Hindawi
Advances in Multimedia
Volume 2018, Article ID 8207201, 8 pages
https://doi.org/10.1155/2018/8207201

Research Article

Scene Understanding Based on High-Order Potentials and Generative Adversarial Networks

Xiaoli Zhao, Guozhong Wang, Jiaqi Zhang, and Xiang Zhang

School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

Correspondence should be addressed to Xiaoli Zhao; evawhy@163.com

Received 31 May 2018; Accepted 19 July 2018; Published 5 August 2018

Academic Editor: Shih-Chia Huang

Scene understanding is to predict a class label at each pixel of an image. In this study, we propose a semantic segmentation framework based on classic generative adversarial nets (GAN) to train a fully convolutional semantic segmentation model along with an adversarial network. To improve the consistency of the segmented image, the high-order potentials, instead of unary or pairwise potentials, are adopted. We realize the high-order potentials by substituting adversarial network for CRF model, which can continuously improve the consistency and details of the segmented semantic image until it cannot discriminate the segmented result from the ground truth. A number of experiments are conducted on PASCAL VOC 2012 and Cityscapes datasets, and the quantitative and qualitative assessments have shown the effectiveness of our proposed approach.

1. Introduction

Scene understanding, based on semantic segmentation, is a core problem in the field of computer vision, which has been applied to 2D image, video, and even volumetric data. Its goal is to assign each pixel a label and then provide complete understanding of a scene. Two examples of scene understanding are shown in Figure 1. The importance of scene understanding is highlighted by the fact that there are increasing applications, such as autonomous driving [1], human-computer interaction [2], robot technology, and augmented reality, to name a few.

The earliest scene parsing [3] is to classify 33 scenes for 2688 images on LMO dataset, which adopts label transfer technology to establish dense correspondences between the input image and each of the nearest neighbors using SIFT flow algorithm. State-of-the-art scene parsing frameworks are mostly based on fully convolutional network (FCN) [4]. FCN transforms the well-known networks-AlexNet, VGG, GooLeNet, and ResNet into fully convolutional ones by replacing the fully connected layers with convolutional ones. The key insight of FCN is to build the "fully convolutional" networks that take input of arbitrary size and produce corresponding-sized output with efficient inference and learning and realize end-to-end and image-to-image system of deep learning. For all these reasons and other contributions, FCN is considered as the milestone of deep learning. Although amounts of pooling operations enlarge the receptive fields of the convolution kernel of FCN, they lose the detailed location information, resulting in coarse segmentation result, which hinders its further application.

In order to refine the segmentation result, a postprocessing stage using conditional random field (CRF) is adopted after the output of system [5], which makes use of the fully connected pairwise CRF to capture the dependencies of pixels and achieve fine local details. Dilated convolution is a generalization of Kronecker-factored convolutional filters [6] which expand exponentially receptive fields without losing resolution by disposing of some pooling layers. The works [7] that make use of this technique allow dense feature extraction on any arbitrary resolution and then combine dilated convolutions of different scales to have wider receptive fields with no additional cost. Combined CRF with dilated convolution, Chen et al. [8] propose the "deeplab" system, which enlarges the receptive fields of filters at multiple scales and overcomes the disadvantage of location accuracy by using a fully connected CRF to response the final layer of network. In order to take the dense CRF with pairwise potentials as an integral part of the network, Zheng et al. [9] propose a model called CRFasRNN to refine the

(a) (b)

Figure 1: Examples of scene parsing: (a) image; (b) ground truth.

segmentation of FCN; they make it possible to fully integrate the CRF with a FCN and train the whole network end to end. Although CRF taking into account the correlation of pixels has improved the segmentation accuracy, it has also increased the computational complexity. To incorporate suitable global features, Zhao et al. [10] propose a pyramid scene parsing network (PSPNet), which extends the pixel-level feature to special designed pyramid pooling one in addition to traditional dilated convolution. This algorithm achieves the champion of ImageNet scene parsing challenge 2016.

