Adversarial Learning for Semi-Supervised Semantic Segmentation

Wei-Chih Hung[1]
whung8@ucmerced.edu
Yi-Hsuan Tsai[2]
ytsai@nec-labs.com
Yan-Ting Liou[34]
lyt@csie.ntu.edu.tw
Yen-Yu Lin[4]
yylin@citi.sinica.edu.tw
Ming-Hsuan Yang[15]
mhyang@ucmerced.edu

[1] University of California, Merced
[2] NEC Laboratories America
[3] National Taiwan University
[4] Academia Sinica, Taiwan
[5] Google Cloud

Abstract

We propose a method for semi-supervised semantic segmentation using an adversarial network. While most existing discriminators are trained to classify input images as real or fake on the image level, we design a discriminator in a fully convolutional manner to differentiate the predicted probability maps from the ground truth segmentation distribution with the consideration of the spatial resolution. We show that the proposed discriminator can be used to improve semantic segmentation accuracy by coupling the adversarial loss with the standard cross entropy loss of the proposed model. In addition, the fully convolutional discriminator enables semi-supervised learning through discovering the trustworthy regions in predicted results of unlabeled images, thereby providing additional supervisory signals. In contrast to existing methods that utilize weakly-labeled images, our method leverages unlabeled images to enhance the segmentation model. Experimental results on the PASCAL VOC 2012 and Cityscapes datasets demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Semantic segmentation aims to assign a semantic label, e.g., person, dog, or road, to each pixel in images. This task is of essential importance to a wide range of applications, such as autonomous driving and image editing. Numerous methods have been proposed to tackle this task [17, 25, 27, 28, 43, 44], and abundant benchmark datasets have been constructed [5, 9, 32, 45] with focus on different sets of scene/object categories as well as various real-world applications. However, this task remains challenging because of large object/scene appearance variations, occlusions, and lack of context understanding. Convolutional Neural Network (CNN) based methods, such as the Fully Convolutional Network (FCN) [28], have recently achieved significant improvement on the task of semantic segmentation, and most state-of-the-art algorithms are based on FCN and additional modules.

Although CNN-based approaches have achieved astonishing performance, they require an enormous amount of training data. Different from image classification and object detection, semantic segmentation requires accurate per-pixel annotations for each training image, which can cost considerable expense and time. To ease the effort of acquiring high-quality data, semi/weakly-supervised methods have been applied to the task of semantic segmentation. These methods often assume that there are additional annotations on the image level [15, 33, 34, 36, 37], box level [6], or point level [2].

In this paper, we propose a semi-supervised semantic segmentation algorithm based on adversarial learning. The recent success of Generative Adversarial Networks (GANs) [10] facilitate effective unsupervised and semi-supervised learning in numerous tasks. A typical GAN consists of two sub-networks, i.e., generator and discriminator, in which these two sub-networks play a min-max game in the training process. The generator takes a sample vector and outputs a sample of the target data distribution, e.g., human faces, while the discriminator aims to differentiate generated samples from target ones. The generator is then trained to confuse the discriminator through back-propagation and therefore generates samples that are similar to those from the target distribution. In this paper, we apply a similar methodology and treat the segmentation network as the generator in a GAN framework. Different from the typical generators that are trained to generate images from noise vectors, our segmentation network outputs the probability maps of the semantic labels given an input image. Under this setting, we enforce the outputs of the segmentation network as close as possible to the ground truth label maps spatially.

To this end, we adopt an adversarial learning scheme and propose a fully convolutional discriminator that learns to differentiate ground truth label maps from probability maps of segmentation predictions. Combined with the cross-entropy loss, our method uses an adversarial loss that encourages the segmentation network to produce predicted probability maps close to the ground truth label maps in a high-order structure. The idea is similar to the use of probabilistic graphical models such as Conditional Random Fields (CRFs) [4, 25, 44], but without the extra post-processing module during the testing phase. In addition, the discriminator is not required during inference, and thus the proposed framework does not increase any computational load during testing. By employing the adversarial learning, we further exploit the proposed scheme under the semi-supervised setting.

