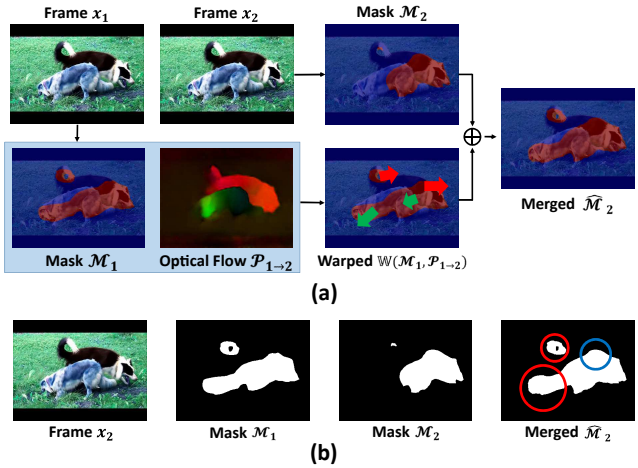

Figure 3: (a) Conceptual description of single-step aggregation. The mask of activated regions $\mathcal{M}_1$ is warped to $\mathbb{W}(\mathcal{M}_1, \mathcal{P}_{1\to2})$, and aggregated with the mask $\mathcal{M}_2$, to generate $\hat{\mathcal{M}}_2$. (b) Examples of generated masks. The red circles denote the regions masked only by $\mathcal{M}_1$, and the blue circle denotes the region masked only by $\mathcal{M}_2$. The mask $\hat{\mathcal{M}}_2$ covers more of the target objects than $\mathcal{M}_1$ and $\mathcal{M}_2$.

3.3.1 Inference localization maps

We use a CAM [53] to obtain a localization map for each class of object in the images. Zhang *et al.* [49] showed empirically, and also proved mathematically that the $c-$th channel of the last feature becomes the CAM for the $c-$th class, within a fully convolutional network of the type described in Section 3.1. We obtain CAMs from images which are scaled or flipped horizontally. After applying inverse transformations to the CAMs, we select the maximum value at each pixel over all the maps. Figure 2 shows some examples of the effect of flipping and scaling on CAMs. The CAM from a small image provides a coarse localization, and a CAM from a larger image identifies more detail of the object. Additionally, different regions are activated by flipped image.

3.3.2 Incremental warping of localization maps

It is widely known that a CAM only identifies the small discriminative part of a target object [49, 17, 44]. Our method obtains information about larger regions of a target object by merging information from successive frames of a video. The masks that indicate different regions of the target object are obtained from successive frames, and then aggregated into a single mask by considering optical flow. An analysis of optical flow between two successive frames provides a warp, which encodes the displacements which relate the pixels in one frame to those in the other. By warping the mask obtained from a frame to the following frame, we can transfer the activated region of the first frame to the second frame, allowing the union of the activated regions of both frames to be considered in a single image.

Let $\mathcal{X} \in \hat{\mathcal{V}}$ be a video containing K frames $\{x_i\}_{i=1}^{K}$. We compute masks $\{\mathcal{M}_i^c\}_{i=1}^{K}$ for each class c in $\mathcal{C}$ by thresholding the localization maps of the $\{x_i\}_{i=1}^{K}$ with the threshold θ_f. We obtain optical flows $\{\mathcal{P}_{i\to i+1}\}_{i=1}^{K-1}$ between each pair of successive frames.

We will now consider a single aggregation step, which combines the masks $\mathcal{M}_1^c$ and $\mathcal{M}_2^c$ obtained from successive frames x_1 and x_2, for a single class c. Let $\mathbb{W}(I, f)$ be the function which warps an image I following the flow field f, using bilinear interpolation. If $\mathcal{P}_{1\to 2}$ is the flow field between x_1 and x_2, then we can use this function to warp $\mathcal{M}_1^c$ to the space of x_2. The warped mask $\mathbb{W}(\mathcal{M}_1^c, \mathcal{P}_{1\to 2})$ expresses regions of x_2 which correspond to the activated regions of x_1. An aggregated mask $\hat{\mathcal{M}}_2^c$ can then be obtained as the union of the warped mask from x_1 and the mask $\mathcal{M}_2^c$ obtained from x_2.

