Robust Semi-Supervised Semantic Segmentation Based on Self-Attention and Spectral Normalization

Jia Zhang[1], Zhixin Li[1,*], Canlong Zhang[1], Huifang Ma[2]
[1]*Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China*
[2] *College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China*
[*]Corresponding Author. E-mail: lizx@gxnu.edu.cn

***Abstract*—The application of adversarial learning for semi-supervised semantic image segmentation based on convolutional neural networks can effectively reduce the number of manually generated labels required in the training process. However, the convolution operator of the generator in the generative adversarial network (GAN) has a local receptive field, so that the long-range dependencies between different image regions can only be modeled after passing through multiple convolutional layers. The present work addresses this issue by introducing a self-attention mechanism in the generator of the GAN to effectively account for relationships between widely separated spatial regions of the input image with supervision based on pixel-level ground truth data. In addition, the adjustment of the discriminator has been demonstrated to affect the stability of GAN training performance. This is addressed by applying spectral normalization to the GAN discriminator during the training process. The proposed stable self-attention adversarial learning semi-supervised semantic image segmentation network is demonstrated to provide superior image segmentation performance compared with the results of current semi-supervised and fully-supervised semantic image segmentation techniques.**

***Index Terms*—Self-Attention, Adversarial Learning, Semi-Supervised, Spectral Normalization**

I. Introduction

Image segmentation is an essential task in the fields of image processing and computer vision. Such tasks conducted at the pixel level are generally referred to as semantic image segmentation, which has seen significant progress in recent years due to the development of convolutional neural networks (CNNs) [1]–[7]. This progress has led to the wide application of semantic segmentation to various fields such as autonomous driving [8] and image editing [9]. However, practical applications of semantic segmentation generally require a large number of object classes and ground truth label data annotated at the pixel level for each class to convey the relationships between object boundaries and their components when conducting fully-supervised CNN training, and these data are generally derived manually at great effort and expense. Efforts have been made to reduce the number of manually generated labels required in the training process, the most common of which has been the application of semi-supervised and weakly-supervised training methods to semantic image segmentation.

The key aspect of semi-supervised learning methods is that they employ weakly tagged data indicating only the existence of objects of a certain class, but provide no ground truth information about the location or boundaries of objects. Naturally, these annotations are weaker than pixel-level labels, but are readily available in a large volume of visual data or can be obtained manually at relatively low cost. Therefore, semi-supervised learning provides an attractive method for training semantic image segmentation models with a limited volume of labeled training data and a large volume of unlabeled data. Various semi-supervised training methods have been proposed for semantic image segmentation. For example, Kalluri et al. [10] combined semi-supervised learning with unsupervised domain adaptation. Stekovic et al. [11] implemented geometric constraints between multiple views of a three-dimensional (3D) scene. Consistency regularization [4], which represents a class of semi-supervised learning algorithms for training deep neural network classifiers, has also been developed to produce state-of-the-art semi-supervised classification results that are conceptually simple and often easy to implement. Finally, we note that the use of image level annotation in semantic image segmentation relies on location maps obtained by a classification network to bridge the gap between image-level and pixel-level annotations [12]. However, these maps focus only on a small part of objects without an accurate representation of their boundaries. As a result, currently available image segmentation methods adopting semi-supervised learning have been demonstrated to produce poor results relative to fully supervised network training methods.

The development of generative adversarial networks (GANs) [13] has led to significant progress in the application of semi-supervised and weakly-supervised learning to semantic image segmentation. This may prevent the learning of long-range dependencies because the optimization algorithm may not find the parameter values that appropriately coordinate the multiple convolutional layers to capture these dependencies. While increasing the size of the convolution kernel can increase the representation capabilities of the network, this also results in a loss of computational and statistical efficiency through the use of local convolutional structures. Contextual dependencies have been addressed in numerous manners. For example, learning contexts have been demonstrated to rely on local features and also contribute to feature representation. A recurrent neural network (RNN) was employed in DAG-RNN [14] to create a directed acyclic graph model for capturing rich contextual dependencies. Pixel-level relationships were captured by PSANet [15] through relative position information in convolutional and spatial dimensions. In addition, EncNet [16] introduced a channel attention mechanism to capture the global context. The capability of attention modules for modeling long-range dependencies has been demonstrated, and these modules have been widely used in many tasks [2], [17]–[19]. At the same time, it is noted that the application of attention modules has become increasingly extensive in the field of computer vision. For example, a self-attention mechanism was applied for training better classification generators [19]. However, these efforts have not been effectively applied to semi-supervised semantic segmentation methods.

The performance control of the discriminator is another continuing challenge in GAN training. Here, the density ratio estimation by the discriminator in high dimensional space is often inaccurate and unstable during training. Optimization seeks to obtain a discriminator that can well distinguish the model distribution from the target distribution [20]. However, once such a discriminator is obtained, the training of the generator is completely stopped. Various efforts have been made to improve the stability of GAN training. For example, the DCGAN model [21] was developed to find a set of better network architecture settings from the perspective of architectural design, and has carried out rich experimental verification in the field of image generation. The Wasserstein GAN model [22] was developed to solve the GAN training instability problem from a theoretical point of view by introducing the concept of Wasserstein distance. The Wasserstein GAN model proposed that the parameter matrix of the discriminator must satisfy Lipschitz constraints. However, the constraint method adopted is relatively simple and crude, and directly restricts the elements in the parameter matrix so that it is not larger than a given value. While this method can guarantee Lipschitz constraints, it also destroys the proportionality relationships between the parameters.

