

Supplementary Materials: Simpler Does It: Generating Semantic Labels with Objectness Guidance

Md Amirul Islam^{1,4}
cs.ryerson.ca/~amirul

Matthew Kowal^{2,4}
mkowal2.github.io

Sen Jia⁵
github.com/SenJia

Konstantinos G. Derpanis^{2,4,6}
www.eecs.yorku.ca/~kosta

Neil D. B. Bruce^{3,4}
socs.uoguelph.ca/~brucen

¹ Ryerson University, Canada

² York University, Canada

³ University of Guelph, Canada

⁴ Vector Institute for AI, Canada

⁵ Toronto AI Lab, LG

⁶ Samsung AI Centre Toronto

S1 Examples of Generated Pseudo-labels on VOC 2012

Figure S1 shows additional examples of the generated pseudo-labels by combining the class agnostic objectness priors with either CAM [18] or bounding box proposals. Our pseudo-label generation technique successfully extracts boundary information from the objectness prior and class information from the CAM or bounding box proposals, resulting in high-quality pseudo-labels with fine-grained details about the object’s shape. For the ignore strategy, we assign values of 255 to the outer regions of a bounding box if it overlaps (above a certain threshold) with the largest bounding box in an image. These overlapped semantic regions have a high degree of uncertainty due to the inherent structure of bounding boxes and ignoring these regions during training results in better predictions (see Fig. S4).

S2 Details of SOnet Architecture

We discussed the SOnet architecture in Sec. 2.2 of the main manuscript. The details of the objectness module in SOnet architecture are shown in Table S1. The input to the objectness module is the segmentation map, $S \in \mathbb{R}^{b \times 21 \times 16 \times 16}$ which is generated by the DeepLabv3-ResNet101 network. The objectness module consists of five convolution layers where first four layers gradually increase the depth (i.e., channel) of the feature map. The last convolution layer predicts the desired objectness map, $O \in \mathbb{R}^{b \times 2 \times 16 \times 16}$. Note that we apply batch normalization and ReLU layers after each convolution layers except the last one which pre-