In this work, we combine two semi-supervised loss terms to leverage the unlabeled data. First, we utilize the confidence maps generated by our discriminator network as the supervisory signal to guide the cross-entropy loss in a self-taught manner. The confidence maps indicate which regions of the prediction distribution are close to the ground truth label distribution so that these predictions can be trusted and trained by the segmentation network via a masked cross-entropy loss. Second, we apply the adversarial loss on unlabeled data as adopted in the supervised setting, which encourages the model to predict segmentation outputs of unlabeled data close to the ground truth distributions.

The contributions of this work are summarized as follows. First, we develop an adversarial framework that improves semantic segmentation accuracy without requiring additional computation loads during inference. Second, we propose a semi-supervised framework and show that the segmentation accuracy can be further improved by adding images without any annotations. Third, we facilitate the semi-supervised learning by leveraging the discriminator network response of unlabeled images to discover trustworthy regions that facilitate the training process for segmentation. Experimental results on the PASCAL VOC 2012 [9] and Cityscapes [5] datasets validate the effectiveness of the proposed adversarial framework for semi-supervised semantic segmentation.

2 Related Work

Semantic segmentation. Recent state-of-the-art methods for semantic segmentation are based on the advances of CNNs. As proposed in [28], one can transform a classification CNN, e.g., AlexNet [21], VGG [39], or ResNet [12], to a fully-convolutional network (FCN) for the semantic segmentation task. However, it is usually expensive and difficult to label images with pixel-level annotations. To reduce the heavy efforts of labeling segmentation ground truth, numerous weakly-supervised approaches are proposed in recent years. In the weakly-supervised setting, the segmentation network is not trained at the pixel level with the fully annotated ground truth. Instead, the network is trained with various weak-supervisory signals that can be obtained easily. Image-level labels are exploited as the supervisory signal in most recent approaches. The methods in [36] and [35] use Multiple Instance Learning (MIL) to generate latent segmentation label maps for supervised training. On the other hand, Papandreou *et al*. [33] use the image-level labels to penalize the prediction of non-existent object classes, while Qi *et al*. [37] use object localization to refine the segmentation. Hong *et al*. [15] leverage the labeled images to train a classification network as the feature extractor for deconvolution. In addition to image-level supervisions, the segmentation network can also be trained with bounding boxes [6, 19], point supervision [2], or web videos [16].

However, these weakly supervised approaches do not perform as well as the fully-supervised methods especially because it is difficult to infer the detailed boundary information from weak-supervisory signals. Hence semi-supervised learning is also considered in some methods to enhance the prediction performance. In such settings, a set of fully-annotated data and weakly-labeled samples are used for training. Hong *et al*. [15] jointly train a network with image-level supervised images and a few fully-annotated frames in the encoder-decoder framework. The approaches in [6] and [33] are generalized from the weakly-supervised to the semi-supervised setting for utilizing additional annotated image data.

Different from the aforementioned methods, the proposed algorithm can leverage unlabeled images in model training, hence greatly alleviating the task of manual annotation. We treat the output of a fully convolutional discriminator as the supervisory signals, which compensate for the absence of image annotations and enable semi-supervised semantic segmentation. On the other hand, the proposed self-taught learning framework for segmentation is related to [34] where the prediction maps of unlabeled images are used as ground truth. However, in [34], the prediction maps are refined by several hand-designed constraints before training, while we learn the selection criterion for self-taught learning based on the proposed adversarial network model.

Generative adversarial networks. Since the GAN framework with its theoretical foundation is proposed [10], it draws significant attention with several improvements in implementation [1, 3, 8, 31, 38] and applciations including image generation [38], super-resolution [22, 24], optical flow [23], object detection [42], domain adaptation [13, 14, 41] and semantic segmentation [29, 40]. The work closest in scope to ours is the one proposed by [29], where the adversarial network is used to help the training process for semantic segmentation. However, this method does not achieve substantial improvement over the baseline scheme and does not tackle the semi-supervised setting. On the other hand, Souly *et al*. [40] propose to generate adversarial examples using GAN for semi-supervised semantic segmentation. However, these generated examples may not be sufficiently close to real images to help the segmentation network since view synthesis from dense labels is still challenging.

