arXiv:1509.03150v2 [cs.CV] 7 Dec 2016

STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation

Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, *Senior Member, IEEE* and Shuicheng Yan *Senior Member, IEEE*

Abstract—Recently, significant improvement has been made on semantic object segmentation due to the development of deep convolutional neural networks (DCNNs). Training such a DCNN usually relies on a large number of images with pixel-level segmentation masks, and annotating these images is very costly in terms of both finance and human effort. In this paper, we propose a simple to complex (STC) framework in which only image-level annotations are utilized to learn DCNNs for semantic segmentation. Specifically, we first train an initial segmentation network called Initial-DCNN with the saliency maps of simple images (*i.e.*, those with a single category of major object(s) and clean background). These saliency maps can be automatically obtained by existing bottom-up salient object detection techniques, where no supervision information is needed. Then, a better network called Enhanced-DCNN is learned with supervision from the predicted segmentation masks of simple images based on the Initial-DCNN as well as the image-level annotations. Finally, more pixel-level segmentation masks of complex images (two or more categories of objects with cluttered background), which are inferred by using Enhanced-DCNN and image-level annotations, are utilized as the supervision information to learn the Powerful-DCNN for semantic segmentation. Our method utilizes 40K simple images from `Flickr.com` and 10K complex images from PASCAL VOC for step-wisely boosting the segmentation network. Extensive experimental results on PASCAL VOC 2012 segmentation benchmark well demonstrate the superiority of the proposed STC framework compared with other state-of-the-arts.

Index Terms—semantic segmentation, weakly-supervised learning, convolutional neural network

1 Introduction

In recent years, deep convolutional neural networks (DCNNs) have demonstrated an outstanding capability in various computer vision tasks, such as image classification [1]–[4], object detection [5], [6] and semantic segmentation [7]–[13]. Most DCNNs for these tasks rely on strong supervision for training, *i.e.*, ground-truth bounding boxes and pixel-level segmentation masks. However, compared with convenient image-level labels, collecting annotations of bounding boxes or pixel-level masks is much more expensive. In particular, for the semantic segmentation task, annotating a large number of pixel-level masks usually requires a considerable amount of financial expenses as well as human effort.

To address this problem, some methods [14]–[18] have been proposed for semantic segmentation by only utilizing image-level labels as the supervised information. However, to the best of our knowledge, the performance of these methods is far from satisfactory compared with fully-supervised schemes (*e.g.*, 40.6% [15] *vs.* 66.4% [13]). Given the complexity of semantic segmentation problems, such as high intra-class variation (*e.g.*, diverse appearance, viewpoints and scale) and different interaction between objects (*e.g.*, partial visibility and occlusion), complex loss functions (*e.g.*, multiple instance learning based loss functions) [14], [15], [18] with image-level annotations may not be adequate for weakly supervised semantic segmentation due to the ignorance of intrinsic pixel-level properties of segmentation masks.

It should be noted that, during the past few years, many salient object detection methods [19]–[22], which do not require high-level supervision information, have been proposed to detect the most visually noticeable salient object in the image. While these methods may not work well for complex images with multiple objects and cluttered background, they often provide satisfactory saliency maps for images with the object(s) of single category and clean background. By automatically retrieving a huge amount of web images and detecting salient objects for relatively simple images, we might be able to obtain a large amount of saliency maps for training semantic segmentation DCNNs at a low cost.

In this work, we propose a simple to complex framework for weakly-supervised segmentation based on the following intuitions. For complex images with clutter background and two or more categories of objects, it is usually difficult to infer the relationship between semantic labels and pixels by only utilizing image-level labels as the supervision. However, for simple images with clean background and a single category of major object(s), foreground and background pixels are easily split based on the salient object detection techniques [20]–[23]. With the indication of the image-level label, it is naturally inferred that pixels belonging to

Yunchao Wei and Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University, China, e-mail: wychao1987@gmail.com; yzhao@bjtu.edu.cn. Xiaodan Liang is with Sun Yatsen University, China, e-mail: xdliang328@gmail.com. Xiaohui Shen is with Adobe Research, U.S., e-mail: xshen@adobe.com. Ming-Ming Cheng is with CCCE, Nankai University, Tianjin, China, e-mail: cmm@nankai.edu.cn. Yunpeng Chen, Jiashi Feng and Shuicheng Yan are with Department of Electrical and Computer Engineering, National University of Singapore, e-mail: qw.2080@gmail.com; elefjia@nus.edu.sg; eleyans@nus.edu.sg.

the foreground can be assigned with the same semantic label. Therefore, an initial segmenter can be learned from simple images based on their foreground/background masks and image-level labels. Furthermore, based on the initial segmenter, more objects from complex images can be segmented so that a more powerful segmenter can be continually learnt for semantic segmentation.