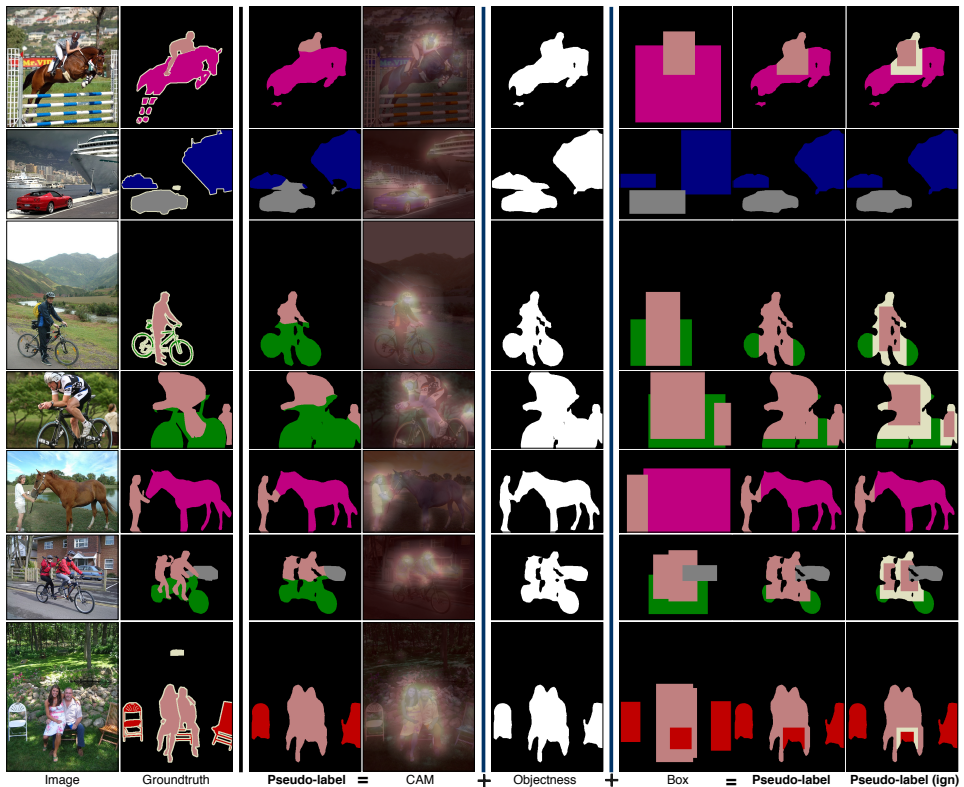


Figure S1: Additional examples of our pseudo-label generation process. We combine a class agnostic objectness prior with either a Class Activation Maps (CAMs) [18] or bounding box proposal to generate a pseudo-label.

dicts the objectness map. The newly introduced convolution layers are trained from scratch.

S3 Supplementary Experiments

In this section, we first provide implementation details of our proposed SONet (Sec. S3.1) and a description of the OpenV5 dataset (Sec. S3.2). Then, we provide anonymous links to the PASCAL VOC 2012 test set results and additional qualitative examples predicted by SONet with different levels of supervision (Sec. S3.3). Further we conduct experiments on video object segmentation (Sec. S3.4). We also show the generality of our proposed pseudo-label generation technique on the Berkeley DeepDrive dataset [24] (Sec. S3.5). Finally, we report a series of ablation studies (Sec. S3.6).

S3.1 Implementation Details

We implement our method using the PyTorch [9] framework trained end-to-end on two NVIDIA GeForce GTX 1080 Ti GPUs. We use the SGD optimizer to train our network.

Input: Segmentation Map $\mathcal{S} \in \mathbb{R}^{b \times 21 \times 16 \times 16}$
Conv2d (3×3), Batch Norm, ReLU $\rightarrow \mathbb{R}^{b \times 32 \times 16 \times 16}$
Conv2d (3×3), Batch Norm, ReLU $\rightarrow \mathbb{R}^{b \times 64 \times 16 \times 16}$
Conv2d (3×3), Batch Norm, ReLU $\rightarrow \mathbb{R}^{b \times 128 \times 16 \times 16}$
Conv2d (3×3), Batch Norm, ReLU $\rightarrow \mathbb{R}^{b \times 256 \times 16 \times 16}$
Conv2d (1×1), Batch Norm, ReLU $\rightarrow \mathbb{R}^{b \times 2 \times 16 \times 16}$
Output: Objectness Map $\mathcal{O} \in \mathbb{R}^{b \times 2 \times 16 \times 16}$

Table S1: Configuration of the Objectness Module in SONet.

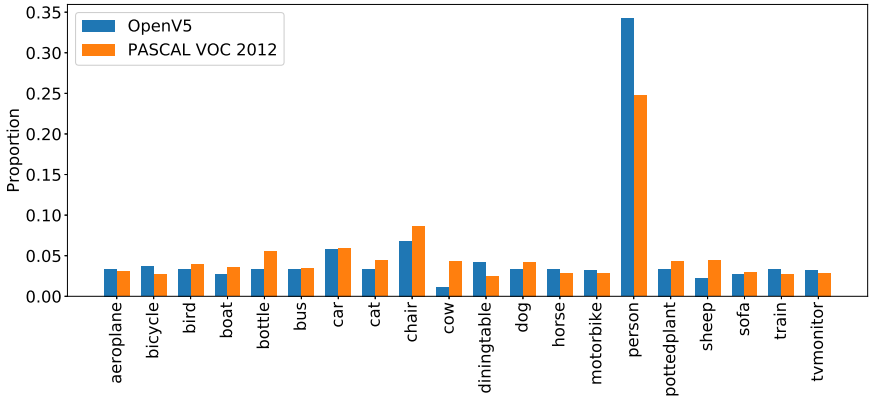


Figure S2: The distribution of semantic object categories in the OpenV5 subset and PASCAL VOC 2012 training sets.