The present study addresses the above-discussed challenges by proposing a stable semi-supervised adversarial learning method for semantic image segmentation. The basis of the segmentation network employed in the present work is the DeepLabv2 [1] framework and the ResNet-101 model pre-trained on the MSCOCO dataset [23]. The presented work makes the following contributions.

- A self-attention mechanism is introduced in the segmentation network of the GAN to directly capture long-range dependencies by calculating the interaction between any two positions of the feature map during semi-supervised adversarial training with supervision based on pixel-level ground truth data.
- Spectral normalization [24] is applied to stabilize the training of the discriminator network. The proposed algorithm requires no intensive adjustment of the hyper-parameters to achieve satisfactory discriminator training performance. Moreover, this new normalization technique is less computationally intensive and easy to integrate into existing implementations than other commonly employed methods.
- The proposed stable self-attention adversarial learning semi-supervised semantic image segmentation network is demonstrated to provide superior image segmentation performance compared with the results of current semi-supervised and fully-supervised semantic image segmentation techniques based on applications to the PASCAL VOC 2012 [25] and Cityscapes [26] datasets. The compared semi-supervised method includes the AdvSemiSeg method proposed by Hung et al. [27], while the fully supervised method is the DeepLabv2 [1] network, which is trained only for specific image regions of the datasets. In addition, we consider the performance of the proposed network without the application of spectral normalization (i.e., only the self-attention modules are applied).

The remainder of this article is organized as follows. Section 2 reviews related work, and the proposed framework is described in detail in Sections 3 and 4. The results of the empirical assessment are presented in Section 5, and a summary is given in Section 6.

II. Related Work

The application of deep CNNs for semantic image segmentation tasks over the past few years has achieved excellent results. Many segmentation methods have employed migration learning, particularly using VGG neural networks [28] with the convolutional layer of the ResNet [29] classification network as the backbone. The effectiveness of deep neural networks in semantic image segmentation has been demonstrated by the application of the full convolutional network (FCN) introduced by Long et al. [3], which linked the convolutional 21-way classifier to the VGG-16 backbone network. Chen et al. [30] applied atrous convolution in the later layers of the VGG-16 network to improve the spatial resolution of the prediction while maintaining the receive domain. Encoder-decoder networks [31] have been newly applied to semantic image segmentation. The encoder is a special neural network used for feature extraction and data dimensionality reduction to generate feature images with semantic information. The role of the decoder network is to map the low resolution feature image output by the encoder network back to the size of the input image for pixel-by-pixel classification. Finally, U-nets [32] uses a permutation convolutional layer to increase resolution, and its jump connection carries a complete feature map.

Goodfellow et al. [13] reintroduced the concept of adversarial learning for conducting image generation tasks, and employed GANs to successfully generate images such as handwritten numbers and faces from random noise. However, random noise and meaningful images apparently derive from different data domains, and the distribution is inconsistent. Accordingly, the GAN model can solve the problem of inconsistent distributions among different data domains.

III. Model Overview

As shown in Figure 1, the primary modules employed in the proposed framework include a segmentation network G and a discriminator D. Given a color input image X_n of dimensions $H \times W \times 3$, the feature map in G models the long-range, multi-level dependencies of the image region through the introduced two-layer self-attention modules. This enables the generator to model the rich contextual relationships based on local features, so that the details of each location and the details of the far ends are well coordinated when generating images. The output probability is a class probability map of dimensions $H \times W \times C$, where C is the number of semantic categories.

The discriminator network is based on the FCN, which accepts inputs of different sizes. The input of D is either the class probability map output by G (i.e., G(Xn)) or a thermally encoded ground truth label map I_n, and the final output is a confidence map of dimensions $H \times W \times 1$. Here, each pixel i of the confidence map is set to 0 if it derives from G, and is set to 1 if it derives from the ground truth label map. Accordingly, the confidence map represents regions of the probabilistic prediction output of G that are closer to the ground truth label distribution. Spectral normalization is applied within D to ensure that its mapping function satisfies the Lipschitz constraint.

Semi-supervised training is conducted using both unlabeled and labeled images. Unlabeled data is applied for training G throughout the training process, while our self-attention modules effectively account for the relationships between widely separated spatial regions of the input image. When using labeled data, the training of G is simultaneously supervised according to the spatial multi-class cross entropy loss $\mathcal{L}_{ce}$ based on I_n and the adversarial loss $\mathcal{L}_{adv}$ obtained from D. Then, the confidence map is employed as a supervisory signal for training G in conjunction with a masked cross entropy loss $\mathcal{L}_{semi}$ in a self-learning manner based on the trusted predictions given in the confidence map.