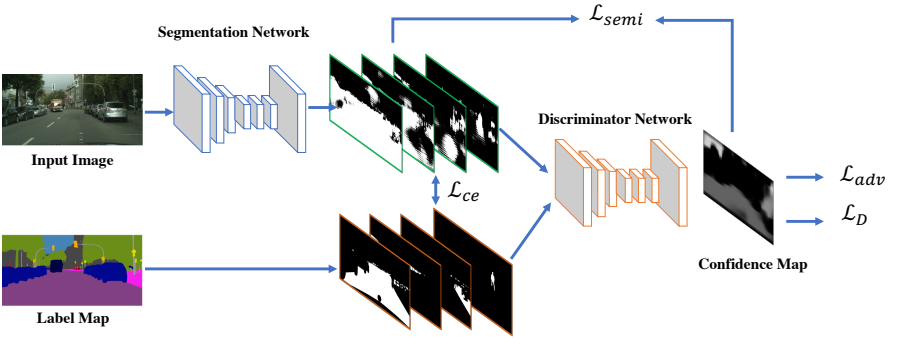

Figure 1: Overview of the proposed system for semi-supervised semantic segmentation. With a fully-convolution discriminator network trained using the loss $\mathcal{L}_D$, we optimize the segmentation network using three loss functions during the training process: cross-entropy loss $\mathcal{L}_{ce}$ on the segmentation ground truth, adversarial loss $\mathcal{L}_{adv}$ to fool the discriminator, and semi-supervised loss $\mathcal{L}_{semi}$ based on the confidence map, i.e., output of the discriminator.

3 Algorithm Overview

Figure 1 shows the overview of the proposed algorithm. The proposed model consists of two modules: segmentation and discriminator networks. The former can be any network designed for semantic segmentation, e.g., FCN [28], DeepLab [4], and DilatedNet [43]. Given an input image of dimension $H \times W \times 3$, the segmentation network outputs the class probability maps of size $H \times W \times C$, where C is the number of semantic categories.

Our discriminator network baed on a FCN, which takes class probability maps as the input, either from the segmentation network or ground truth label maps and then outputs spatial probability maps of size $H \times W \times 1$. Each pixel p of the discriminator outputs map represents whether that pixel is sampled from the ground truth label ($p = 1$) or the segmentation network ($p = 0$). In contrast to the typical GAN discriminators which take fix-sized input images (64×64 in most cases) and output a single probability value, we transform our discriminator to a fully-convolutional network that can take inputs of arbitrary sizes. More importantly, we show this transformation is essential for the proposed adversarial learning scheme.

During the training process, we use both labeled and unlabeled images under the semi-supervised setting. When using the labeled data, the segmentation network is supervised by both the standard cross-entropy loss $\mathcal{L}_{ce}$ with the ground truth label map and the adversarial loss $\mathcal{L}_{adv}$ with the discriminator network. Note that we train the discriminator network only with the labeled data. For the unlabeled data, we train the segmentation network with the proposed semi-supervised method. After obtaining the initial segmentation prediction of the unlabeled image from the segmentation network, we compute a confidence map by passing the segmentation prediction through the discriminator network. We in turn treat this confidence map as the supervisory signal using a self-taught scheme to train the segmentation network with a masked cross-entropy loss $\mathcal{L}_{semi}$. This confidence map indicates the quality of the predicted segmented regions such that the segmentation network can trust during training.

4 Semi-Supervised Training with Adversarial Network

In this section, we present the proposed network architecture and learning schemes for the segmentation as well as discriminator modules.