We train all the variants of our SONet for 40 epochs with an initial learning rate of $2e-3$. We use a random crop of 513×513 and 321×321 during training for SONet and SONet*, respectively. Similarly, we use an output stride of 16 and 8 during training for SONet and SONet*, respectively. During inference, we use a crop of 513×513 and rescale to the original size using simple bilinear interpolation before calculating the mIoU. Following the current practice [4, 2, 1], to report test set results on PASCAL VOC 2012, we first train on the augmented training set followed by fine-tuning on the original trainval set with the generated pseudo-labels.

S3.2 OpenV5 Dataset

We have shown experimental results using the OpenV5 dataset for the task of semantic segmentation in Table 1(c) of the main paper. We compare against the state-of-the-art by using the standard protocol (training on PASCAL VOC 2012 augmented train set and evaluate on PASCAL VOC 2012 val/test set). As mentioned in Sec. 3.1 of the main manuscript, we use a subset of the OpenV5 dataset, where each semantic category is contained in a large number of images, consisting of 42,621 total images and 20 semantic categories. Figure S2 shows the comparison of the object instance distribution of the OpenV5 subset and PASCAL VOC

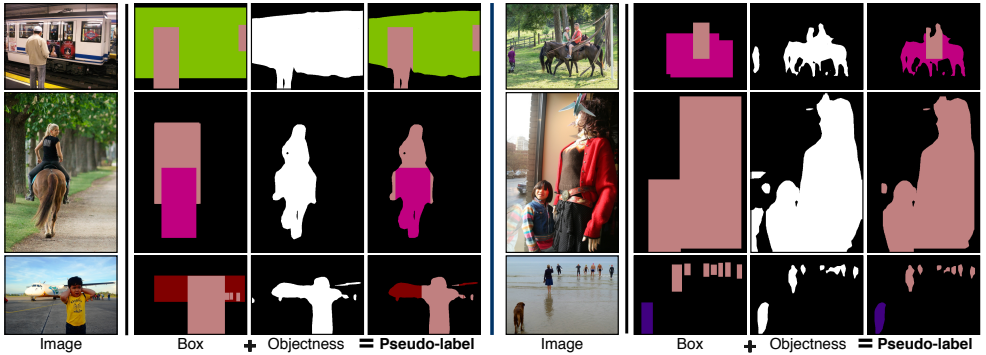


Figure S3: OpenV5 pseudo-labels. Examples of generated pseudo-labels with our proposed approach on the OpenV5 images.

2012 dataset. It is evident from the table that there are a considerable number of instances for each semantic category and *person* is by far the most dominant category as expected since it co-occurs with most of other categories. Figure S3 shows examples of the generated pseudo-labels for OpenV5, by combining the class agnostic objectness priors with bounding box proposals.

S3.3 PASCAL VOC 2012 Test Set Results

We illustrate additional visual examples predicted by SONet with different levels (CAMs and box-driven) of supervision on PASCAL VOC 2012 validation images in Fig. S4. The segmentation mask generated by SONet produces more accurate results when trained with CAMs or box-driven pseudo-labels than SONet trained solely with CAMs or bounding box annotations.

S3.4 Video Object Segmentation Results

We also experiment on the YouTube-Object (YTO) dataset [10] to show the effectiveness of our method in segmenting objects from videos by simply evaluating the results produced by SONet. Following prior works [6, 8, 12], we use the groundtruth segmentation masks provided by [6] to evaluate the performance of SONet and also compare our method with recent video segmentation methods with weak supervision in Table S2. Note that all the baseline methods are explicitly trained on video datasets and use temporal cues, while our method is trained on static images without temporal information. Our SONet method outperforms the existing methods which use different levels of supervision. This may be because objectness-driven pseudo-labels provide more fine-grained localization with sharper object boundaries than coarse bounding boxes. Samples of the predicted masks for the YTO dataset are shown in Fig. S5.