In the above-mentioned algorithms, a common property is that all label variables are predicted either using unary potentials such as FCN or using pairwise potentials such as methods based on CRF. Despite the fact that pairwise potentials refine the accuracy of semantic segmentation, they only consider the correlation of two pixels. In an image, many pixels have the consistency across superpixels; high-order potentials should be effective in refining the segmentation accuracy. Arnab et al. [11] have integrated specific classes of high-order potentials in CNN-based segmentation models. This specific class may be object or superpixel and so on, for which we need to design different energy function to calculate high-order potentials, whose computation is complicated.

The generative adversarial nets (GAN) proposed by Goodfellow et al. [12] in 2014 can be characterized by training a pair of networks in competition with each other, in which an adversarial network can estimate the generative model without approximating many intractable probability computation. Because there is no need for any Markov chains or unrolled approximate inference network, GAN has drawn many researchers' attention in the domains of superresolution [13], image-to-image translation [14, 15], and image synthesis [16, 17], etc. We are interested in higher-order consistency without confining to a certain class. We also do not want to have complex probability or inference computation. Motivated by all kinds of GAN, we proposed a semantic segmentation framework based on GAN, which consists of two components: generative network and adversarial network. The former one generates the segmented image, and the latter one encourages the segmentation model to improve continuously the semantic segmentation result until it cannot be distinguished from the ground truth according to the value of loss function. Different from the classic GAN, we take the original image as the input of the generative network and and the output of generative network or corresponding ground truth as the input of the adversarial network; then adversarial network discriminates the similarity of two inputs. If the value of loss function of the framework is large, backpropagation is performed to adjust the parameters of the network; if the value of loss function satisfies the termination criterion, the output of the generative network is the final semantic segmentation result. The semantic segmentation framework based on GAN is shown in Figure 2. This approach takes into account the high-order potentials of an image because it differentiates the similarity between the segmented image and the corresponding ground truth in the whole image.

2. The Proposed Semantic Segmentation Approach

The aim of the proposed framework is to generate the semantic image $G(x_n)$ from an original image x_n. To achieve this goal, we design a generator network G and an adversarial network D. The generator is trained as a network parameterized by θ_G. These parameters denote the weights and are obtained by minimizing the loss function; then the output $G(x_n)$ of generator and the ground truth y_n are fed into the adversarial network parameterized by θ_D, in which the discriminator is trained to distinguish real or fake value. In order to achieve the desired result, it is important to design the architecture network and loss function.

2.1. The Architecture of Networks. Some works have shown that deeper network model can improve the performance of

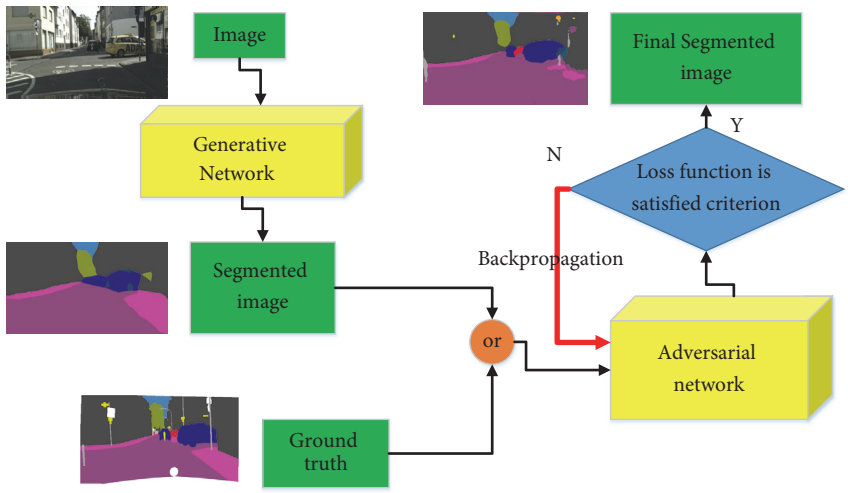


Figure 2: Overview of the proposed scene parsing framework.