IV. Semi-supervised Training

A. Loss Functionst

The segmentation network is trained by minimizing the following multi-task loss function.

$$\mathcal{L}_G = \mathcal{L}_{ce} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{semi}\mathcal{L}_{semi} \tag{1}$$

Where, λ_{adv} and λ_{semi} are the weights applied to adjust the significance of the corresponding loss function components in the minimization process, and thereby minimize multitasking losses. The first loss component in (1) is defined as follows.

$$\mathcal{L}_{ce} = -\sum_{h,w}\sum_{c\in C} \mathbf{I}_n^{(h,w,c)} \log\left(G\left(\mathbf{X}_n\right)^{(h,w,c)}\right) \tag{2}$$

Here, we convert the discrete labels of I_n into a C-channel probability map by applying a hot coding scheme, where the map entry $\mathbf{I}_n^{(h,w,c)}$ for the $(h,w)^{th}$ pixel and the c^{th} class is assigned a value of 1 if the pixel $\mathbf{X}_n^{(h,w,c)}$ belongs to class c, and is otherwise assigned a value of 0. The second loss component is defined as follows.

$$\mathcal{L}_{adv} = -\sum_{h,w} \log\left(D\left(G\left(\mathbf{X}_n\right)\right)^{(h,w)}\right) \tag{3}$$

Here, $D\left(G\left(\mathbf{X}_n\right)\right)^{(h,w)}$ is the confidence map value of X_n for the $(h,w)^{th}$ pixel. We note that unlabeled data incurs no loss associated with $\mathcal{L}_{ce}$ because unlabeled data includes no ground truth annotation. Therefore, only $\mathcal{L}_{adv}$ is applicable under these training circumstances. Finally, we define the third loss component in (1) using an index function $F(\cdot)$ and a threshold $\mathcal{T}_{semi}$ to binarize the confidence map, such that the trusted area can be better displayed. This is given as follows.

$$\begin{aligned}\mathcal{L}_{semi} = & -\sum_{h,w}\sum_{c\in C} F\left(D\left(G\left(\mathbf{X}_n\right)\right)^{(h,w)} > T_{semi}\right) \\ & \cdot \quad \hat{\mathbf{I}}_n^{(h,w,c)} \log\left(G\left(\mathbf{X}_n\right)^{(h,w,c)}\right)\end{aligned} \tag{4}$$

Here, the self-taught, one-hot encoded ground truth $\hat{\mathbf{I}}_n$ is an element-wise set with $\hat{\mathbf{I}}_n^{(h,w,e^*)} = 1$ if $c^* = \arg\max_c G\left(\mathbf{X}_n\right)^{(h,w,c)}$. In the training process, the product of the self-learning target $\hat{\mathbf{I}}_n$ and the value of $F(\cdot)$ is regarded as a constant. Experiments have demonstrated that a value of $T_{semi} = 0.2$ yields good robustness in the training process.

The discriminator network is trained by minimizing the spatial cross-loss function $\mathcal{L}_D$, which is given as follows.

$$\begin{aligned}\mathcal{L}_D = & -\sum_{h,w} \left(1 - y_n\right) \log\left(1 - D\left(G\left(\mathbf{X}_n\right)\right)^{(h,w)}\right) \\ & + \quad y_n \log\left(D\left(\mathbf{I}_n\right)^{(h,w)}\right)\end{aligned} \tag{5}$$

Here, $y_n = 0$ if the discriminator input is $G\left(\mathbf{X}_n\right)$ and $y_n = 1$ if the discriminator input is I_n, while $\mathrm{D}\left(\mathbf{I}_n\right)^{(h,w)}$ is the confidence map value of I_n for the $(h,w)^{th}$ pixel.

B. Self-attention Module

The framework of the proposed two-layer self-attention module is illustrated in Figure 2. The module takes the local feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ of the previous layer as its input, and generates two feature maps $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{H\times W\times C}$, performs matrix multiplication after transposing Q and K, and calculates the attention map $\mathbf{S} \in \mathbb{R}^{N\times N}$ with a softmax layer, where $N = H \times W$ is the number of pixels. Here, the elements of S provide the following measure of the dependency of the i^{th} pixel on the j^{th} pixel.

$$S_{ji} = \frac{\exp\left(Q_i \cdot K_j\right)}{\sum_{i=1}^{N} \exp\left(Q_i \cdot K_j\right)} \tag{6}$$

Additional non-local features are determined by adding a convolution map to X to obtain a new feature map $\mathbf{V} \in \mathbb{R}^{H\times W\times C}$, and S and V are subjected to matrix multiplication after transposition. Then, the product is multiplied by a proportional parameter α to form an attention-weighted feature map, which is added to the original feature map X as follows.

$$O_j = \alpha \sum_{i=1}^{N} \left(s_{ji} V_i\right) + X_j \tag{7}$$

We initialize α as 0, and α is adjusted to assign more weight to non-local features in a self-learning manner. Thus, the final self-attention feature map O is the weighted sum of the features at all locations and the original features. This models