4.1 Network Architecture

Segmentation network. We adopt the DeepLab-v2 [4] framework with ResNet-101 [12] model pre-trained on the ImageNet dataset [7] and MSCOCO [26] as our segmentation baseline network (see Figure 1). However, we do not use the multi-scale fusion proposed in [4] since it would occupy all memory of a single GPU and make it impractical to train the discriminator. Similar to the recent semantic segmentation method [4, 43], we remove the last classification layer and modify the stride of the last two convolution layers from 2 to 1, thereby making the resolution of the output feature maps effectively $1/8$ of the input image size. To enlarge the receptive fields, we apply the dilated convolution [43] in conv4 and conv5 layers with strides of 2 and 4, respectively. In addition, we use the Atrous Spatial Pyramid Pooling (ASPP) method [4] in the last layer. Finally, we apply an up-sampling layer along with the softmax output to match the size of the input image.

Discriminator network. We use the structure similar to [38] for the discriminator network. It consists of 5 convolution layers with 4×4 kernel and $\{64, 128, 256, 512, 1\}$ channels in the stride of 2. Each convolution layer is followed by a Leaky-ReLU [30] parameterized by 0.2 except the last layer. To transform the model into a fully convolutional network, an up-sampling layer is added to the last layer to rescale the output to the size of the input map. Note that we do not employ any batch-normalization layer [18] as it only performs well when the batch size is sufficiently large.

4.2 Loss Function

Given an input image $\mathbf{X}_n$ of size $H \times W \times 3$, we denote the segmentation network by $S(\cdot)$ and the predicted probability map by $S(\mathbf{X}_n)$ of size $H \times W \times C$ where C is the category number. We denote the fully convolutional discriminator by $D(\cdot)$ which takes a probability map of size $H \times W \times C$ and outputs a confidence map of size $H \times W \times 1$. In the proposed method, there are two possible inputs to the discriminator network: segmentation prediction $S(\mathbf{X}_n)$ or one-hot encoded ground truth vector $\mathbf{Y}_n$.

Discriminator network. To train the discriminator network, we minimize the spatial cross-entropy loss $\mathcal{L}_D$ with respect to two classes using:

$$\mathcal{L}_D = -\sum_{h,w} (1-y_n) \log(1 - D(S(\mathbf{X}_n))^{(h,w)}) + y_n \log(D(\mathbf{Y}_n)^{(h,w)}), \tag{1}$$

where $y_n = 0$ if the sample is drawn from the segmentation network, and $y_n = 1$ if the sample is from the ground truth label. In addition, $D(S(\mathbf{X}_n))^{(h,w)}$ is the confidence map of $\mathbf{X}$ at location (h,w), and $D(\mathbf{Y}_n)^{(h,w)}$ is defined similarly. To convert the ground truth label map with discrete labels to a C-channel probability map, we use the one-hot encoding scheme on the ground truth label map where $\mathbf{Y}_n^{(h,w,c)}$ takes value 1 if pixel $\mathbf{X}_n^{(h,w)}$ belongs to class c, and 0 otherwise.

One potential issue with the discriminator network is that it may easily distinguish whether the probability maps come from the ground truth by detecting the one-hot probability [29]. However, we do not encounter this problem during the training phase. One reason is that we use a fully-convolutional scheme to predict spatial confidence, which increases the difficulty of learning the discriminator. In addition, we evaluate the *Scale* scheme [29] where the ground truth probability channel is slightly diffused to other channels according to the distribution of segmentation network output. However, the results show no difference, and thus we do not adopt this scheme in this work.

Segmentation network. We train the segmentation network by minimizing a multi-task loss function:

$$\mathcal{L}_{seg} = \mathcal{L}_{ce} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{semi}\mathcal{L}_{semi}, \tag{2}$$

where $\mathcal{L}_{ce}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{semi}$ denote the spatial multi-class cross entropy loss, adversarial loss, and semi-supervised loss, respectively. In (2), λ_{adv} and λ_{semi} are two weights for minimizing the proposed multi-task loss function.