$$\hat{\mathcal{M}}_2^c = \mathcal{M}_2^c \cup \mathbb{W}(\mathcal{M}_1^c, \mathcal{P}_{1\to 2}). \quad (1)$$

We repeat this procedure for the remaining annotated object classes in $\mathcal{C}$, and call the result $\hat{\mathcal{M}}_2$. This procedure is illustrated in Figure 3(a), and Figure 3(b) shows that a unioned mask map $\hat{\mathcal{M}}_2$ contains the activated regions from both frame x_1 and x_2. $\hat{\mathcal{M}}_2$ can then be warped using the optical flow $\mathcal{P}_{2\to 3}$ and aggregated with $\mathcal{M}_3$, producing $\hat{\mathcal{M}}_3$. By repeating this procedure until the $K-$th frame, we can obtain $\hat{\mathcal{M}}_K$, which contains the union of all the activated regions from all K frames. Aggregated masks from all the videos in $\hat{\mathcal{V}}$ can then be used as a proxy ground truth to train a segmentation network.

3.4. Segmentation Network

Many weakly supervised segmentation methods are built upon existing weakly supervised segmentation networks. For example, MDC [46] is based on a slightly modified version of the AE-PSL [44], and GAIN [26], TPL [22], and Boost-Web [35] are based on SEC [23]. We used FickleNet [24], but we also experimented with two other popular weakly supervised semantic segmentation networks: SEC [23] and DSRG [17]. All these segmentation networks are trained using the proxy ground truth of $\hat{\mathcal{V}}$ generated by the method described in Section 3.3. The background of $\hat{\mathcal{V}}$ is identified by saliency detection [13] and includes all pixels with saliency values lower than θ_b. When each segmentation network is trained, we use the proxy ground truth of $\mathcal{I}$ provided by the authors of each segmentation method.

For each iteration in training time, we create a batch; half of the elements in the batch come from $\mathcal{I}$, and the other half from $\hat{\mathcal{V}}$. We obtain the segmentation losses $L_{\mathcal{I}}$ and $L_{\hat{\mathcal{V}}}$ from the data from $\mathcal{I}$ and $\hat{\mathcal{V}}$ respectively. We then update the segmentation network by $L_{\mathcal{I}} + L_{\hat{\mathcal{V}}}$. Since $\mathcal{I}$ and $\hat{\mathcal{V}}$ may have different data distributions, we perform the domain adaptation according to [12]: the segmentation network is trained using both $\mathcal{I}$ and $\hat{\mathcal{V}}$ to predict segmentation masks for the images from $\mathcal{I}$, and those maps are used as a proxy ground truth to fine-tune the network.

4. Experiments

4.1. Experimental Setup

Image Dataset: We conducted experiments on the PASCAL VOC 2012 image segmentation benchmark [7], which contains 20 foreground object classes and one background class. Using the same protocol as other work on weakly supervised semantic segmentation, we trained our network using augmented 10,582 training images with image-level annotations. We determined mean intersection-over-union (mIoU) values for 1,449 validation images and 1,456 test images. The results for the test images were obtained from the official PASCAL VOC evaluation server.

Web Video Dataset: Starting with the Web-Crawl [12] dataset, we filtered out irrelevant frames and refine inaccurate labels with a threshold $\tau = 0.9$, and aggregated masks from $K = 5$ frames into a single mask, yielding 15,000 final samples for training a segmentation network. If several sets of frames can be selected from each video in Section 3.2, we chose only one to avoid similarity of samples. We obtained optical flows using PWC-Net [38]. The foreground and background thresholds, θ_f and θ_b, were set to 0.2 and 0.12 respectively.

Classification Network: Our classifier is based on the VGG-16 network [36], pre-trained using the Imagenet [5] dataset. The VGG-16 network was modified by removing all the fully connected layers and the last pooling layer, and we replaced the convolutional layers of the last block with convolutions dilated with a rate of 2. We added two convolutional layers with 1024 channels and a kernel size of 3 with 2D dropout [37].

Segmentation Network: As already stated, we experimented with FickleNet [24], SEC [23], and DSRG [17]. We followed the settings recommended by the authors of those methods, except for the optional domain adaptation process, during which the learning rate was reduced to 0.01 from the default learning rate.