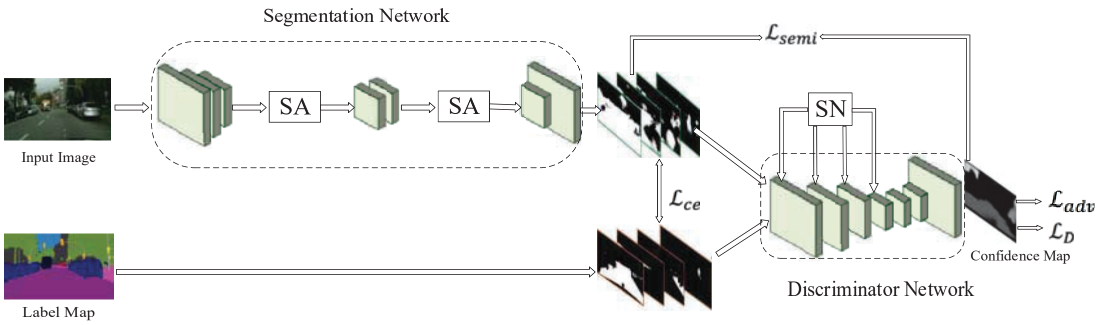

Fig. 1. An overview of the proposed system for semi-supervised semantic image segmentation, where the segmentation network G outputs a class probability map, SA represents the self-attention modules, SN represents the application of the spectral normalization technique, the discriminator network D outputs a confidence map, $\mathcal{L}_{ce}$ is the spatial multi-class cross entropy loss based on the ground truth label map, $\mathcal{L}_{adv}$ is the adversarial loss of D, and $\mathcal{L}_{semi}$ is the masked cross entropy loss.

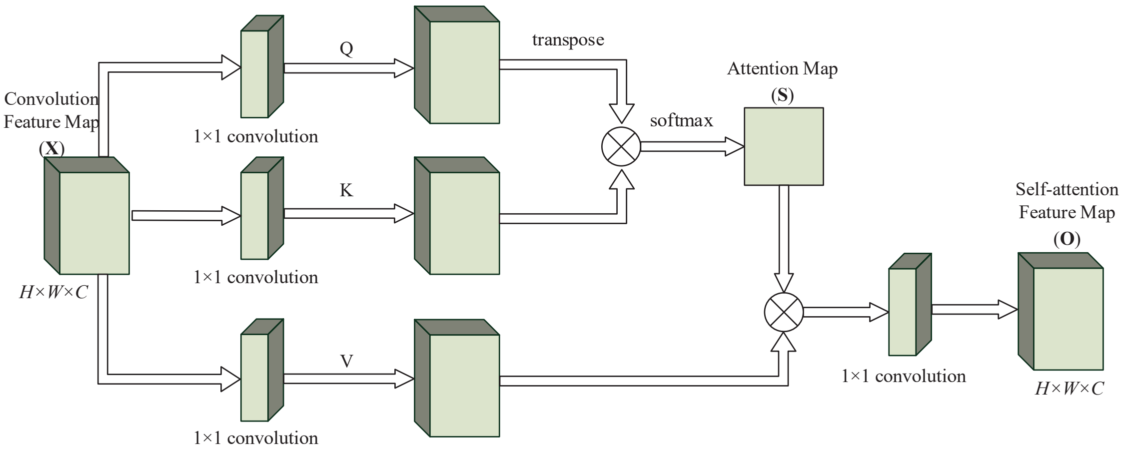

Fig. 2. Schematic illustrating the two-layer self-attention module framework. Here, the symbol $\otimes$ represents matrix multiplication, which is conducted after transposing the feature maps Q, K, V, and S, as indicated by the transposition symbol T.

the long-range semantic dependencies between feature maps, and thereby helps to improve the distinguishability of features.

C. Spectral Normalization

The optimal form of D for a standard GAN is determined as follows.

$$D_G^*(x)=\frac{q_{\text{data}}\left(x\right)}{q_{\text{data}}\left(x\right)+p_G(x)}=\operatorname{sigmoid}\left(f^*(x)\right) \tag{8}$$

Here, q_{data} is the distribution of data x, p_G is the distribution of the generated model corresponding to x that is learned through the confrontational minimum-maximum optimization process, and $f^*(x)=\log q_{\text{data}}\left(x\right)-\log p_G(x)$, where its derivative is given as follows.

$$\nabla_x f^*(x)=\frac{1}{q_{\text{data}}\left(x\right)}\nabla_{\boldsymbol{x}} q_{\text{data}}\left(\boldsymbol{x}\right)-\frac{1}{p_G(\boldsymbol{x})}\nabla_{\boldsymbol{x}} p_G(\boldsymbol{x}) \tag{9}$$

However, this derivative term is unbounded or even incalculable, and regular restrictions must be added in practice. Therefore, a mechanism is needed to define the derivative of $f^*(x)$. We first note that, if the bias added to each layer of D is omitted, the upper bound of $f^*(x)$ can be determined according to the following inequality.

$$\begin{aligned}\|f\|_{\text{Lip}} \quad &\leq \quad \left\|\left(\boldsymbol{h}_L\mapsto W^{L+1}\boldsymbol{h}_L\right)\right\|_{\text{Lip}}\\ &\cdot \quad \|a_L\|_{\text{Lip}}\cdot\left\|\left(\boldsymbol{h}_{L-1}\mapsto W^L\boldsymbol{h}_{L-1}\right)\right\|_{\text{Lip}}\\ &\cdots \quad \|a_1\|_{\text{Lip}}\cdot\left\|\left(\boldsymbol{h}_0\mapsto W^1\boldsymbol{h}_0\right)\right\|_{\text{Lip}}\\ &= \quad \prod_{l=1}^{L+1}\left\|\left(\boldsymbol{h}_{l-1}\mapsto W^l\boldsymbol{h}_{l-1}\right)\right\|_{\text{Lip}}=\prod_{l=1}^{L+1}\sigma\left(W^l\right)\end{aligned} \tag{10}$$