Specifically, semantic labels are firstly employed as queries to retrieve images on the image hosting websites, *e.g.*, `Flickr.com`. The retrieved images from the first several pages usually meet the definition of a simple image. With these simple images, high quality saliency maps are generated by the state-of-the-art saliency detection technique [22]. Based on the supervision of image-level labels, we can easily assign a semantic label to each foreground pixel and learn a semantic segmentation DCNN supervised by the generated saliency maps by employing a multi-label cross-entropy loss function, in which each pixel is classified to both the *foreground* class and *background* according to the predicted probabilities embedded in the saliency map. Then, a simple to complex learning process is utilized to gradually improve the capability of DCNN, in which the predicted segmentation masks of simple images by initially learned DCNN are in turn used as the supervision to learn an enhanced DCNN. Finally, with the enhanced DCNN, more difficult and diverse masks from complex images are further utilized for learning a more powerful DCNN. Particularly, the contributions of this work are summarized as follows:

- We propose a simple to complex (STC) framework that can effectively train the segmentation DCNN in a weakly-supervised manner (*i.e.*, only image-level labels are provided). The proposed framework is general, and any state-of-the-art fully-supervised network structure can be incorporated to learn the segmentation network.
- A multi-label cross-entropy loss function is introduced to train a segmentation network based on saliency maps, where each pixel can adaptively contribute to the *foreground* class and *background* with different probabilities.
- We evaluate our method on the PASCAL VOC 2012 segmentation benchmark [24]. The experimental results well demonstrate the effectiveness of the STC framework, achieving the state-of-the-art performance.

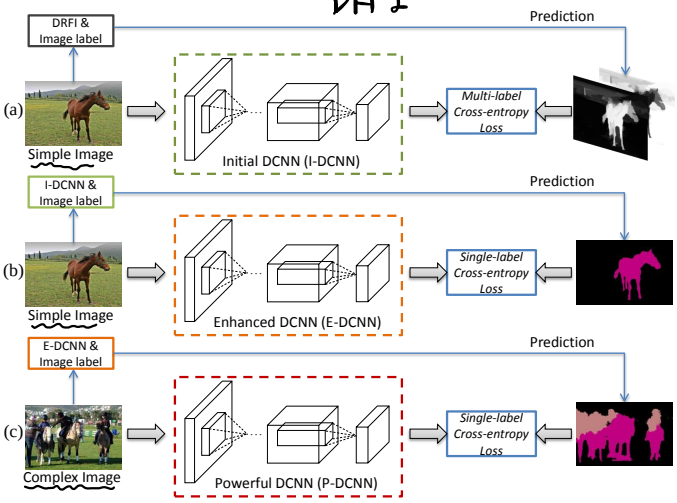

Fig. 1. An illustration of the proposed simple to complex (STC) framework. (a) High quality saliency maps of simple images are first generated by DRFI [22] as the supervised foreground/background masks to train the Initial-DCNN using the proposed loss function. (b) Then, a better Enhanced-DCNN is learned, supervised with the segmentation masks predicted by Initial-DCNN. (c) Finally, more masks of complex images are predicted to train a more powerful network, called Powerful-DCNN.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

To reduce the burden of the pixel-level mask annotation, some weakly-supervised methods have been proposed for semantic segmentation. Dai *et al.* [8] and Papandreou *et al.* [14] proposed to estimate semantic segmentation masks by utilizing annotated bounding boxes. For example, by incorporating pixel-level masks from the Pascal VOC [24] and annotated bounding boxes from the COCO [25], state-of-the-art results on PASCAL VOC 2012 benchmark were achieved by [8]. To further reduce the burden of the bounding boxes collection, some works [14]–[16], [18], [26]–[28] proposed to train the segmentation network by only using image-level labels. Pathak *et al.* [16] and Pinheiro *et al.* [15] proposed to utilize multiple instance learning (MIL) [29] framework to train the DCNN for segmentation. In [14], an alternative training procedure based on Expectation-Maximization (EM) algorithm was presented to dynamically predict foreground (with semantics)/background pixels. Pathak *et al.* [18] introduced constrained convolutional neural networks for weakly-supervised segmentation. Specifically, by utilizing object size as additional supervision, significant improvements were made by [18]. Most recently, three kinds of loss functions, *i.e.*, seeding, expansion and constrain-to-boundary, were leveraged in [28] to train the segmentation network. Saleh *et al.* [27] also proposed a relevant approach using foreground/background prior for learning to segment, which is able to evidence the effectiveness of our framework.

2.2 Self-paced Learning

Our framework first learns from simple images and then applies the learned network to complex ones, which is related to self-paced learning [30]. Recently, various computer vision applications [31]–[33] based on self-paced learning have been proposed. In specific, Tang *et al.* [31] adapted object detectors learned from images to videos by starting with easy samples. Jiang *et al.* [32] addressed