Reproducibility: PyTorch [29] was used for training the classifier, extracting CAMs, and obtaining optical flow [38], and we used the Caffe deep learning framework [20] in the segmentation step.

4.2. Experimental Results

4.2.1 Results on Image Segmentation

Weakly supervised segmentation: Table 1 shows comparison of recently introduced weakly supervised semantic segmentation methods with various levels of supervision. These methods all use a segmentation model based on VGG-16 [36]. Our method achieves mIoU values of 63.9

Table 1: Comparison of the performance of weakly supervised segmentation methods using VGG16-based segmentation model on VOC 2012 validation and test sets.

Methods	*val*	*test*
Supervision: Image-level and additional annotations		
MIL-seg $_{\text{CVPR '15}}$ [30]	42.0	40.6
STC $_{\text{TPAMI '17}}$ [45]	49.8	51.2
TransferNet $_{\text{CVPR '16}}$ [11]	52.1	51.2
AISI $_{\text{ECCV '18}}$ [16]	61.3	62.1
Supervision: Image-level annotations only		
SEC $_{\text{ECCV '16}}$ [23]	50.7	51.1
CBTS-cues $_{\text{CVPR '17}}$ [33]	52.8	53.7
TPL $_{\text{ICCV '17}}$ [22]	53.1	53.8
AE_PSL $_{\text{CVPR '17}}$ [44]	55.0	55.7
DCSP $_{\text{BMVC '17}}$ [2]	58.6	59.2
MEFF $_{\text{CVPR '18}}$ [9]	-	55.6
GAIN $_{\text{CVPR '18}}$ [26]	55.3	56.8
MCOF $_{\text{CVPR '18}}$ [43]	56.2	57.6
AffinityNet $_{\text{CVPR '18}}$ [1]	58.4	60.5
DSRG $_{\text{CVPR '18}}$ [17]	59.0	60.4
MDC $_{\text{CVPR '18}}$ [46]	60.4	60.8
SeeNet $_{\text{NIPS '18}}$ [15]	61.1	60.7
FickleNet $_{\text{CVPR '19}}$ [24]	61.2	61.9
Ours	**63.9**	**65.0**

and 65.0 for PASCAL VOC 2012 validation and test images respectively, which is 94.4% of that of DeepLab [3], trained with fully annotated data, which achieved an mIoU of 67.6 on validation images. Our method is 3.1% better on test images than the best method which uses only image-level annotations for supervision. Our method also significantly outperformed several methods which have additional, as well as image-level annotations. These methods include TransferNet [11] which is trained on pixel-level annotations of 60 classes not included in the PASCAL VOC. AISI [16] has a salient instance detector, which is trained on well-annotated instance-level saliency maps, which are one of the most difficult forms of annotation to obtain. Our method also thoroughly outperformed existing methods based on the ResNet backbone [10] as shown in Table 2.

Figure 4 shows some examples of predicted segmentation masks produced by Bootstrap-Web [35], FickleNet [24], and our system, using both VGG-16 and ResNet-based segmentation models. In general, our proxy ground truth covers larger regions of the target object than those produced by other methods, so that the segmentation masks produced by our method tend to be more accurate.

Webly supervised segmentation: Table 3 shows mIoU values achieved on the PASCAL VOC 2012 dataset of webly supervised segmentation methods and the total number of training samples of each method. Our method showed the best performance despite being trained on a relatively small amount of data: 10.5k PASCAL VOC images and 15k frames of web video, whereas Web-Crawl [12] and Boot-Web [35] used 971k and 87.3k samples for training, respectively.

Table 2: Comparison of the performance of weakly supervised segmentation methods using the ResNet-based segmentation model on the VOC 2012 validation and test sets.

Methods	Backbone	*val*	*test*
Weakly supervised methods:			
MCOF [43]	ResNet 101	60.3	61.2
DCSP [2]	ResNet 101	60.8	61.9
DSRG [17]	ResNet 101	61.4	63.2
AffinityNet [1]	ResNet 38	61.7	63.7
SeeNet [15]	ResNet 101	63.1	62.8
CIAN [8]	ResNet 101	64.1	64.7
FickleNet [24]	ResNet 101	64.9	65.3
Webly supervised methods:			
Boot-Web [35]	ResNet 50	63.0	63.9
Ours	ResNet 101	**66.5**	**67.4**

Table 3: Comparison of webly supervised segmentation methods on VOC 2012 validation and test images. The 'Samples' column contains the total number of samples used for training, including the VOC images.