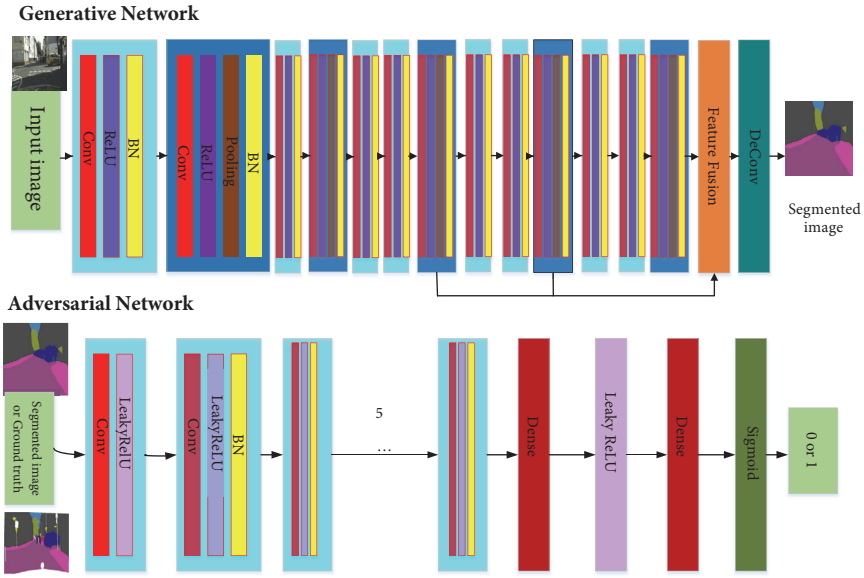


Figure 3: Architecture of generative and adversarial networks.

the segmentation and meanwhile make the architecture of the network complex, resulting in difficult training [18]. We make a compromise between the depth of the network and the performance of the algorithm.

In the generative network, which is shown in the first row of Figure 3, there are two modules of convolution and deconvolution. The role of convolution module is to extract the feature maps of an image, which consists of 10 layers. Each layer is composed of convolution, activation function, and batch normalization. The convolution is performed with 3×3 kernels and 64 feature maps followed by ReLU layer as the activation function, whose role is to conduct the nonlinear operation. Batch normalization is performed to avoid the network overfitting in each layer. Although pooling operations enlarge the receptive field of the network, they also reduce the accuracy of the segmentation. To improve the fine details of feature maps, the last three pooling outputs are integrated into one, on which deconvolution is performed to achieve the same size output with the original image.

To discriminate the ground truth from the segmented image, we train a discriminator network, which is illustrated in the second row of Figure 3. This architecture follows literature [13] to solve (4) in an alternating manner along with the generator. It contains eight convolution layers and uses LeakyReLU as the activation function. The convolution is conducted by 3×3 kernels, resulting in final feature maps of size 512, which are followed by two dense layers and a final sigmoid activation function to achieve a probability for classification.

2.2. Loss Function. In terms of information theory, cross entropy denotes the similarity of two variables; the more

similar the distribution of two variables, the smaller the cross entropy, so we adopt the cross entropy as the loss function. The definition of cross entropy is shown in the following:

$$CE\left(p,\widehat{p}\right) = -\sum_{i} p_i \log \widehat{p}_i. \tag{1}$$

where p and $\widehat{p}$ are the real value and predicted value. Equation (1) is Shannon entropy when p and $\widehat{p}$ are equal. In the multiple classification task, we use one-hot encoding cross entropy. Equation (1) can be rewritten as follows:

$$CE\left(y,\widehat{p}\right) = -\sum_{i} y_i \log \widehat{p}_i = -\log \widehat{p}_i. \tag{2}$$

where y specifies one pixel of ground truth and y_i represents 0 or 1.