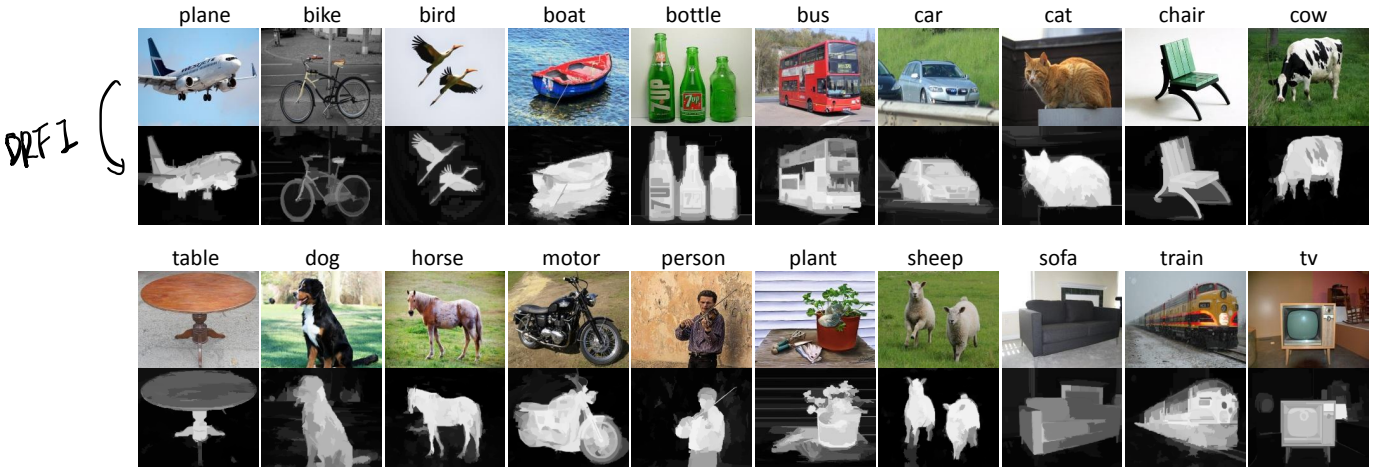

Fig. 2. Examples of simple images and the corresponding saliency maps generated by DRFI on the 20 classes of PASCAL VOC.

the data diversity. In [33], very few samples were used as seeds to train a weak object detector, and then more instances were iteratively accumulated to enhance the object detector, which can be considered as a slightly-supervised self-paced learning method. However, different from self-paced learning where each iteration automatically selects samples for training, the simple or complex samples are defined according to their appearance (*e.g.*, single/multiple object(s) or clean/cluttered background) before training in this work.

Besides, many other works [17], [34]–[37] have also addressed this task. These methods are usually applied on simple or small scale datasets, *e.g.*, MSRA [38] and SIFT-flow [39]. Specifically, Liu *et al.* [35] proposed a graph propagation method to automatically assign the annotated labels at image level to those contextually derived semantic regions. Xu *et al.* [34] presented a latent structured prediction framework, where the graphical model encodes the presence and absence of a class as well as assignments of semantic labels to super-pixels. Vezhnevets *et al.* [37] proposed a maximum expected agreement model selection principle that evaluates the quality of a model from the parametric family of structured models for semantic segmentation.

3 Proposed Method

Figure 1 shows the architecture of the proposed simple to complex (STC) framework. We utilize the state-of-the-art saliency detection method, *i.e.*, discriminative regional feature integration (DRFI) [22], to generate the saliency maps of simple images. The produced saliency maps are first employed to train an initial DCNN with a multi-label cross-entropy loss function. Then the simple to complex framework is proposed to gradually improve the capability of segmentation DCNN.

3.1 Initial-DCNN

For the generated saliency map of each image, the larger pixel value means it is more likely that this pixel belongs to foreground. Figure 2 shows some instances of simple images and the corresponding saliency maps generated by DRFI. It can be observed that there exists explicit association between the foreground pixels and the semantic object(s). Since each simple image is accompanied with a semantic label, it can be easily inferred that foreground candidate pixels can be assigned with the corresponding image-level label. Then, a multi-label cross-entropy loss function is proposed to train the segmentation network supervised by saliency maps.

Suppose there are C classes in the training set. We denote $\mathcal{O}_I = \{1, 2, \cdots, C\}$ and $\mathcal{O}_P = \{0, 1, 2, \cdots, C\}$ as the category sets for image-level label and pixel-level label, respectively, where 0 indicates the *background* class. We denote the segmentation network filtering by $f(\cdot)$, where all the convolutional layers filter the given image I. The $f(\cdot)$ produces a $h \times w \times (C+1)$ dimensional output of activations, where h and w are the height and the width of the feature map for each channel, respectively. We utilize the softmax function to compute the posterior probability of each pixel of I belonging to the k^{th} $(k \in \mathcal{O}_P)$ class, which is formulated as follows,

$$p_{ij}^k = \frac{\exp\left(f_{ij}^k(I)\right)}{\sum_{l \in \mathcal{O}_P} \exp\left(f_{ij}^l(I)\right)}, \tag{1}$$

where $f_{ij}^k(I)$ is the activation value at location (i, j) $(1 \le i \le h, 1 \le j \le w)$ of the k^{th} feature map. In general, we define the probability obtained from the saliency map of the l^{th} class at the location (i, j) as $\hat{p}_{ij}^l$ $(\sum_{l \in \mathcal{O}_P} \hat{p}_{ij}^l = 1)$. Then, the multi-label cross-entropy loss function for semantic segmentation is then defined as

$$-\frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{l \in \mathcal{O}_P} \hat{p}_{ij}^l \log(p_{ij}^l). \tag{2}$$

Specifically, for each simple image, we assume that only one semantic label is included. Suppose that the simple image I is annotated by the c^{th} $(c \in \mathcal{O}_I)$ class, and then the normalized value from the saliency map is

taken as the probability of each pixel belonging to the class c. We resize the saliency map to the same size of the output feature map from the DCNN and Eqn. (2) can then be re-formulated as

$$-\frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} (\hat{p}_{ij}^{c} \log(p_{ij}^{c}) + \hat{p}_{ij}^{0} \log(p_{ij}^{0})), \quad (3)$$

where p_{ij}^{0} indicates the probability of the pixel at location (i, j) belonging to the *background* ($p_{ij}^{0} = 1 - p_{ij}^{c}$). We denote the segmentation network learned in this stage as Initial-DCNN (I-DCNN for short).