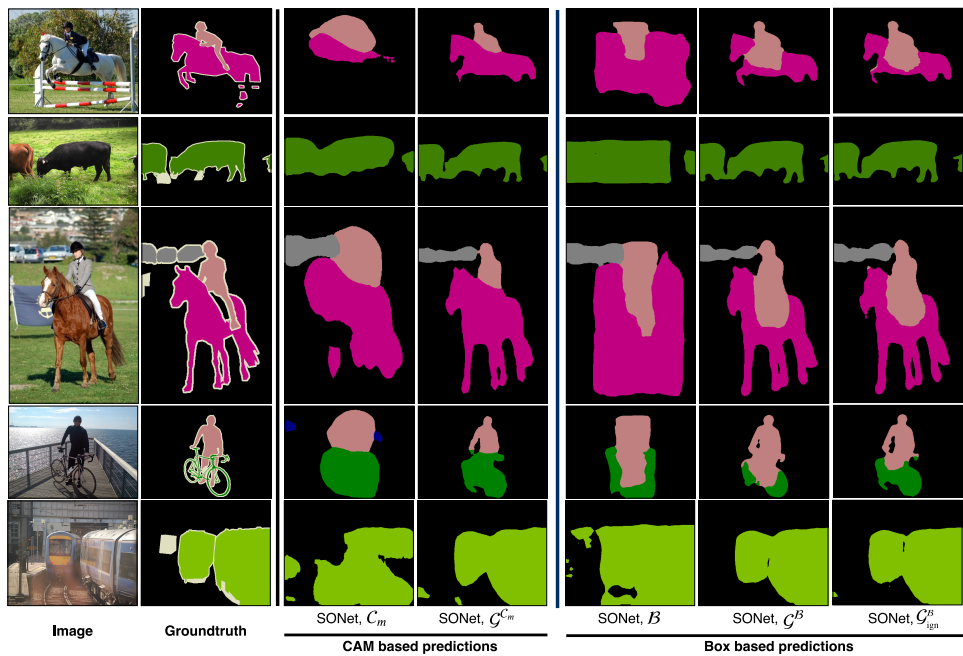


Figure S4: **Examples of predicted segmentation maps** with our proposed approach on PASCAL VOC 2012 validation images.

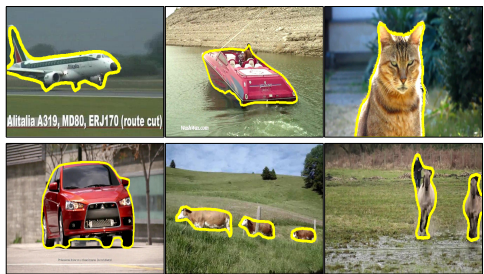


Figure S5: Predicted masks for frames of the YouTube-Object dataset. The outline of segmented regions are shown as yellow curves. Note that the object boundaries of SONet’s predictions are detailed and smooth.

	SOSD [14]	OVS [14]	DPM [14]	BBF [14]	Crawl [14]	SROW [14]	AAR [14]	SONet
Temporal	✓	✓	✓	✓	✓	✓	✓	X
Sup.	\mathcal{B}	\mathcal{B}	\mathcal{B}	\mathcal{I}	\mathcal{I}	\mathcal{I}	\mathcal{I}	\mathcal{I}
mIoU	54.1	56.2	61.7	53.3	58.6	61.9	62.1	64.3

Table S2: Quantitative comparison of recent video object segmentation methods with various methods of supervision on the YouTube-Object dataset. \mathcal{B} : bounding box, \mathcal{I} : image-level supervision. Note that the baseline numbers are taken from [14] for fair comparison.