The loss function of the proposed networks is a weighted sum of two terms. The first is a multiclass cross entropy term of a generator that encourages the segmented output similar to the input. We use $G(x)$ to denote the class probability map over C classes of size $H \times W \times C$ that the segmentation model generates given an input image x of size $H \times W \times C$. This segmentation model predicts the right class label at each pixel independently, which is described in the following:

$$GL\left(\theta_G\right) = l_{mce}\left(y, G\left(x\right)\right) = -\sum_{i=1}^{H\times W} \log \widehat{p}_i. \tag{3}$$

where $l_{mce}(G(x), y)$ represents the cross entropy loss function of multiple classification on an image of size $H \times W$, in which the class probability of per-pixel is predicted as $\widehat{p}_i$.

The second loss term represents the loss of the adversarial network. If the adversarial network can distinguish the output of generator from the ground truth, the loss value is large; otherwise, the loss is small. Because the loss is calculated based on the whole image or a large portion of it, this high-order statistics dissimilarity can be penalized by the adversarial loss term. We take the output of the adversarial network as $D(\cdot) \in [0, 1]$. Training the adversarial model is equivalent to minimizing the following binary classification loss:

$$AL\left(\theta_D\right) = l_{bce}\left(D\left(y\right), 1\right) + l_{bce}\left(D\left(G\left(x\right)\right), 0\right). \tag{4}$$

where l_{bce} denotes the binary cross entropy loss and $D(y)$ and $D(G(x))$ represent the label maps of adversarial network when the network input is the ground truth y or the output of a generator $G(x)$.

Given a data set of N original images x_n and the corresponding ground truth y_n, we define the total loss functions of the proposed semantic segmentation networks based on GAN as in the following:

$$\begin{aligned} TL\left(\theta_G, \theta_D\right) &= GL\left(\theta_G\right) + \alpha \times AL\left(\theta_D\right) \\ &= \sum_{n=1}^{N} \left(l_{mce}\left(y_n, G\left(x_n\right)\right) + \alpha\right. \\ &\left. \times \left(l_{bce}\left(D\left(y_n\right), 1\right) + l_{bce}\left(D\left(G\left(x_n\right)\right), 0\right)\right)\right). \end{aligned} \tag{5}$$

where α denotes weight factor. In this paper, we set it as 0.01.

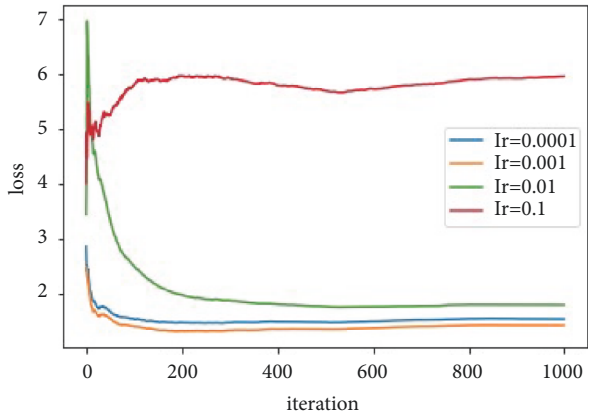


Figure 4: Divergence of different learning rate.

3. Experiments

To evaluate the proposed scene understanding algorithm based on GAN, we conduct some experiments on two widely used datasets, including PASCAL VOC 2012 [19] and urban scene understanding dataset Cityscapes [1]. We train networks on a NVIDIA Tesla K40 GPU and Intel Xeon E5 CPU using 2000 iterations and the batch size of size 16.