Here, $\|\cdot\|_{\text{Lip}}$ represents the Lipschitz norm, spectral normalization controls the Lipschitz constant of the discriminant function f by strictly constraining each layer g: $h_{\text{in}}\rightarrow h_{\text{out}}$, $\left\{\mathrm{W}^1,\ldots,\mathrm{W}^{\mathrm{L}},\mathrm{W}^{\mathrm{L+1}}\right\}$ is the learning parameters set, α_1 is an element-wise non-linear activation function, and $\sigma(W)$ represents the two-norm of W, which is regarded as a constant.

Using the linearity of properties, where, for any coefficient β and matrix A with the property $\|\beta\mathbf{A}\| = |\beta| \cdot \|\mathbf{A}\|$. Therefore, the upper bound of f is 1. Accordingly, the spectral normalization of matrix W is given as follows.

$$\bar{W}_{\mathrm{SN}}(W) := W/\sigma(W) \tag{11}$$

Then, applying the spectral normalization in (11) within the inequality (10) for each layer weight W of D ensures that D can be regarded as a function of the implicit function f, and its Lipschitz norm can be constrained to be less than 1, which achieves the required limiting effect during the training of the discriminator.

V. Experiments

A. Experimental Setup

The PASCAL VOC 2012 dataset contains 21 object classes. We included the labeled image segmentation boundary dataset (SBD) [33] to obtain a total of 10,582 training images. The testing set included 1,449 verified images. The Cityscapes dataset contains 50 videos in driving scenes from which 2975, 500, and 1525 images were extracted and annotated with 19 classes for training, validation, and testing, respectively. The average crossover (mean IU) was applied as the evaluation index for both datasets. The image segmentation performances of all trained networks were evaluated for both datasets with different proportions of labeled data, including 1/8, 1/4, 1/2, and the full set of labeled data, with the remainder being unlabeled data. Both unlabeled and labeled data were randomly extracted, and the same data were used for all networks.

Random scaling and cropping operations of size 321×321 were employed during the training process for the PASCAL VOC dataset. Each iteration was 20k times and the batch size was 8. For the Cityscapes dataset, we adjusted the size of the input images to 512×1024 without random clipping/scaling, and applied 40k times per iteration with a batch size of 2. For semi-supervised training, the training process began after 5,000 iterative training of labeled data to avoid the model being affected by the initial noise mask and prediction. The discriminator network and the segmentation network were trained jointly. In each iteration, only the batch containing the ground truth data are used for training the discriminator.

Training was conducted using a PyTorch toolbox operating on a Titan X GPU with 12 GB of memory. A stochastic gradient descent (SGD) optimizer was used with a momentum of 0.9 and a weight decay of 10^{-4}. The initial learning rate was set to 2.5×10^{-4}, which decreased according to a polynomial decay with a power of 0.9. The Adam optimizer [12] was used for training the discriminator, and the learning rate was set to 10^{-4}. When conducting training using both unlabeled and labeled data, λ_{adv} and λ_{semi} were set to 0.001 and 0.1, respectively, while T_{semi} was set to 0.2.

B. Result

Table 1 gives the results of the semi-supervised and full-supervised evaluations on the Cityscapes dataset. In the case of self-attention alone (i.e., Ours (SA)), the performance increased by 1.4% to 2.4% over that of the AdvSemiSeg model, which serves as a baseline, and the performance increased by 2.3% to 3.2% after adding spectral normalization (i.e., Ours (SA+AN). We observed a 3.2% increase in the segmentation performance for the 1/8 proportion of labeled data compared to that of the baseline. It is speculated that the two-stage GAN training is poor under the condition of low-labeled data, in which the discriminator is only updated according to the labeled samples. This reduces the amount of data seen during training, resulting in overfitting.

TABLE I

Image segmentation performance results (mean IU) for the Cityscapes dataset with different proportions of labeled data randomly sampled for training

Methods	Labeled Data			
	1/8	1/4	1/2	Full
FCN-8s [3]	-	-	-	65.3
Dilation10 [5]	-	-	-	67.1
CutMix [34]	63.4	65.2	67.7	-
DeepLabv2	55.5	59.9	64.1	66.4
AdvSemiSeg [27]	58.8	62.3	65.7	67.7
Ours(SA)	61.2	63.7	67.5	70.4
Ours(SA+SN)	62.0	64.6	68.3	71.1

TABLE II

Image segmentation performance results (mean IU) for the PASCAL VOC dataset with different proportions of labeled data randomly sampled for training

Methods	Labeled Data			
	1/8	1/4	1/2	Full
FCN-8s [3]	-	-	-	67.2
Dilation10 [5]	-	-	-	73.9
Mittal S et al. [35]	71.4	-	-	75.6
DeepLabv2	66.0	68.3	69.8	73.6
AdvSemiSeg [27]	69.5	72.1	73.8	74.9
Ours(SA)	70.3	72.7	75.9	78.4
Ours(SA+SN)	71.8	73.5	76.3	78.9

The mean IU performance results of the semi-supervised and fully-supervised networks for the PASCAL VOC dataset are presented in Table 2. In the case of self-attention alone, the performance increased by 0.6% to 2.1% over that of the AdvSemiSeg model, and the performance increased by 1.4% (for a 1/4 proportion of labeled data) to 2.5% (for a 1/2 proportion of labeled data) after adding spectral normalization. Figure 3 presents a comparison of the ground truth (GT) image segmentation data and the segmentation results obtained by the proposed network under different proportions of labeled data during training.