It should be noted that we can also utilize SaliencyCut [20] to generate the foreground/background segmentation masks based on the generated saliency maps. Then, single-label cross-entropy loss can be employed for training. We compare this scheme with our proposed method, and find that the performance on VOC 2012 *val* set will drop by 3%. The reason is that some saliency detection results are inaccurate. Therefore, directly applying SaliencyCut [20] to generate segmentation masks will introduce many nosies, which is harmful for training the I-DCNN. However, based on the proposed multi-label cross-entropy loss, correct semantic labels will still contribute to the optimization, which can decrease the negative effect caused by low quality saliency maps.

3.2 Simple to Complex Framework

In this section, a progressively training strategy is proposed by incorporating more complex images with image-level labels to enhance the segmentation capability of DCNN. Based on the trained I-DCNN, segmentation masks of images can be predicted, which can be used to further improve the segmentation capability of DCNN. Similar to the definition in Section 3.1, we denote the predicted probability for the k^{th} class at the location (i, j) as p_{ij}^{k}. Then, the estimated label g_{ij} of the pixel at location (i, j) by the segmentation DCNN can be formulated as

$$g_{ij} = \arg\max_{k \in \mathcal{O}_P} p_{ij}^{k}. \quad (4)$$

3.2.1 Enhanced-DCNN

However, incorrect predictions from the I-DCNN may lead to the drift in semantic segmentation when used as the supervision for training DCNN. Fortunately, for each simple image in the training set, the image-level label is given, which can be utilized to refine the predicted segmentation mask. Specifically, if the simple image I is labeled with c ($c \in \mathcal{O}_I$), the estimated label of the pixel can be re-formulated as

$$g_{ij} = \arg\max_{k \in \{0, c\}} p_{ij}^{k}, \quad (5)$$

where 0 indicates the category of *background*. In this way, some false predictions for simple images in the training set can be eliminated. Then, a more powerful segmentation DCNN called Enhanced-DCNN (E-DCNN for short) is trained by utilizing the predicted segmentation masks as the supervised information. We train the E-DCNN with the single-label cross-entropy loss function, which is widely used by fully-supervised schemes [11].

3.2.2 Powerful-DCNN

In this stage, complex images with image-level labels, in which more semantic objects and cluttered background are included, are utilized to train the segmentation DCNN. Compared with I-DCNN, E-DCNN possesses a more powerful semantic segmentation capability due to the usage of the large number of predicted segmentation masks. Although E-DCNN is trained with simple images, the semantic objects in those images have large variety in terms of appearance, scale and viewpoint, which is consistent with their appearance variation in complex images. Therefore, we can apply E-DCNN to predict the segmentation masks of complex images. Similar as Eqn. (5), to eliminate false predictions, the estimated label for each pixel of image I is formulated as

$$g_{ij} = \arg\max_{k \in \Omega} p_{ij}^{k}, \quad (6)$$

where Ω indicates the set of ground-truth semantic labels (including *background*) for each image I. We denote the segmentation network trained in this stage as Powerful-DCNN (P-DCNN for short).

In this work, two kinds of cross-entropy losses are utilized to train segmentation networks. In particular, cross-entropy loss in the fully convolutional network is a pixel-wise one. For the fully supervised scheme, each pixel can only be assigned to one class and the corresponding cross-entropy is a single-label one. This matches the target of E-DCNN and P-DCNN. Therefore, we train these two networks using the single-label loss. For training the I-DCNN, the class information of each pixel can not be exactly obtained. To address this issue, each pixel is softly associated with two classes (one is background and the other is one of the 20 foreground classes) with different probabilities according to the produced saliency map and image-level label. We consider the loss function for this scheme as the multi-label cross-entropy loss. To illustrate the effectiveness of each step, some segmentation results generated by I-DCNN, E-DCNN and P-DCNN are shown in Figure 3. It can be seen that the segmentation results are progressively becoming better based on the proposed simple to complex framework.

4 Experimental Results

4.1 Dataset

Flickr-Clean: We construct a new dataset called `Flickr-Clean` to train the segmentation network of I-DCNN. The keywords, whose semantics are consistent with those from PASCAL VOC, are employed as queries to retrieve images on the image hosting website `Flickr.com`. We crawl images in the first several pages of searching results and use the state-of-the-art saliency

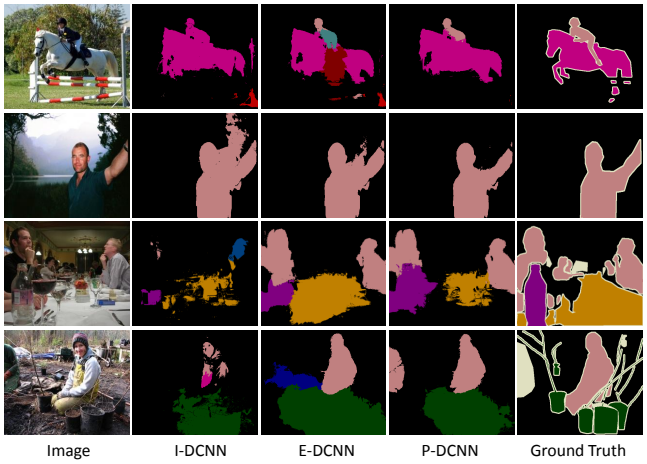

Fig. 3. Examples of segmentation results generated by I-DCNN, E-DCNN and P-DCNN on the PASCAL VOC 2012 *val* set, respectively.