Methods	Samples	val	test
Methods using web images:			
WebS-i2 $_{\text{CVPR '17}}$ [21]	20.3K	53.4	55.3
Boot-Web $_{\text{CVPR '18}}$ [35]	87.3K	58.8	60.2
Methods using web videos:			
M-CNN $_{\text{ECCV '16}}$ [42]	13.6K	38.1	39.8
Web-Crawl $_{\text{CVPR '17}}$ [12]	971K	58.1	58.7
Ours	25.5K	**63.9**	**65.0**

4.2.2 Results on Video Segmentation

In Table 4, we assessed the segmentation results produced by our system on the YouTube-Object dataset [31], and compared with state-of-the-art video segmentation methods with various degrees of supervision. We used segmentation masks annotated by Jain *et al.* [18] for ground-truth of evaluation. We also report the mIoU of DSRG [17] and FickleNet [24] trained on $\mathcal{I}$ alone as baselines. Our method showed better performance than the existing methods and even surpassed the methods which use stronger supervision, such as bounding boxes. A few examples of predicted segmentation masks for the YouTube-Object dataset are shown in Figure 6.

4.3. Ablation Study

Number of frames aggregated: Figure 5(a) shows mIoU scores for the PASCAL VOC 2012 validation images with

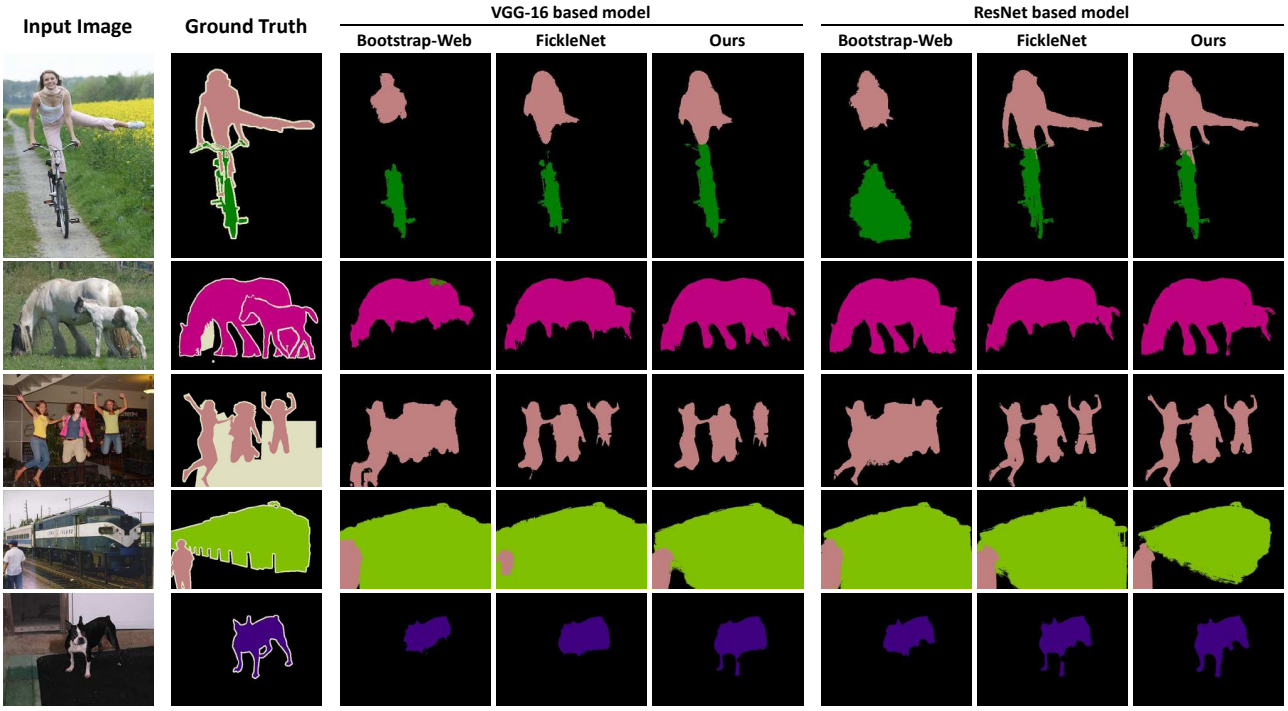

Figure 4: Examples of predicted segmentation masks for PASCAL VOC 2012 validation images.