We first consider the scenario of using annotated data. Given an input image $\mathbf{X}_n$, its one-hot encoded ground truth $\mathbf{Y}_n$ and prediction result $S(\mathbf{X}_n)$, the cross-entropy loss is obtained by:

$$\mathcal{L}_{ce} = -\sum_{h,w}\sum_{c\in C} \mathbf{Y}_n^{(h,w,c)} \log(S(\mathbf{X}_n)^{(h,w,c)}). \tag{3}$$

We use the adversarial learning process through the loss $\mathcal{L}_{adv}$ given a fully convolutional discriminator network $D(\cdot)$:

$$\mathcal{L}_{adv} = -\sum_{h,w} \log(D(S(\mathbf{X}_n))^{(h,w)}). \tag{4}$$

With this loss, we train the segmentation network to fool the discriminator by maximizing the probability of the predicted results being generated from the ground truth distribution.

Training with unlabeled data. In this work, we consider the adversarial training under the semi-supervised setting. For unlabeled data, we do not apply $\mathcal{L}_{ce}$ since there is no ground truth annotation. The adversarial loss $\mathcal{L}_{adv}$ is still applicable as it only requires the discriminator network. However, we find that it is crucial to choose a smaller λ_{adv} than the one used for labeled data. It is because the adversarial loss may over-correct the prediction to fit the ground truth distribution without the cross entropy loss.

In addition, we use the trained discriminator with unlabeled data within a self-taught learning framework. The main idea is that the trained discriminator can generate a confidence map $D(S(\mathbf{X}_n))$ which can be used to infer the regions sufficiently close to those from the ground truth distribution. We then binarize this confidence map with a threshold to highlight the trustworthy region. Furthermore, the self-taught, one-hot encoded ground truth $\hat{\mathbf{Y}}_n$ is element-wise set with $\hat{\mathbf{Y}}_n^{(h,w,c^*)} = 1$ if $c^* = \arg\max_c S(\mathbf{X}_n)^{(h,w,c)}$. The resulting semi-supervised loss is defined by:

$$\mathcal{L}_{semi} = -\sum_{h,w}\sum_{c\in C} I(D(S(\mathbf{X}_n))^{(h,w)} > T_{semi}) \cdot \hat{\mathbf{Y}}_n^{(h,w,c)} \log(S(\mathbf{X}_n)^{(h,w,c)}), \tag{5}$$

where $I(\cdot)$ is the indicator function and T_{semi} is the threshold to control the sensitivity of the self-taught process. Note that during training we treat both the self-taught target $\hat{\mathbf{Y}}_n$ and the value of indicator function as constant, and thus (5) can be simply viewed as a masked spatial cross entropy loss. In practice, we find that this strategy works robustly with T_{semi} ranging between 0.1 and 0.3.

5 Experimental Results

Implementation details. We implement the proposed algorithm using the PyTorch framework. We train the proposed model on a single TitanX GPU with 12 GB memory. To train the segmentation network, we use the Stochastic Gradient Descent (SGD) optimization method, where the momentum is 0.9, and the weight decay is 10^{-4}. The initial learning rate is set as

Table 1: Results on the VOC 2012 *val* set.

Methods	Data Amount 1/8	1/4	1/2	Full
FCN-8s [28]	N/A	N/A	N/A	67.2
Dilation10 [43]	N/A	N/A	N/A	73.9
DeepLab-v2 [4]	N/A	N/A	N/A	77.7
our baseline	66.0	68.3	69.8	73.6
baseline + $\mathcal{L}_{adv}$	67.6	71.0	72.6	74.9
baseline + $\mathcal{L}_{adv}$ + $\mathcal{L}_{semi}$	69.5	72.1	73.8	N/A

Table 2: Results on the Cityscapes *val* set.

Methods	Data Amount 1/8	1/4	1/2	Full
FCN-8s [28]	N/A	N/A	N/A	65.3
Dilation10 [43]	N/A	N/A	N/A	67.1
DeepLab-v2 [4]	N/A	N/A	N/A	70.4
our baseline	55.5	59.9	64.1	66.4
baseline + $\mathcal{L}_{adv}$	57.1	61.8	64.6	67.7
baseline + $\mathcal{L}_{adv}$ + $\mathcal{L}_{semi}$	58.8	62.3	65.7	N/A

2.5×10^{-4} and is decreased with polynomial decay with power of 0.9 as mentioned in [4]. For training the discriminator, we adopt Adam optimizer [20] with the learning rate 10^{-4} and the same polynomial decay as the segmentation network. For the hyper-parameters in the proposed method, λ_{adv} is set as 0.01 and 0.001 when training with labeled and unlabeled data, respectively. We set λ_{semi} as 0.1 and T_{semi} as 0.2.