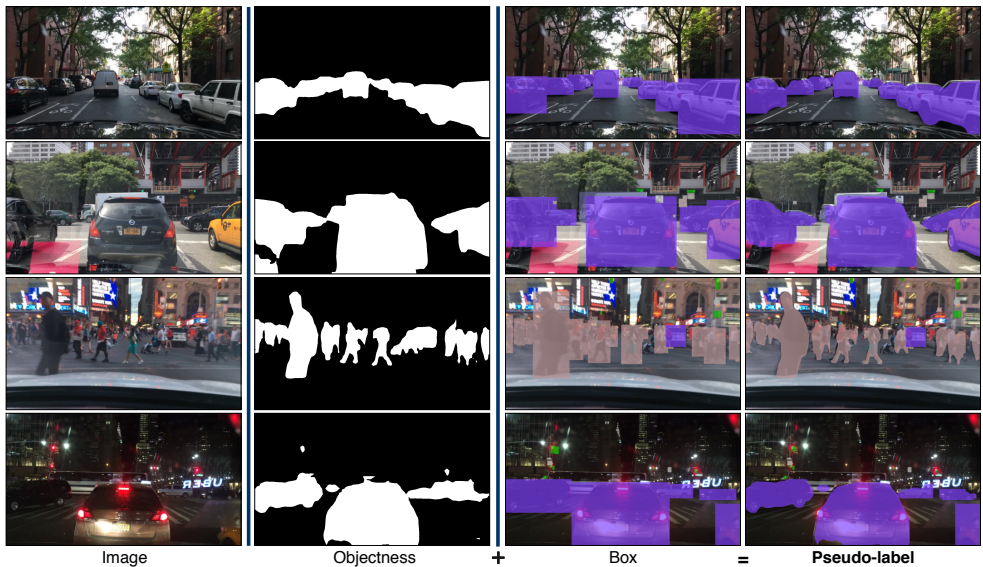


Figure S6: **Berkeley DeepDrive pseudo-labels.** Examples of generated pseudo-labels with our proposed approach on the Berkeley DeepDrive video frames.

S3.5 Generalization to Different Domains: Berkeley DeepDrive

We further apply our bounding box-driven pseudo-label generation technique on a recent driving dataset, Berkeley DeepDrive [14], to validate whether our procedure can generalize well on a dataset from a different domain. The Berkeley DeepDrive dataset [14] is composed of images of diverse road scenes (with motion blur) taken from various locations throughout the USA. We generate pseudo-labels for 100k frames which have bounding box annotation available for the 10 different categories: bus, light, sign, person, bike, truck, motor, car, train, and rider. Figure S6 presents examples of generated pseudo-labels of DeepDrive video frames. It is clear that our class agnostic objectness model can generate masks with sharp boundaries in complex driving scenarios, resulting in high-quality pseudo-labels. Since the DeepDrive dataset does not provide pixel-wise annotation for these 100k frames we can not evaluate the quality of generated pseudo-labels in terms of mIoU.

Name	Sup.	Architecture (ϕ_O)	Input	mIoU
SONet	\mathcal{G}^B	smaller network discussed in Sec. 3.3	semantic (\mathcal{S})	73.8
v1	\mathcal{G}^B	single 1×1 convolution layer	res5C	72.2
v2	\mathcal{G}^B	single 1×1 convolution layer	semantic (\mathcal{S})	73.5
v3	\mathcal{G}^B	smaller network discussed in Sec. 3.3	res5C	73.6

Table S3: Comparison of objectness branch variants for SONet on the PASCAL VOC 2012 validation set.

S3.6 Ablation Studies

We conduct further ablation studies to analyze our design and the effectiveness of the objectness branch (Sec. S3.6.1 & Sec. S3.6.2).

S3.6.1 Design Choices of Objectness Branch.

We vary the design of the objectness branch, ϕ_O , of SONet and compare the architectures against each other. The results are reported in Table S3. We evaluate three different variants: (v1) a single 1×1 convolutional layer which predicts the objectness and takes as input the final feature representation (`res5C`), (v2) a single 1×1 convolution layer which takes as input the semantic prediction (\mathcal{S}), and (v3) a smaller network is applied (as discussed in Sec. 3.2 of the main manuscript) but takes as input the features from `res5C`.

S3.6.2 Effectiveness of Objectness Branch in SONet

We provide additional qualitative examples in Fig. S7 to show the objectness branch’s effect on SONet’s semantic segmentation predictions. Note that SONet without the objectness branch is equivalent to DeepLabv3 [14]. As can be seen from the examples, the objectness branch can guide the segmentation network to produce more accurate and smooth predictions.