The mean IU performance results of the semi-supervised and fully-supervised networks for each class in the PASCAL VOC dataset with different proportions of labeled data are presented in Table 3. In addition, the mean IU values for all classes originally reported in Table 2 are included in the final column. We note from the results that the proposed self-attention modules and spectral normalization significantly

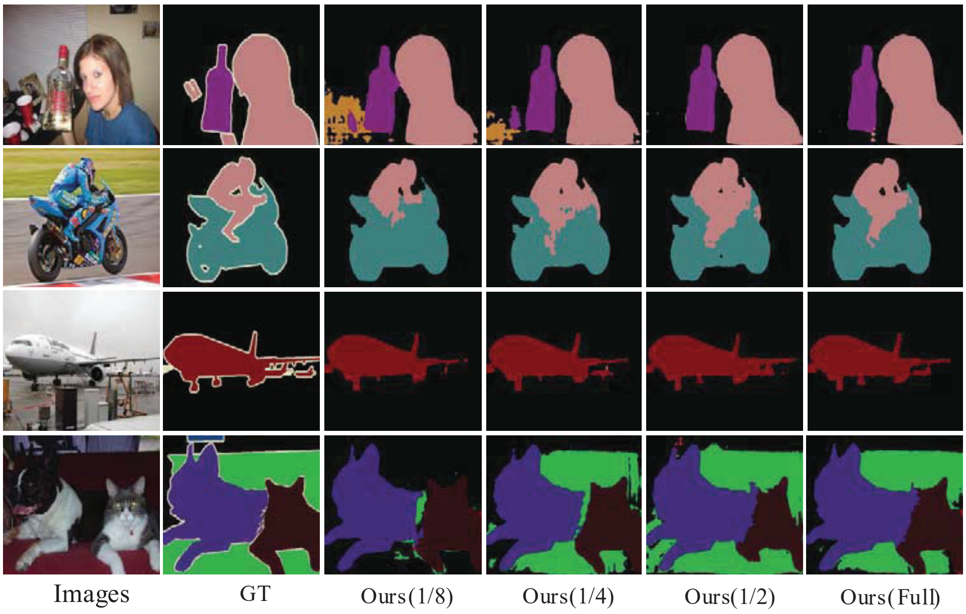

Fig. 3. Comparisons of the ground truth (GT) image segmentation data and the segmentation results obtained by the proposed network under different proportions of labeled data for selected images in the PASCAL VOC dataset. It can be seen that a better segmentation effect is achieved when 1/2 of the labeled data is randomly selected for training.

TABLE III
IMAGE SEGMENTATION PERFORMANCE RESULTS (MEAN IU) FOR THE PASCAL VOC DATASET WITH DIFFERENT PROPORTIONS OF LABELED DATA RANDOMLY SAMPLED FOR TRAINING

Models	*bkg*	*aero*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*mbk*	*prsn*	*plnt*	*sheep*	*sofa*	*train*	*tv*	*mIoU*
1/8 Labeled Data																						
AdvSemiSeg	0.93	0.82	0.40	0.82	0.63	0.71	0.90	**0.84**	0.84	0.30	0.72	0.40	0.76	0.73	**0.79**	0.83	0.52	0.75	0.43	0.79	0.71	69.5
Ours(SA)	0.93	0.85	0.41	0.85	0.67	0.74	0.88	0.82	0.86	0.27	**0.80**	0.42	0.78	**0.78**	0.77	0.83	0.42	0.77	0.42	0.78	**0.73**	70.3
Ours(SA+SN)	0.93	**0.87**	**0.42**	**0.85**	**0.67**	**0.76**	**0.91**	0.82	**0.87**	**0.32**	0.79	**0.45**	**0.80**	0.74	0.76	0.82	0.52	**0.82**	**0.45**	**0.80**	0.70	71.8
1/4 Labeled Data																						
AdvSemiSeg	0.93	0.88	0.40	0.85	0.67	0.76	0.91	0.83	0.86	0.31	0.79	0.45	0.79	0.75	0.78	0.84	**0.56**	0.79	**0.47**	0.79	0.74	72.1
Ours(SA)	0.94	0.87	**0.42**	0.86	0.69	**0.78**	0.92	0.84	0.89	0.32	0.76	**0.48**	**0.81**	0.74	0.80	**0.85**	0.53	0.78	0.44	**0.82**	0.72	72.7
Ours(SA+SN)	0.94	**0.89**	0.41	**0.87**	**0.70**	0.74	**0.93**	**0.85**	0.89	**0.35**	**0.82**	0.45	0.80	**0.83**	0.80	0.84	0.55	**0.80**	0.44	0.81	0.74	73.5
1/2 Labeled Data																						
AdvSemiSeg	0.94	0.88	0.41	0.86	0.67	0.79	0.91	0.85	0.87	0.34	0.81	0.52	0.80	0.80	**0.82**	0.84	0.57	0.82	0.47	0.81	0.73	73.8
Ours(SA)	0.94	0.88	0.42	**0.87**	0.67	0.82	**0.93**	**0.88**	0.89	**0.40**	0.86	0.59	0.84	**0.84**	0.81	0.86	0.60	0.81	**0.48**	0.84	0.75	75.9
Ours(SA+SN)	0.94	**0.89**	0.42	0.85	0.67	**0.83**	0.92	0.86	0.89	0.39	0.86	**0.61**	**0.85**	0.82	0.81	0.86	0.60	**0.88**	0.46	0.84	0.75	76.3
Full Labeled Data																						
AdvSemiSeg	0.94	0.89	0.41	0.87	0.67	0.81	0.91	0.85	0.88	0.36	0.83	0.53	0.82	0.80	0.83	0.85	0.59	0.83	0.49	0.83	0.74	74.9
Ours(SA)	0.94	0.89	0.43	**0.88**	0.73	0.82	**0.94**	**0.87**	0.90	0.41	0.85	**0.59**	0.86	0.83	**0.86**	0.87	**0.66**	0.87	**0.56**	**0.88**	0.74	78.4
Ours(SA+SN)	**0.95**	**0.90**	**0.43**	0.87	**0.75**	0.82	0.93	0.86	**0.91**	**0.42**	**0.87**	0.57	**0.86**	**0.86**	0.85	**0.90**	0.65	**0.89**	0.55	0.86	**0.77**	78.9