Table 4: Comparison of video object segmentation methods with various supervision on the YouTube-Object dataset.

Method	sup.	mIoU
Tang *et al.* CVPR '13 [39]	$\mathcal{U}$	23.9
Papazoglou *et al.* ICCV '13 [28]	$\mathcal{U}$	46.8
Jang *et al.* CVPR '16 [19]	$\mathcal{U}$	53.0
Zhang *et al.* CVPR '15 [50]	$\mathcal{B}$	54.1
Drayer *et al.* ArXiv '16 [6]	$\mathcal{B}$	56.2
Zhang *et al.* TPAMI '18 [51]	$\mathcal{B}$	61.7
Saleh *et al.* ICCV '17 [34]	$\mathcal{I}$	53.3
Web-Crawl CVPR '17 [12]	$\mathcal{I}$	58.6
SROWN TIP '18 [47]	$\mathcal{I}$	61.9
Ours	$\mathcal{I}$	**62.1**

$\mathcal{U}$−Unsupervised, $\mathcal{B}$−Bounding boxes, $\mathcal{I}$−Image labels

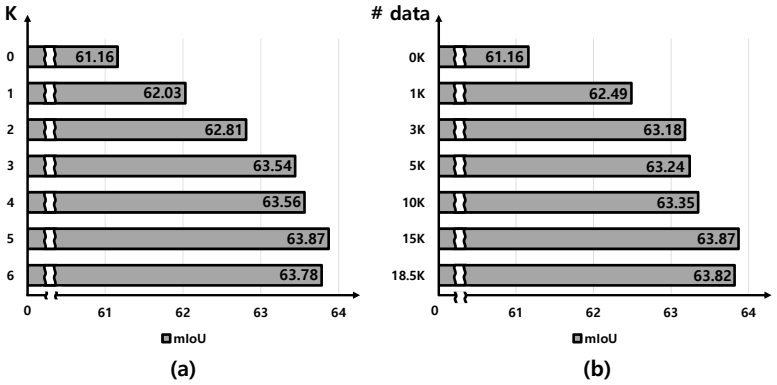

Figure 5: (a) Comparison of mIoU scores using different K. $K = 0$ denotes the result trained without web videos. (b) Comparison of mIoU scores using different numbers of web data.

different number of aggregated frames K. Using a single frame ($K = 1$) results in only slight performance improvement on the score without any web videos ($K = 0$). Increasing the number of successive frames across which maps are aggregated improves performance, which we expect, as larger regions of the target are represented by the final mask. But we find that aggregation above $K = 5$ is not beneficial. This can be attributed to the approximations of the optical flow involved in frame-to-frame warping. Errors can be expected to build up as regions are warped across more frames, undermining the accuracy of the proxy ground truth (see Appendix). The examples of aggregated mask at each incremental warping step are shown in Figure 7.

Number of web samples: The effect of the number of web samples is shown in Figure 5(b). Without any web videos, FickleNet [24] is trained with the PASCAL VOC data alone, and the mIoU is 61.2. The mIoU value increases monotonically up to 15,000 samples. More samples produce little change in the performance.

Other weakly supervised segmentation networks: In addition to FickleNet [24], we experimented with SEC [23] and DSRG [17] with our method. Table 5 shows the performance of those three segmentation networks with image data $\mathcal{I}$ alone, with an additional video dataset $\mathcal{V}$, and also with domain adaptation $\mathcal{DA}$. SEC [23] does not have a retraining process, so we add a retraining step, before domain