detection method, *i.e.*, discriminative regional feature integration (DRFI), to generate the saliency maps of the crawled images. In order to ensure that the images are simple ones, we adopt a method similar to that proposed in [40], [41] to filter the crawled images. We measure the impreciseness and incompleteness of the Saliency-Cut [20] segmentation as in [40]. Denote the number of pixels of the given image I as N_I and the number of foreground pixels of the corresponding segmentation mask as N_f. We reserve those images whose foreground regions fit $0.3*N_I < N_f < 0.5*N_I$. Without such a filtering scheme to clean up the training set, the performance of using all 100K crawled images for training results in 4% performance drop for the I-DCNN. In the end, 41,625 images are collected to train the segmentation network.

PASCAL VOC 2012: The proposed weakly-supervised method is evaluated on the PASCAL VOC 2012 segmentation benchmark [24]. The original training data contain 1,464 images. In [42], 10,582 extra images ($train\_aug$) are annotated for training. In our experiment, 10,582 images with only image-level labels are utilized as the complex image set for training. The *val* and *test* sets have 1,449 and 1,456 images, respectively. For both *val* and *test* sets, we only use the simple images from `Flickr-Clean` and the complex images from $train\_aug$ for training. The performance is measured in terms of pixel intersection-over-union (IoU) averaged on 21 classes (20 *object* and one *background*). Extensive evaluation of the weakly-supervised method is primarily conducted on the *val* set and we also report the result on the *test* set (whose ground-truth masks are not released) by submitting the results to the official PASCAL VOC 2012 server.

4.2 Training Strategies

We employ the proposed simple to complex framework to learn the DCNN component of the DeepLab-CRF model [13], whose parameters are initialized by the VGG-16 model [2] pre-trained on ImageNet [43]. For the training of segmentation DCNNs (I-DCNN, E-DCNN and P-DCNN), we use a mini-batch size of 8 images. Every training image is resized to $330 \times n$ and patches with the size of 321×321 are randomly cropped during the training stage. The initial learning rate is set as 0.001 (0.01 for the last layer) and divided by 10 after almost every 5 epochs. The momentum and the weight decay are set as 0.9 and 0.0005. The training progress is performed for about 15 epochs. To fairly compare our results with those from [14], [18], dense CRF inference is adopted to post-process the predicted results. Each segmentation DCNN is trained based on a NVIDIA GeForce Titan GPU with 6GB memory. All the experiments are conducted using DeepLab code [13], which is implemented based on the publicly available Caffe framework [44].

TABLE 1
Comparison of I-DCNN models trained with different saliency maps on VOC 2012 *val* set (mIoU in %).

Saliency Method	HS [21]	DRFI [22]
I-DCNN	42.5	44.9

TABLE 2
Comparison of I-DCNN models trained on different numbers of images on VOC 2012 *val* set (mIoU in %).

`Flickr-Clean`	1/16	1/8	1/4	1/2	All
I-DCNN	39.8	42.1	45.7	45.6	44.9

4.3 Justifications

Justifications of Different Saliency Detecton Methods: DRFI achieves the state-of-the-art performance on 6 popular benchmark datasets indicated by [19]. To investigate the quality of saliency maps generated by DRFI and how the performance of our proposed method varies with adopting different saliency detection methods, we train another I-DCNN model based on saliency maps produced by one of the latest methods, *i.e.*, Hierarchical Saliency (HS) [21] detection method. Table 1 shows the segmentation results of I-DCNN models trained by different saliency maps. It can be seen that using DRFI saliency maps to train I-DCNN is effective.

Justifications of the Number of Training Images: To investigate when the increasing number of collected training images will saturate the performance of the proposed method, we train I-DCNN models using varying numbers of images from `Flickr-Clean` dataset (see also Table 2). Each smaller set is a subset of the following larger set. For example, the 1/16 set is a subset of the 1/8 set. We firstly observe performance improvements when incorporating more training samples, which is quite intuitive. After getting best performance when using 1/4 of the training samples, further increasing the training samples hurts the performance. We believe the reason is that Flickr images ranked in the last few pages are quite noisy, and cannot be efficiently utilized by our current scheme. In this paper, our experiments are based on all images from `Flickr-Clean`.

TABLE 3
Comparison of different segmentation DCNNs on VOC 2012 *val* set.

Method	Training Set	mIoU
I-DCNN	Flickr-Clean	44.9
E-DCNN	Flickr-Clean	46.3
P-DCNN	Flickr-Clean + VOC	49.8

Justifications of the Simple to Complex Framework: Table 3 shows the comparisons of different segmentation DCNNs. It can be observed that, based on the proposed multi-label cross-entropy loss, saliency maps of simple images accompanied with image-level labels can be conveniently employed to train an effective neural network for semantic segmentation. The performance of I-DCNN is 44.9%, which can outperform most state-of-the-arts. Besides, training with the segmentation masks predicted by I-DCNN can further improve the capability of semantic segmentation, *i.e.*, 46.3% *vs.* 44.9%. In addition, based on the enhanced neural network (*i.e.*, E-DCNN), the performance can be further boosted, *i.e.*, 49.8% *vs.* 46.3%, by adding more complex images for training. Therefore, for the weakly-supervised semantic segmentation task, the proposed simple to complex (STC) framework is effective. In addition, we also conduct experiments of training I-DCNN with complex images to validate the necessity of using simple images. The mIoU score is 17.6%, which is far below the result of ours. Please refer to the supplementary marital for more details.