For semi-supervised training, we randomly interleave labeled and unlabeled data while applying the training scheme described in Section 4.2. Note that, to prevent the model suffering from initial noisy masks and predictions, we start the semi-supervised learning after training for 5000 iterations with labeled data. We update both the segmentation network and discriminator network jointly. In each iteration, only the batch containing the ground truth data are used for training the discriminator. When randomly sampling partial labeled and unlabeled data from the datasets, we average several experiment results with different random seeds to ensure the evaluation robustness. The code and model are available at https://github.com/hfslyc/AdvSemiSeg.

Evaluation datasets and metric. In this work, we conduct experiments on two semantic segmentation datasets: PASCAL VOC 2012 [9] and Cityscapes [5]. On both datasets, we use the mean intersection-over-union (mean IU) as the evaluation metric.

The PASCAL VOC 2012 dataset comprises 20 common objects with annotations on daily captured photos. In addition, we utilize the extra annotated images from the Segmentation Boundaries Dataset (SBD) [11] and obtain a set of total 10,582 training images. We evaluate our models on the standard validation set with 1,449 images. During training, we use the random scaling and cropping operations with size 321×321. We train each model on the PASCAL VOC dataset for 20k iterations with batch size 10.

The Cityscapes dataset contains 50 videos in driving scenes from which 2975, 500, 1525 images are extracted and annotated with 19 classes for training, validation, and testing, respectively. Each annotated frame is the 20-th frame in a 30-frames snippet, where only these images with annotations are considered in the training process. We resize the input image to 512×1024 without any random cropping/scaling. We train each model on the Cityscapes dataset for 40k iterations with batch size 2.

PASCAL VOC 2012. Table 1 shows the evaluation results on the PASCAL VOC 2012 dataset. To validate the semi-supervising scheme, we randomly sample 1/8, 1/4, 1/2 images as labeled data, and use the rest of training images as unlabeled data. We compare the proposed algorithm against the FCN [28], Dilation10 [43], and DeepLab-v2 [4] methods. to demonstrate that our baseline model performs comparably with the state-of-the-art schemes. Note that our baseline model is equivalent to the DeepLab-v2 model without multi-scale fusion. The adversarial loss brings consistent performance improvement (from 1.6%*to*2.8%) over different amounts of training data. Incorporating the proposed semi-supervised learning scheme brings overall 3.5% to 4.0% improvement. Figure 2 shows visual comparisons of the segmentation results generated by the proposed method. We observe that the segmentation boundary achieves significant improvement when compared to the baseline model.

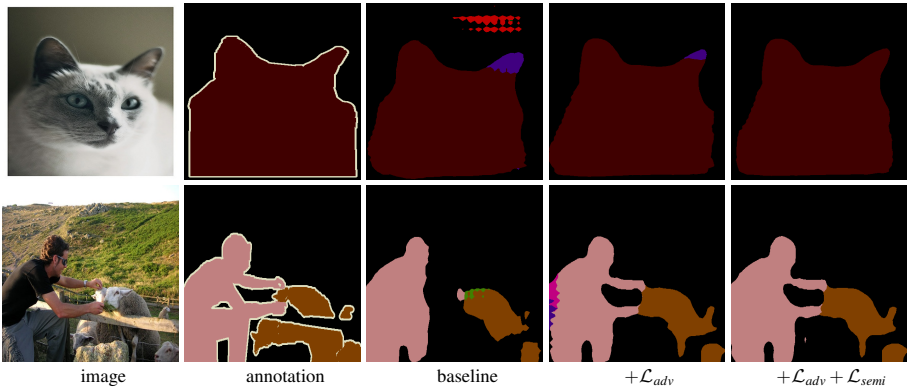

Figure 2: Comparisons on the PASCAL VOC 2012 dataset using 1/2 labeled data.