improved the segmentation performance for the 21 classes of the PASCAL VOC dataset. Clearly, the addition of the self-attention modules to the segmentation network has a good effect on capturing the long-range contextual information between any two pixels of the feature map, and thereby improves the feature representation of the model. In addition, adding spectral normalization to the discriminator is further beneficial for training the GAN network. This is further illustrated by the image segmentation results presented in Figure 4, which compare the GT image segmentation data and the segmentation results obtained by the AdvSemiSeg model and the proposed model when using 1/2 of the labeled data during training. The segmentation results obtained by the proposed model after introducing either self-attention or spectral normalization are both better qualitatively than those of the AdvSemiSeg model, and particularly after introducing both self-attention and spectral normalization components. From this we can see the effectiveness of the self-attention module for capturing the global dependence of the input and the spectrally normalized stable GAN.

The image segmentation results presented thus far were based on the DeepLabv2 framework and the ResNet-101 model pre-trained on the MSCOCO dataset. In addition, we substituted the DeepLabv2 framework with the DeepLabv3

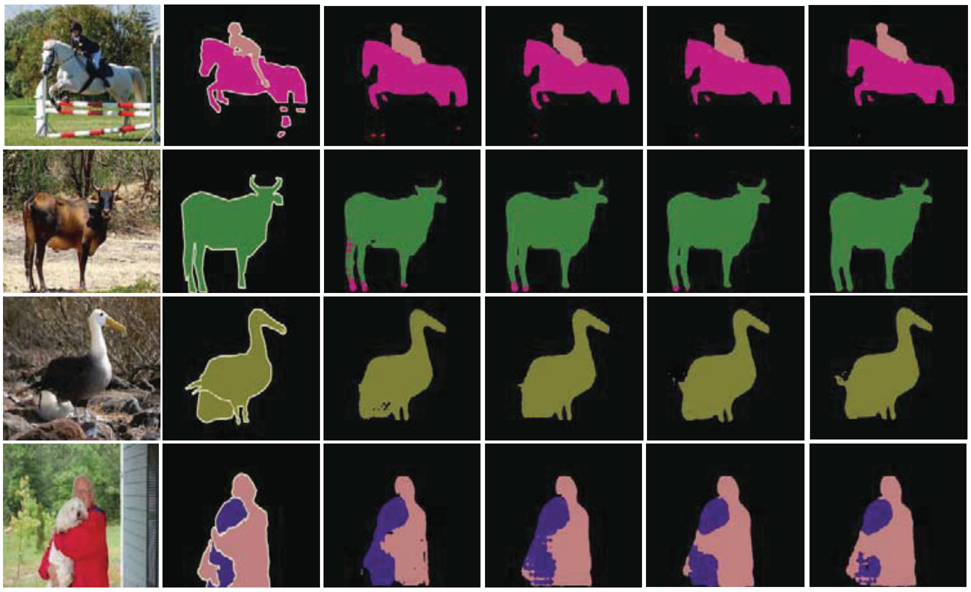

Fig. 4. Comparisons of the GT image segmentation data and the segmentation results obtained by the AdvSemiSeg model and the proposed model for selected images in the PASCAL VOC dataset when using 1/2 of the labeled data during training. The proposed model is qualitatively observed to provide better segmentation results after introducing self-attention (SA) and spectral normalization (SN) than the AdvSemiSeg model.

TABLE IV
IMAGE SEGMENTATION PERFORMANCE RESULTS (MEAN IU) OBTAINED FOR THE PASCAL VOC DATASET USING DIFFERENT BACKBONE ARCHITECTURES WITH DIFFERENT PROPORTIONS OF LABELED DATA.