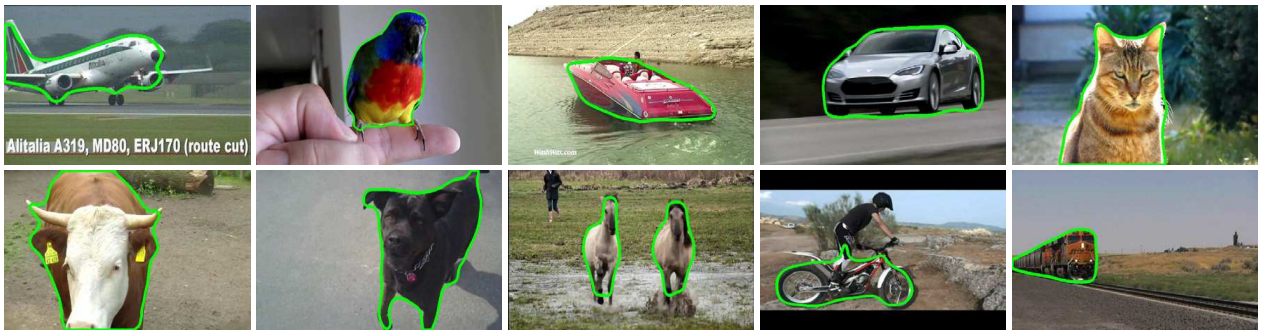

Figure 6: Predicted masks for frames of the YouTube-Object dataset. Segmented regions are outlined by green curves.

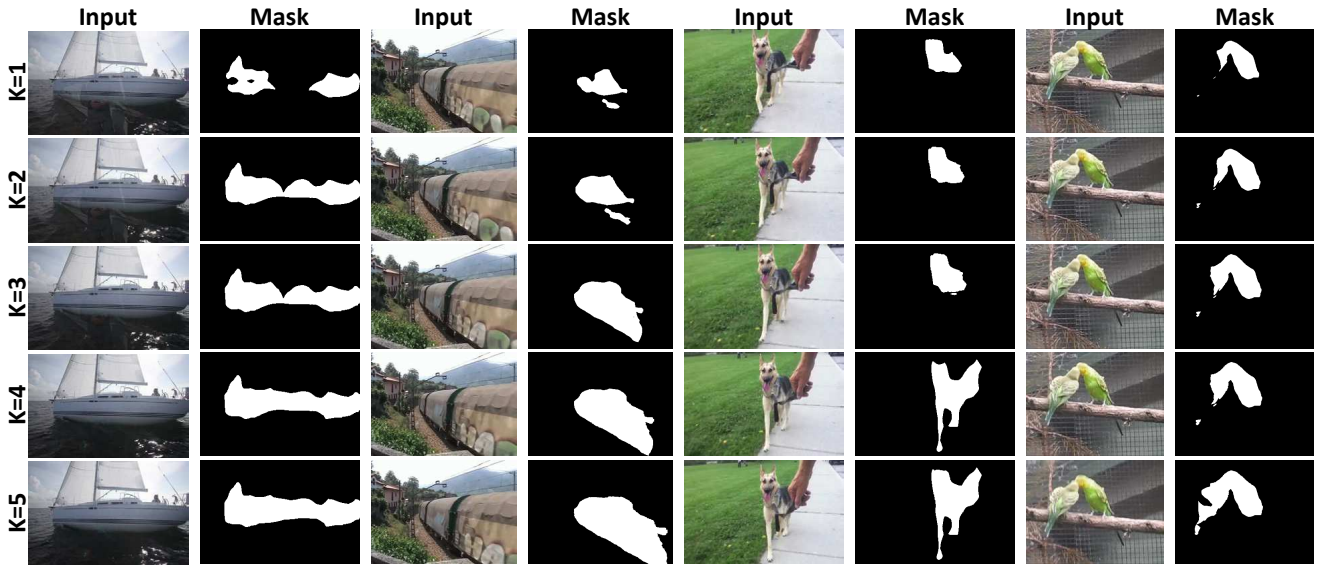

Figure 7: Examples of aggregated mask at each incremental warping step.

Table 5: Effects of addingvideo data $\mathcal{V}$ and domain adaptation $\mathcal{DA}$ on three weakly supervised segmentation models.

		SEC [23]	DSRG [51]	FickleNet [24]
mIoU	$\mathcal{I}$	50.7	59.0	61.2
	$\mathcal{I}$+$\mathcal{V}$	59.5[1]	62.1	63.2
	$\mathcal{I}$+$\mathcal{V}$+$\mathcal{DA}$	61.1	62.9	63.9

[1] A retrain process is included

adaptation, along the line of that in DSRG [17]. The results in Table 5 show that our method works effectively with the three weakly supervised semantic segmentation networks.