4.4 Comparison with State-of-the-art Methods

Table 4 shows the detailed results of ours compared with those of state-of-the-art methods. * indicates those methods that use additional images to train the segmentation network. For MIL-FCN [16], EM-Adapt [14], CCNN [18], DCSM [26], BFBP [27] and SEC [28], the segmentation networks are trained on $train\_aug$ taken from VOC 2012. For MIL-ILP-* [15], the segmentation network is trained with 700K images for 21 classes taken from ILSVRC 2013. Image-level prior (ILP), and some smooth priors, *i.e.*, superpixels (-sppxl), BING [45] boxes (-bb) and MCG [46] segmentations (-seg), are utilized for post-processing to further boost the segmentation results. The proposed framework is learned on 50K (40K simple images from `Flickr-Clean` and 10K complex images from PASCAL VOC) images, which are much fewer compared with those of [15] (700K). Surprisingly, our result can make a significant improvement compared with the best result of [15] (49.8% *vs.* 42.0%). It can be observed that SEC [28] achieves the state-of-the-art performance on this challenging task. The superiority of SEC mainly benefits from using CRF-based constrain-to-boundary loss for network optimizing. By only using cross-entropy loss, the mIoU score reported in [28] is 45.4%. Based on simple images that are cheap to obtain, our STC framework can easily achieve the competitive performance (49.8% *vs.* 50.7%) by simply employing cross-entropy loss.

Table 5 reports our results on PASCAL VOC 2012 *test* set and compare them with the state-of-the-art weakly-supervised methods. It can be observed that our result is competitive compared with the state-of-the-art performance (51.2% *vs.* 51.7%). For EM-Adapt [14], the segmentation network is learned based on $train\_aug$ and val sets. In [18], by adding additional supervision of object size information, the performance can be improved from 35.5% to 45.1%. We also compare our result with several fully-supervised methods in Table 5. It can be observed that we have made a significant improvement to approach those results learned with fully supervised schemes. In particular, our weakly-supervised framework achieves similar results compared with SDS [9], which is learned in a fully supervised manner. Besides, we conduct additional experiments based on the semi-supervised setting. The experimental results demonstrate that STC can also boost the segmentation performance when only a small number of fully-supervised images is available. More detailed comparative analyses are provided in the supplementary material.

Qualitative segmentation results obtained by the proposed framework are shown in Figure 4. Some failure cases are shown in the last row of Figure 4. In the first case (row: 6, column: 1), the *chair* has a similar appearance as *sofa* and the pixels of foreground segmentation are totally predicted as *sofa*. In the second (row: 6, column: 2) and the third case (row: 6, column: 3), *sofa* which occupies a large region of the image is wrongly predicted as *background*. Including more samples with clean background and various appearances for training or using classification results for post-processing may help solve these issues.

4.5 Discussion

The comparison between the proposed STC and [15] is a little unfair. The deep neural network utilized in [15] is based on `OverFeat` [47], in which there are 10 weight layers, while in this paper, we utilize the VGG-16 model, which has 16 weight layers, as the basic architecture of the segmentation network. Both two models are pre-trained on ImageNet and the VGG-16 model works better than the `OverFeat` model on the ILSVRC [43] classification task. However, Pinheiro *et al.* [15] utilized 700K images with image-level labels for training, which is a much larger number compared with the training set (50K) of ours. In addition, the performance of [15] highly depends on complex post-processing. Without any post-processing step, the performance of [15] is 17.8%, which is far below the result of ours, *i.e.*, 49.8%.

ACKNOWLEDGMENTS

This work was sponsored by the National Key Research and Development of China (NO. 2016YFB0800404), the National Natural Science Foundation of China (NO. 61532005, NO. 61210006, NO. 61402268, NO. 61572264), and CAST young talents plan.

TABLE 4
Comparison of weakly-supervised semantic segmentation methods on VOC 2012 *val* set.