Table 3: Adversarial learning comparison with [29] on the VOC 2012 *validation* set.

	Baseline	Adversarial
[29]	71.8	72.0
ours	73.6	74.9

Table 4: Semi-supervised learning comparisons on the VOC 2012 *validation* set without using additional labels of the SBD.

	Data Amount	Fully-supervised	Semi-supervised
[33]	Full	62.5	64.6
[40]	Full	59.5	64.1
ours	Full	66.3	68.4
[40]	30%	38.9	42.2
ours	30%	57.4	60.6

Cityscapes. Table 2 shows evaluation results on the Cityscapes dataset. By applying the adversarial loss $\mathcal{L}_{adv}$, the model achieves 0.5% to 1.9% gain over the baseline model under the semi-supervised setting. This shows that our adversarial training scheme can encourage the segmentation network to learn the structural information from the ground truth distribution. Combining with the adversarial learning and the proposed semi-supervised scheme, the proposed algorithm achieves the performance gain of 1.6% to 3.3%.

Comparisons with state-of-the-art methods. Table 3 shows comparisons with [29] which utilizes adversarial learning for segmentation. There are major differences between [29] and our method in the adversarial learning processes. First, we design a universal discriminator for various segmentation tasks, while [29] utilizes one network structures for each dataset. Second, our discriminator does not require RGB images as additional inputs but directly operates on the prediction maps from the segmentation network. Table 3 shows that our method achieves 1.2% gain in mean IU over the method in [29].

We present the results under the semi-supervised setting in Table 4. To compare with [33] and [40], our model is trained on the original PASCAL VOC 2012 train set (1,464 images) and use the SBD [11] set as unlabeled data. It is worth noticing that in [33], image-level labels are available for the SBD [11] set, and in [40] additional unlabeled images are generated through their generator during the training stage.

Hyper-parameter analysis. The proposed algorithm is governed by three hyper parameters: λ_{adv} and λ_{semi} for balancing the multi-task learning in (2), and T_{semi} used to control the sensitivity in the semi-supervised learning described in (5). Table 5 shows sensitivity analysis on hyper parameters using the PASCAL VOC dataset under the semi-supervised setting. More analysis and results are provided in the supplementary material.

We first show comparisons of different values of λ_{semi} with 1/8 amount of data under the semi-supervised setting. We set $\lambda_{adv} = 0.01$ and $T_{semi} = 0.2$ for the comparisons. Overall, the

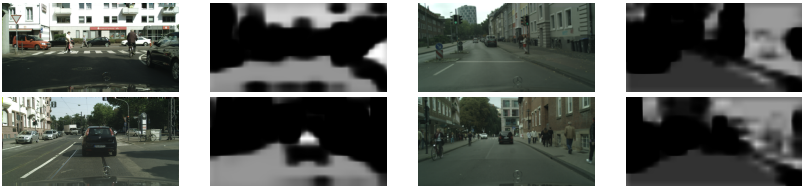

Figure 3: Visualization of the confidence maps. Given the prediction results generated by the segmentation network, the confidence maps are obtained from the discriminator. In the confidence maps, the brighter regions indicate that they are closer to the ground truth distribution, and we utilize these brighter regions for semi-supervised learning.

Table 5: Hyper parameter analysis.

Data Amount	λ_{adv}	λ_{semi}	T_{semi}	Mean IU
1/8	0.01	0	N/A	67.6
1/8	0.01	0.05	0.2	68.4
1/8	0.01	0.1	0.2	69.5
1/8	0.01	0.2	0.2	69.1
1/8	0.01	0.1	0	67.2
1/8	0.01	0.1	0.1	68.8
1/8	0.01	0.1	0.2	69.5
1/8	0.01	0.1	0.3	69.2
1/8	0.01	0.1	1.0	67.6

Table 6: Ablation study of the proposed method on the PASCAL VOC dataset.