Methods	1/8	1/4	1/2	Full
Deeplabv2 (v2)	66.0	68.3	69.8	73.6
Ours v2	71.8	73.5	76.3	78.9
Deeplabv3 (v3)	unstable	69.4	70.9	75.2
Ours v3	72.8	75.3	77.0	79.5

TABLE V
IMAGE SEGMENTATION PERFORMANCE RESULTS (MEAN IU) OBTAINED BY THE PROPOSED MODEL UNDER DIFFERENT INCLUSIONS OF MODEL COMPONENTS FOR THE PASCAL VOC DATASETS WITH 1/2 AND THE FULL SET OF LABELED DATA DURING TRAINING.

			Labeled Data	
SA1	SA2	SN	1/2	Full
			73.82	74.98
		✓	74.52	76.72
✓			75.17	77.69
✓	✓		75.94	78.45
✓		✓	76.13	78.68
	✓	✓	76.20	78.74
✓	✓	✓	76.36	78.93

framework, and the image segmentation performances obtained with different proportions of labeled data are listed in Table 4. Interestingly, the DeepLabv3 framework was found to be unstable when training with a 1/8 proportion of labeled data, although the training process was stable under larger proportions of labeled data, and the image segmentation performances were always better than those obtained using the DeepLabv2 framework. We also note that the training instability observed with a 1/8 proportion of labeled data was well mitigated by the application of spectral normalization. At the same time, our semi-supervised model performed even better using the Deeplabv3 backbone.

The effects of the different components included within the training framework were evaluated by conducting an ablation study of our proposed method using 1/2 and the full set of labeled data during training, and the results are listed in Table 5. It can be observed from the table that the inclusion of spectral normalization during training definitely enhances the image segmentation performance of the model. In addition, the inclusion of SA2 appears to have a somewhat more profound effect on the segmentation performance than SA1, although both modules definitely improve the segmentation performance. Overall, our approach is demonstrated to be very effective for improving the image segmentation performance of semi-supervised GAN networks.

VI. Conclusions

This study presented an improved process for conducting semi-supervised training of GAN frameworks for semantic image segmentation. First, we introduced self-attention modules in the segmentation network to effectively account for relationships between widely separated spatial regions of the input image and thereby capture long-range contextual information. In addition, we applied spectral normalization in the discriminator network to enhance the stability of the GAN training process, making the generated examples more diverse than those obtained using conventional weight normalization. The proposed stable self-attention adversarial learning semi-supervised semantic image segmentation network was demonstrated to provide superior image segmentation performance compared with the results of current semi-supervised and fully-supervised semantic image segmentation techniques based on applications to two datasets.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Nos. 61966004, 61663004, 61866004, 61762078), the Guangxi Natural Science Foundation (Nos. 2019GXNSFDA245018, 2018GXNSFDA281009, 2017GXNSFAA198365), the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, the Guangxi Talent Highland Project of Big Data Intelligence and Application, Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

REFERENCES

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets,atrous convolution, and fully connected crfs, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 834–848.
[2] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in:Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3194–3203.
[3] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[4] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, I. J. Goodfellow, Realistic evaluation of semi-supervised learning algorithms, 2018.
[5] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122, 2015.
[6] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
[7] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1529–1537.
[8] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361.
[9] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, M.-H. Yang, Deep image harmonization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3789–3797.
[10] T. Kalluri, G. Varma, M. Chandraker, C. Jawahar, Universal semi-supervised semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5259–5270.
[11] S. Stekovic, F. Fraundorfer, V. Lepetit, S4-net: Geometry-consistent semi-supervised semantic segmentation, arXiv preprint arXiv:1812.10717, 2018.
[12] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in:Advances in neural information processing systems, 2014, pp. 2672–2680.
[14] B. Shuai, Z. Zuo, B. Wang, G. Wang, Scene segmentation with dag-recurrent neural networks, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 1480–1493.
[15] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, J. Jia, Psanet: Point-wise spatial attention network for scene parsing, in:Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283.
[16] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[17] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, arXiv preprint arXiv:1703.03130, 2017.
[18] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, C. Zhang, Disan: Directional self-attention network for rnn/cnn-free language understanding, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[20] M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks. arxiv 2017.
[21] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434, 2015.
[22] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, arXiv preprint arXiv:1701.07875, 2017.
[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[24] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, arXiv preprint arXiv:1802.05957, 2018.
[25] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303–338.
[26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp.3213–3223.
[27] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, M.-H. Yang, Adversarial learning for semi-supervised semantic segmentation, arXiv preprint arXiv:1802.07934, 2018.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062, 2014.
[31] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, G. Wang, Context contrasted feature and gated multi-scale aggregation for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2393–2402.
[32] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[33] B. Hariharan, P. Arbel´aez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 991–998.
[34] G. French, T. Alia, S. Laine, M. Mackiewicz, C. Finlayson, Consisency regularization and cutmix for semi-supervised semantic segmentation, arXiv preprint arXiv:1906.01916, 2019.
[35] S. Mittal, M. Tatarchenko, T. Brox, Semi-supervised semantic segmentation with high-and low-level consistency, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.