To quantitatively assess the accuracy of scene parsing, four performance indices are adopted: pixel accuracy (PA), mean pixel accuracy (MPA), mean intersection over union (MeanIoU), and frequency weighted intersection over union (FWIoU), whose formulations [20] are in (6)–(9). We assume a total of $k + 1$ classes, and p_{ij} is the amount of pixels of class i inferred to belong to class j. p_{ii} denotes the number of true positives, while p_{ij} and p_{ji} are usually represented as false positives and false negatives, respectively:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}. \tag{6}$$

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}. \tag{7}$$

$$MeanIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{8}$$

$$FWIoU = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}}\sum_{i=0}^{k}\frac{\sum_{j=0}^{k} p_{ij}p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{9}$$

We use adaptive estimates of first-order moments (ADAM) [21] to optimize the algorithm because it requires little parameter-tuning, in which β_1 and β_2 are set to 0.9 and 0.999, respectively. We have also compared the divergence of different learning rate on the algorithm to select the optimal value, which is shown in Figure 4. According to this figure, we select 10^{-3} as the rate learning in these experiments.

3.1. Experiment 1: PASCAL VOC 2012. We carry out experiments on PASCAL VOC 2012 segmentation dataset, which

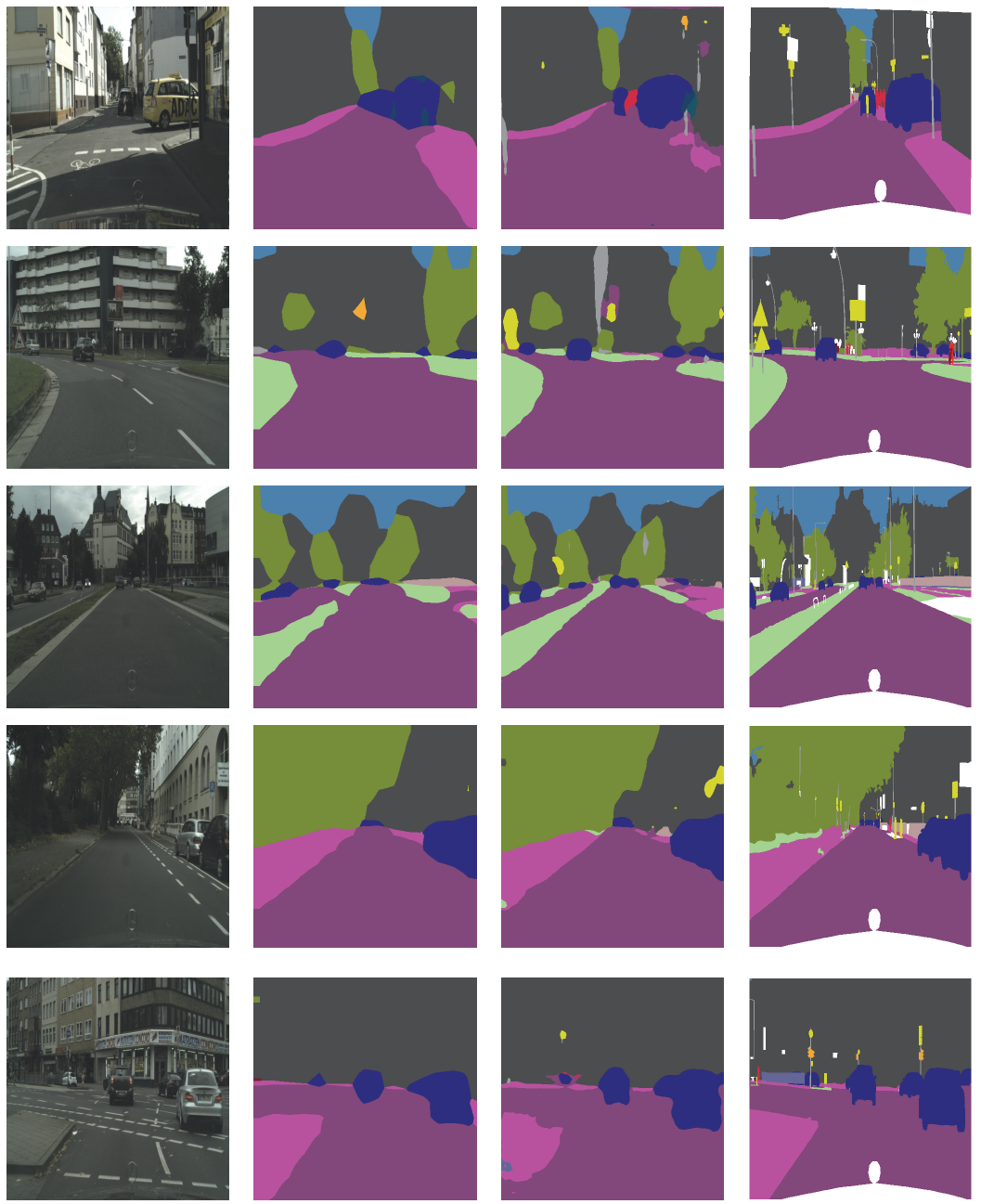
Figure 5: Comparison of different semantic segmentation algorithms on PASCAL VOC 2012: (a) original image; (b) FCN-8S; (c) DeepLab; (d) our method; (e) ground truth.