$\mathcal{L}_{adv}$	$\mathcal{L}_{semi}$	FCD	Data Amount 1/8	Data Amount Full
			66.0	73.6
✓		✓	67.6	74.9
✓			66.6	74.0
	✓	✓	65.7	N/A
✓	✓	✓	69.5	N/A

proposed method achieves the best mean IU of 69.5% with 1.9% gain. when λ_{semi} is set to 0.1. Second, we perform the experiments with different values of T_{semi} by setting $\lambda_{adv} = 0.01$ and $\lambda_{semi} = 0.1$. With higher T_{semi}, the proposed model only trusts regions of high structural similarity as the ground truth distribution. Overall, the proposed model achieves the best results when $T_{semi} = 0.2$ and performs well for a wide range of T_{semi} (0.1 to 0.3). When $T_{semi} = 0$, we trust all the pixel predictions in unlabeled images, which results in performance degradation. Figure 3 shows sample confidence maps from the predicted probability maps.

Ablation study. We present the ablation study of our proposed system in Table 6 on the PASCAL VOC dataset. First, we examine the effect of using fully convolutional discriminator (FCD). To construct a discriminator that is not fully-convolutional, we replace the last convolution layer of the discriminator with a fully-connected layer that outputs a single neuron as in typical GAN models. Without using FCD, the performance drops 1.0% and 0.9% with all and one-eighth data, respectively. This shows that the use of FCD is essential to adversarial learning. Second, we apply the semi-supervised learning method without the adversarial loss. The results show that the adversarial training on the labeled data is important to our semi-supervised scheme. If the segmentation network does not seek to fool the discriminator, the confidence maps generated by the discriminator would be meaningless, providing weaker supervisory signals.

6 Conclusions

In this work, we propose an adversarial learning scheme for semi-supervised semantic segmentation. We train a discriminator network to enhance the segmentation network with both labeled and unlabeled data. With labeled data, the adversarial loss for the segmentation network is designed to learn higher order structural information without post-processing. For unlabeled data, the confidence maps generated by the discriminator network act as the self-taught signal for refining the segmentation network. Extensive experiments on the PASCAL VOC 2012 and Cityscapes datasets validate the effectiveness of the proposed algorithm.

Acknowledgments. W.-C. Hung is supported in part by the NSF CAREER Grant #1149783, gifts from Adobe and NVIDIA.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. WhatâĂŹs the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 2, 3

[3] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 3

[4] 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. In *TPAMI*, 2017. 2, 4, 5, 7

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2, 7

[6] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 2, 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[8] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015. 3

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *IJCV*, 2010. 1, 2, 7

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 3

[11] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 7, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[13] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. In *arXiv preprint arXiv:1612.02649*, 2016. 3

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *ICML*, 2018. 3

[15] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NIPS*, 2015. 2, 3

[16] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017. 3

[17] Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Scene parsing with global context embedding. In *ICCV*, 2017. 1

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[19] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 3

[20] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*, 2014. 7

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[22] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *CVPR*, 2017. 3

[23] Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. In *NIPS*, 2017. 3

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *arXiv preprint arXiv:1609.04802*, 2016. 3

[25] Guosheng Lin, Chunhua Shen, Anton van dan Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 1, 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[27] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen Change Loy, and Xiaoou Tang. Semantic Image Segmentation via Deep Parsing Network. In *ICCV*, 2015. 1

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 3, 4, 7

[29] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. In *NIPS Workshop on Adversarial Training*, 2016. 3, 5, 8

[30] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 5

[31] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv preprint arXiv:1611.04076*, 2016. 3

[32] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *CVPR*, 2014. 1

[33] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015. 2, 3, 8

[34] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 2, 3

[35] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015. 3

[36] Pedro O Pinheiro and Ronan Collobert. Weakly supervised semantic segmentation with convolutional networks. In *CVPR*, 2015. 2, 3

[37] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016. 2, 3

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 3, 5

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[40] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi and weakly supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017. 3, 8

[41] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 3

[42] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*, 2017. 3

[43] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1, 4, 5, 7

[44] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 1, 2

[45] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. In *CVPR*, 2017. 1