Methods	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
Other state-of-the-art methods:																						
MIL-FCN [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25.7
EM-Adapt [14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.2
CCNN [18]	65.9	23.8	17.6	22.8	19.4	36.2	47.3	46.9	47.0	**16.3**	36.1	22.2	43.2	33.7	44.9	39.8	29.9	33.4	22.2	38.8	36.3	34.5
MIL* [15]	37.0	10.4	12.4	10.8	5.3	5.7	25.2	21.1	25.2	4.8	21.5	8.6	29.1	25.1	23.6	25.5	12.0	28.4	8.9	22.0	11.6	17.8
MIL-ILP* [15]	73.2	25.4	18.2	22.7	21.5	28.6	39.5	44.7	46.6	11.9	40.4	11.8	45.6	40.1	35.5	35.2	20.8	41.7	17.0	34.7	30.4	32.6
MIL-ILP-sppxl* [15]	77.2	37.3	18.4	25.4	28.2	31.9	41.6	48.1	50.7	12.7	45.7	14.6	50.9	44.1	39.2	37.9	28.3	44.0	19.6	37.6	35.0	36.6
MIL-ILP-bb* [15]	78.6	46.9	18.6	27.9	30.7	38.4	44.0	49.6	49.8	11.6	44.7	14.6	50.4	44.7	40.8	38.5	26.0	45.0	20.5	36.9	34.8	37.8
MIL-ILP-seg* [15]	79.6	50.2	21.6	40.6	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	**36.9**	42.0
DCSM [26]	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	38.2	55.4	32.2	42.6	39.6	44.1
BFBP [27]	79.2	60.1	20.4	50.7	41.2	46.3	62.6	49.2	62.3	13.3	49.7	38.1	58.4	49.0	57.0	48.2	27.8	55.1	29.6	54.6	26.6	46.6
SEC [28]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62 3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
Ours:																						
STC*	84.5	68.0	19.5	60.5	42.5	44.8	68.4	64.0	64.8	14.5	52.0	22.8	58.0	55.3	57.8	60.5	40.6	56.7	23.0	57.1	31.2	49.8

TABLE 5
Comparison of fully- and weakly- supervised semantic segmentation methods on VOC 2012 *test* set.

Methods	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
Fully Supervised:																						
SDS [9]	86.3	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	51.6
FCN-8s [11]	-	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLab-CRF [13]	92.1	78.4	33.1	78.2	55.6	65.3	81.3	75.5	78.6	25.3	69.2	52.7	75.2	69.0	79.1	77.6	54.7	78.3	45.1	73.3	56.2	66.4
Weakly Supervised (other state-of-the-art methods):																						
MIL-FCN [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.9
EM-Adapt [14]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
CCNN [18]	-	21.3	17.7	22.8	17.9	38.3	51.3	43.9	51.4	15.6	38.4	17.4	46.5	38.6	53.3	40.6	34.3	36.8	20.1	32.9	38.0	35.5
MIL-ILP-sppxl* [15]	74.7	38.8	19.8	27.5	21.7	32.8	40.0	50.1	47.1	7.2	44.8	15.8	49.4	47.3	36.6	36.4	24.3	44.5	21.0	31.5	41.3	35.8
MIL-ILP-bb* [15]	76.2	42.8	20.9	29.6	25.9	38.5	40.6	51.7	49.0	9.1	43.5	16.2	50.1	46.0	35.8	38.0	22.1	44.5	22.4	30.8	43.0	37.0
MIL-ILP-seg* [15]	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6
DCSM [26]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
BFBP [27]	80.3	57.5	24.1	66.9	31.7	43.0	67.5	48.6	56.7	12.6	50.9	42.6	59.4	52.9	65.0	44.8	41.3	51.1	33.7	44.4	33.2	48.0
SEC [28]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
Weakly Supervised (ours):																						
STC*	85.2	62.7	21.1	58.0	31.4	55.0	68.8	63.9	63.7	14.2	57.6	28.3	63.0	59.8	67.6	61.7	42.9	61.0	23.2	52.4	33.1	51.2

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[4] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE TPAMI*, vol. 38, no. 9, pp. 1901–1907, 2016.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE CVPR*, 2014, pp. 580–587.

[6] R. Girshick, "Fast r-cnn," in *arXiv preprint arXiv:1504.08083*, 2015.

[7] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan, "Semantic segmentation without annotating segments," in *IEEE ICCV*, 2013, pp. 2176–2183.

[8] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," *arXiv preprint arXiv:1503.01640*, 2015.

[9] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *ECCV*, 2014, pp. 297–312.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," *arXiv preprint arXiv:1502.03240*, 2015.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015.

[12] J. Dai, K. He, and J. Sun, "Convolutional feature masking for joint object and stuff segmentation," in *IEEE CVPR*, 2015.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *preprint arXiv:1412.7062*, 2014.

[14] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a dcnn for semantic image segmentation," *arXiv preprint arXiv:1502.02734*, 2015.

[15] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *IEEE CVPR*, 2015, pp. 1713–1721.

[16] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," *arXiv preprint arXiv:1412.7144*, 2014.

[17] J. Xu, A. G. Schwing, and R. Urtasun, "Learning to segment under various forms of weak supervision," in *IEEE CVPR*, 2015.

[18] D. Pathak, P. Krähenbühl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," *arXiv preprint arXiv:1506.03648*, 2015.

[19] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE TIP*, vol. 24, no. 12, pp. 5706–5722, 2015.

[20] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE TPAMI*, vol. 37, no. 3, pp. 569–582, 2015.

[21] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *IEEE TPAMI*, vol. 38, no. 4, pp. 717–729, 2016.