S3.6.3 Transferring Semantic Knowledge from Source to Target Dataset.

As an additional baseline, we directly transfer the semantic information from COCOStuff to the VOC12 dataset. Towards this goal, we first train DeepLabv3 [14] on COCOStuff to output semantic segmentation (i.e., multi-class) masks instead of objectness masks (i.e., binary). Note, similar to the objectness training, we only consider the *things* classes and use the pretrained model to generate pseudo-label (quality: 50.8% mIoU) for the VOC12 train set. Then, we train DeepLabv3 using the generated pseudo-labels, resulting in 53.4% mIoU on VOC12 val set.

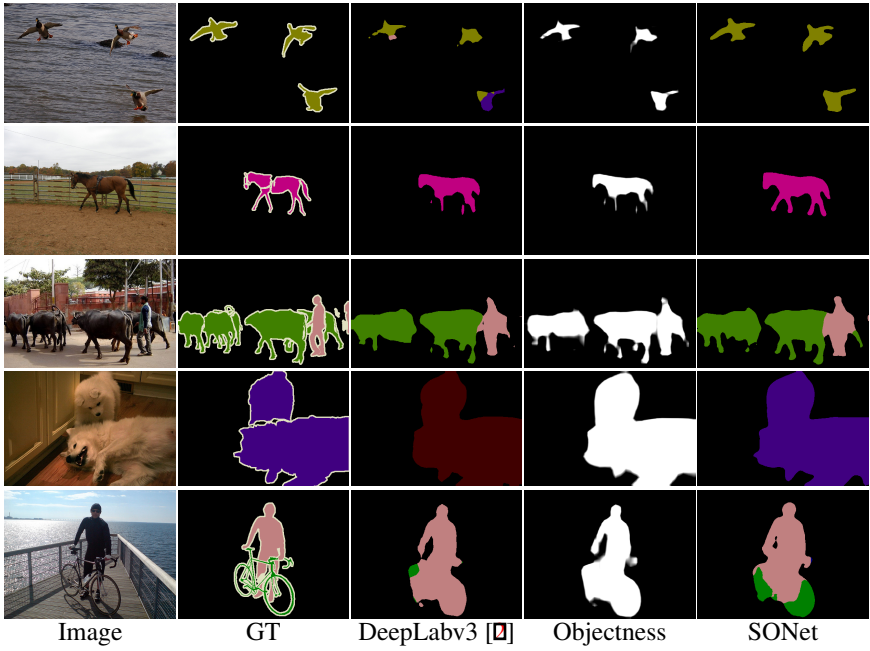


Figure S7: **Visualization of the effect of the objectness branch** on the segmentation results. Note that the difference between the baseline DeepLabv3 [1] and SONet is the objectness branch. Images are taken from the PASCAL VOC 2012 validation set.

References

- [1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017.
- [2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018.
- [3] Benjamin Drayer and Thomas Brox. Object detection, tracking, and motion segmentation for object-level video segmentation. *arXiv preprint arXiv:1608.03066*, 2016.
- [4] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017.
- [5] Suyog Dutt Jain and Kristen Grauman. Supervoxel-consistent foreground propagation in video. In *ECCV*, 2014.
- [6] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In *ICCV*, 2019.
- [7] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.
- [8] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.
- [9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- [10] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.
- [11] 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 *ICCV*, 2017.
- [12] Kevin Tang, Rahul Sukthankar, Jay Yagnik, and Li Fei-Fei. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013.
- [13] Le Yang, Junwei Han, Dingwen Zhang, Nian Liu, and Dong Zhang. Segmentation in weakly labeled videos via a semantic ranking and optical warping network. *TIP*, 2018.
- [14] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.
- [15] Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, and Changqun Xia. Semantic object segmentation via detection in weakly labeled video. In *CVPR*, 2015.
- [16] Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, Changqun Xia, and Jun Li. Semantic object segmentation in tagged videos via detection. *TPAMI*, 2017.

- [17] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.
- [18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.