The results for SEC [23] offer the possibility of a fairer comarison with Bootstrap-Web [35], which is based on SEC. For the PASCAL VOC validation images, our method achieved mIoU values of 61.1, while Bootstrap-Web [35] achieved 58.8.

5. Conclusions

We have proposed a method to use videos automatically obtained from the web as additional data in weakly supervised semantic segmentation. We obtain activated regions from each frame of a video and aggregate them on a single image, so that our proxy ground truth covers large regions of the target object. This method does not require additional supervision, it can be realized without complicated optimization processes or off-the-shelf segmentation methods, and it requires relative few samples, because a lot of information extracted from many frames can be aggregated into a single frame. We have demonstrated that our method produces better results than those from other state-of-the-art weakly and webly supervised approaches. We have also demonstrated that our method works effectively with several weakly supervised semantic segmentation networks.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [2018R1A2B3001628], AIR Lab (AI Research Lab) in Hyundai Motor Company through HMC-SNU AI Consortium Fund, Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01367), Samsung Electronics (DS and Foundry), and the Brain Korea 21 Plus Project in 2019.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[2] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *British Machine Vision Conference*, 2017.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[4] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[6] Benjamin Drayer and Thomas Brox. Object detection, tracking, and motion segmentation for object-level video segmentation. *arXiv preprint arXiv:1608.03066*, 2016.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[8] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *arXiv preprint arXiv:1811.10842*, 2018.

[9] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1277–1286, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3204–3212, 2016.

[12] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7322–7330, 2017.

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

[14] Qibin Hou, Ming-Ming Cheng, Jiangjiang Liu, and Philip HS Torr. Webseg: Learning semantic segmentation from web searches. *arXiv preprint arXiv:1803.09859*, 2018.

[15] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 547–557, 2018.

[16] Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 2018.

[17] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018.

[18] Suyog Dutt Jain and Kristen Grauman. Supervoxel-consistent foreground propagation in video. In *European Conference on Computer Vision*, pages 656–671. Springer, 2014.

[19] Won-Dong Jang, Chulwoo Lee, and Chang-Su Kim. Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 696–704, 2016.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[21] Bin Jin, Maria V Ortiz Segovia, and Sabine Susstrunk. Webly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3635, 2017.

[22] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[23] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016.

[24] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019.

[25] Sungmin Lee, Jangho Lee, Jungbeom Lee, Chul-Kee Park, and Sungroh Yoon. Robust tumor localization with pyramid grad-cam. *arXiv preprint arXiv:1805.11393*, 2018.

[26] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[28] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784, 2013.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[30] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015.

[31] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3282–3289. IEEE, 2012.

[32] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[33] Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3529–3538, 2017.

[34] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation. In *2017 IEEE international conference on computer vision (ICCV)*, pages 2125–2135. IEEE, 2017.

[35] Tong Shen, Guosheng Lin, Chunhua Shen, and Ian Reid. Bootstrapping the performance of webly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1371, 2018.

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

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[38] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[39] Kevin Tang, Rahul Sukthankar, Jay Yagnik, and Li Fei-Fei. Discriminative segment annotation in weakly labeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2483–2490, 2013.

[40] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[41] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *European Conference on Computer Vision*, 2018.

[42] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Weakly-supervised semantic segmentation using motion cues. In *European Conference on Computer Vision*, pages 388–404. Springer, 2016.

[43] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2018.

[44] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017.

[45] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2314–2320, 2017.

[46] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018.

[47] Le Yang, Junwei Han, Dingwen Zhang, Nian Liu, and Dong Zhang. Segmentation in weakly labeled videos via a semantic ranking and optical warping network. *IEEE Transactions on Image Processing*, 27(8):4025–4037, 2018.

[48] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1857–1866, 2018.

[49] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[50] Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, and Changqun Xia. Semantic object segmentation via detection in weakly labeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3641–3649, 2015.

[51] Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, Changqun Xia, and Jun Li. Semantic object segmentation in tagged videos

via detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1741–1754, 2018.

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.