[22] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient

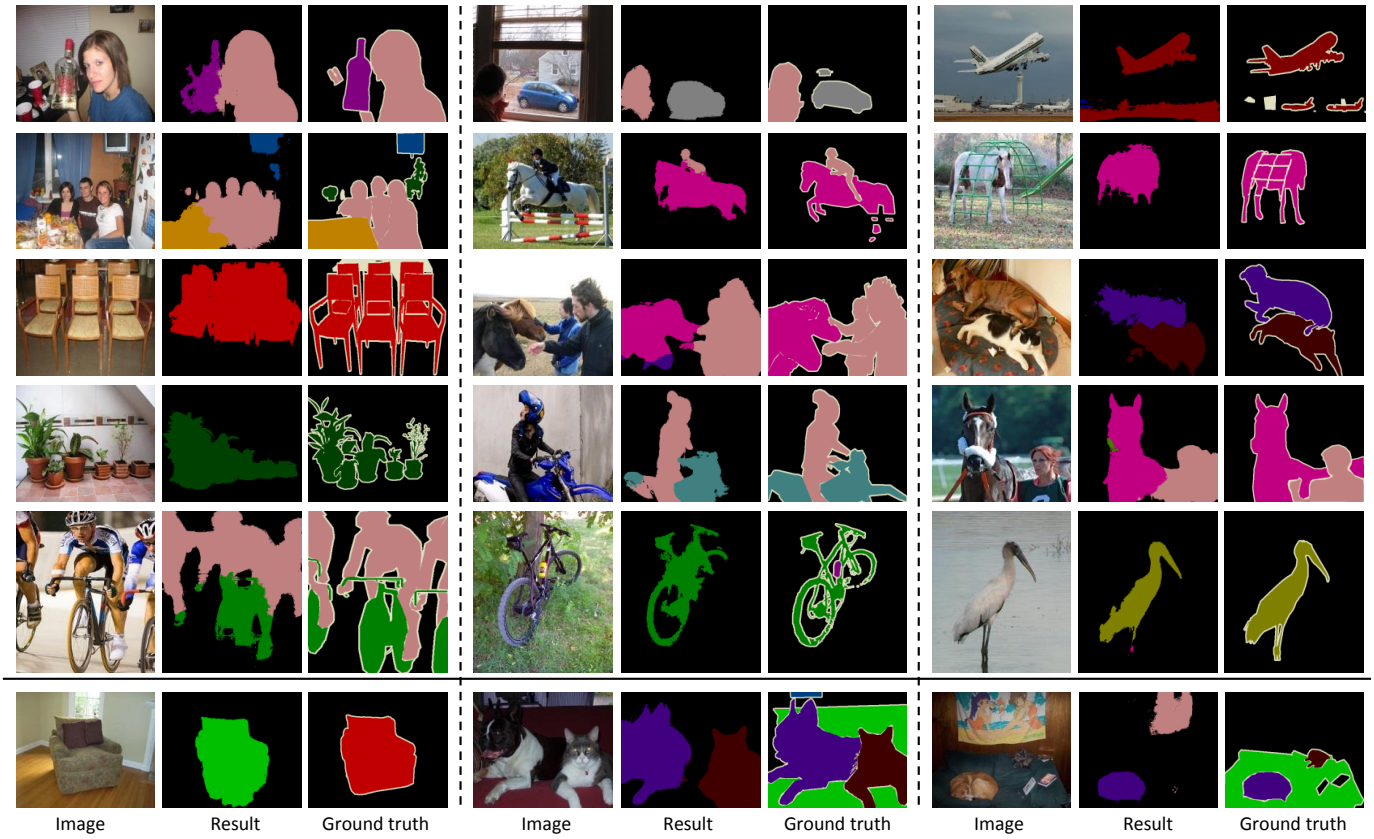

Fig. 4. Qualitative segmentation results on PASCAL VOC 2012 *val* set. Some failure cases are shown in the last row.

object detection: A discriminative regional feature integration approach," in *IEEE CVPR*, 2013, pp. 2083–2090.
[23] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, no. 11, pp. 1254–1259, 1998.
[24] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2014.
[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[26] W. Shimoda and K. Yanai, "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in *ECCV*, 2016, pp. 218–234.
[27] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *ECCV*, 2016, pp. 413–432.
[28] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *ECCV*, 2016, pp. 695–711.
[29] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *NIPS*, 1998, pp. 570–576.
[30] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *NIPS*, 2010, pp. 1189–1197.
[31] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *NIPS*, 2012, pp. 638–646.
[32] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *NIPS*, 2014, pp. 2078–2086.
[33] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan, "Towards computational baby learning: A weakly-supervised approach for object detection," in *IEEE ICCV*, 2015, pp. 999–1007.
[34] J. Xu, A. G. Schwing, and R. Urtasun, "Tell me what you see and i will show you where it is," in *IEEE CVPR*, 2014, pp. 3190–3197.
[35] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, and H. Lu, "Weakly supervised graph propagation towards collective image parsing," *IEEE Transactions on Multimedia*, vol. 14, no. 2, pp. 361–373, 2012.
[36] M. Rubinstein, C. Liu, and W. T. Freeman, "Annotation propagation in large image databases via dense image correspondence," in *ECCV*, 2012, pp. 85–99.
[37] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *IEEE CVPR*, 2012, pp. 845–852.
[38] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *IJCV*, vol. 81, no. 1, pp. 2–23, 2009.
[39] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *IEEE CVPR*, 2009, pp. 1972–1979.
[40] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: internet image montage," *ACM TOG*, vol. 28, no. 5, p. 124, 2009.
[41] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: Group saliency in image collections," *The Visual Computer*, vol. 30, no. 4, pp. 443–453, 2014.
[42] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *IEEE ICCV*, 2011, pp. 991–998.
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.
[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014, pp. 675–678.
[45] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *IEEE CVPR*, 2014, pp. 3286–3293.
[46] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *IEEE CVPR*, 2014, pp. 328–335.
[47] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *ICLR*, 2014.