contains 20 object categories and 1 background class. Its augmented dataset [22] includes 10582, 1449, and 1456 images for training, validation, and testing. We have compared our method with the classic FCN [4] and popular DeepLab [5]: the accuracy of every class is shown in Table 1. Except for bicycle class, our approach achieves the highest accuracy on other 20 classes. Table 2 illustrates the four performance indices of different algorithms, PA, MPA, MeanIoU, and FWIoU. It is obvious that, from the left to right column, the accuracy of the algorithm gradually increases. The proposed approach gets the highest accuracy on these four performance indices.

To qualitatively validate the proposed method, several examples are exhibited in Figure 5. For "cat" in row one, our method gets the cat in accordance with the ground truth; however, FCN and DeepLab segment other noise regions. For "cow" and "child" in rows two and five, the details, such as leg, can be segmented in our method, while leg cannot be found in images using other two methods. In the fourth image, little cow and person are segmented in fine contour comparing with other two methods. In a word, the subjective quality of the segmented image using DeepLab is better than that using FCN; the segmented result using our method outperforms those using FCN and DeepLab.

3.2. Experiment 2: Cityscapes. Cityscapes [1] is a dataset for semantic urban scene understanding which was released in 2016. It contains 5000 high quality pixel-level finely annotated images collected from 50 cities in different seasons. The images, which consists of 2975, 500, and 1524 images for training, validation, and testing, are divided into 19 categories. Because this dataset is recently released, previous algorithms have not issued code for this dataset. We only do subjective assessment for Cityscapes using our method and FCN.

Several examples are shown in Figure 6. It is clear that our proposed method outperforms FCN and can achieve more details and distinguish road, building, cars, etc.



(a) (b) (c) (d)

Figure 6: Comparison of different semantic segmentation algorithms on Cityscapes: (a) original image; (b) FCN-8S; (c) our method; (d) ground truth.

4. Conclusion

In this paper, we propose a scene understanding framework based on generative adversarial networks, which trains the fully convolutional semantic segmentation network by adversarial network, and adopt high-order potentials to achieve the fine details and consistency of the segmented semantic image. We perform a number of experiments on two famous datasets, PASCAL VOC 2012 and Cityscapes. We analyze not only each class accuracy but also four accuracy indices by using different semantic segmentation algorithms. The quantitative and qualitative assessments have shown our proposed method achieves the best accuracy among all algorithms. In the future, we will do more experiments on Cityscapes dataset and address the misclassification caused by class imbalance.

Data Availability

The data used to support the findings of this study are included within the article.

Table 1: Each class accuracy.

Class Label	FCN-32s	FCN-16s	FCN-8s	DeepLab	Our Method
Background	92.8	92.8	91.2	92.6	93.8
aeroplane	75.4	76.2	76.8	83.5	86.1
bicycle	33.6	34.3	34.4	36.6	35.9
bird	67.7	68.2	68.9	82.5	87.7
boat	48.6	49.4	49.4	62.3	63.5
bottle	58.4	59.2	60.3	66.5	67.2
bus	73.4	74.6	75.3	85.4	87.1
car	74.2	73.2	74.4	78.5	82.3
cat	77.6	78.4	77.6	73.7	86.8
chair	21.8	22.5	21.4	30.4	32.3
cow	62.1	62.5	62.5	72.9	76.5
Dining-table	46.3	46.7	46.8	60.4	62.0
dog	68.4	69.8	71.8	78.5	81.1
horse	63.4	63.8	63.9	75.5	77.9
motorbike	76.2	76.4	76.5	82.1	84.3
person	72.3	72.4	73.9	79.7	82.4
Potted-plant	44.5	44.5	45.2	58.2	59.6
sheep	71.2	71.6	72.4	82.0	84.3
sofa	37.4	37.2	37.4	48.8	54.9
train	69.4	69.8	70.9	73.7	76.2
tv/monitor	54.3	54.5	55.1	63.3	64.2

Table 2: Four accuracy indices using different algorithms.

Accuracy	FCN-32s	FCN-16s	FCN-8s	DeepLab	Our Method
PA(%)	82.6	83.7	85.4	87.4	88.2
MPA(%)	61.3	61.8	62.1	69.8	72.6
Mean IOU(%)	63.5	64.5	67.2	70.3	73.9
FW IOU(%)	83.6	84.1	84.7	86.9	88.4

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by Shanghai Science and Technology Committee (no. 15590501300).

References

[1] M. Cordts, M. Omran, S. Ramos et al., "The Cityscapes dataset for semantic urban scene understanding," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 3213–3223, USA, July 2016.

[2] M. Oberweger, P. Wohlhart, and V. Lepetit, *Hands deep in deep learning for hand pose estimation*, Computer Science, 2015.

[3] C. Liu, J. Yuen, and A. Torralba, "Nonparametric Scene Parsing via Label Transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2368–2382, 2011.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 3431–3440, USA, June 2015.

[5] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *Computer Science*, vol. 4, pp. 357–361, 2014.

[6] S. Zhou, J. N. Wu, Y. Wu, and X. Zhou, Exploiting local structures with the kronecker layer in convolutional networks, 2015.

[7] F. Yu and V. Koltun, *Multi-scale context aggregation by dilated convolutions*, arXiv preprint, 2015.

[8] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and 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*, vol. 40, no. 4, pp. 834–848, 2018.

[9] S. Zheng, S. Jayasumana, B. Romera-Paredes et al., "Conditional random fields as recurrent neural networks," in *Proceedings of the 15th IEEE International Conference on Computer Vision, ICCV 2015*, pp. 1529–1537, Chile, December 2015.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, Honolulu, HI, July 2017.

[11] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr, "Higher Order Conditional Random Fields in Deep Neural Networks," in *Computer Vision – ECCV 2016*, vol. 9906 of *Lecture Notes in Computer Science*, pp. 524–540, Springer International Publishing, Cham, 2016.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*, pp. 2672–2680, Canada, December 2014.

[13] C. Ledig, L. Theis, F. Huszar et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, Honolulu, HI, July 2017.

[14] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, Honolulu, HI, July 2017.

[15] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, Venice, October 2017.

[16] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked Generative Adversarial Networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1866–1875, Honolulu, HI, July 2017.

[17] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, *Generative adversarial text to image synthesis*, arXiv preprint, 2016.

[18] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 1–9, Boston, Mass, USA, June 2015.

[19] M. Everingham, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[20] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey

on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018.

[21] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint, 2014.

[22] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 991–998, Spain, November 2011.