GAN and DCN Based Multi-step Supervised Learning for Image Semantic Segmentation

Jie Fang[1,2(✉)] and Xiaoqian Cao[3]

[1] Center for OPTical IMagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, People's Republic of China
fangjie2015@opt.cn
[2] University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, People's Republic of China
[3] College of Electrical and Information Engineering, Shaanxi University of Science and Technology, Xi'an 710021, Shaanxi, People's Republic of China
caoxiaoqian@sust.edu.cn

Abstract. Image semantic segmentation contains two sub-tasks, segmenting and labeling. However, the recent *fully convolutional network* (FCN) based methods often ignore the first sub-task and consider it as a direct labeling one. Even though these methods have achieved competitive performances, they obtained spatially fragmented and disconnected outputs. The reason is that, pixel-level relationships inside the deepest layers become inconsistent since traditional FCNs do not have any explicit pixel grouping mechanism. To address this problem, a multi-step supervised learning method, which contains image-level supervised learning step and pixel-level supervised learning step, is proposed. Specifically, as for the visualized result of image semantic segmentation, it is actually an image-to-image transformation problem, from RGB domain to category label domain. The recent *conditional generative adversarial network* (cGAN) has achieved significant performance for image-to-image generation task, and the generated image remains good regional connectivity. Therefore, a cGAN supervised by RGB-category label map is used to obtain a coarse segmentation mask, which avoids generating disconnected segmentation results to a certain extent. Furthermore, an interaction information (II) loss term is proposed for cGAN to remain the spatial structure of the segmentation mask. Additionally, *dilated convolutional networks* (DCNs) have achieved significant performance in object detection field, especially for small objects because of its special receptive field settings. Specific to image semantic segmentation, if each pixel is seen as an object, this task can be transformed to object detection. In this case, combined with the segmentation mask from cGAN, a DCN supervised by the pixel-level label is used to finalize the category recognition of each pixel in the image. The proposed method achieves satisfactory performances on three public and challenging datasets for image semantic segmentation.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 28–40, 2018.
https://doi.org/10.1007/978-3-030-03335-4_3

Keywords: cGAN · DCN · Image semantic segmentation
Multi-step supervised learning

Image semantic segmentation, which aims to parse image into several semantic regions, specifically, attach one of the annotated semantic category labels to each pixel or super-pixel in the image automatically, is an important task for understanding objects in a scene. As a bridge towards high-level tasks, image semantic segmentation is adopted in various applications, such as human pose estimation [11], visual tracking [9], *etc.* Even though remarkable efforts [7,12,20] have been made for image semantic segmentation during the past decades, this task is still a challenging problem (Fig. 1).

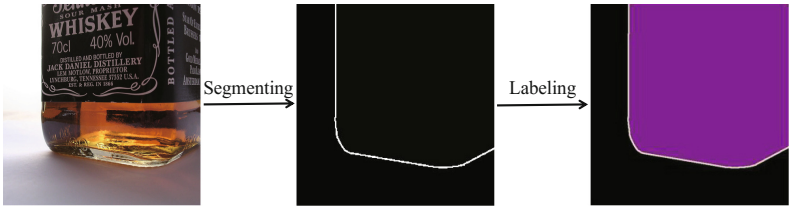

Fig. 1. The task of image semantic segmentation, which includes two steps: segmenting and labeling.

Most recent methods for image semantic segmentation are formulated to solve structured pixel-wise labeling problem on CNNs [1,14,21]. These methods convert an existing CNN architecture for classification to a *fully convolutional network* (FCN) [15]. They obtain a coarse label map from the network by classifying each local region in the image, and perform a simple deconvolution for pixel-level labeling. *Conditional random field* (CRF) [24] is optionally applied to output map for better segmentation. The main advantage of the FCN based methods that the network accepts a whole image as an input and performs fast and accurate inference. Adopting the FCN, many present subsequent methods have solved the challenging task to a certain extent and achieved even better performance. However, all of these methods still rely on traditional FCN architecture. As a result, their segmentation maps are spatially fragmented and disconnected, because traditional FCNs do not any explicit pixel grouping mechanism and then pixel-level relationships inside the deepest layers are inconsistent.

Actually, the task of image semantic segmentation includes two sub-tasks: image segmenting and semantic category labeling, the former focus on the relationships among different pixels while the latter emphasizes on the labeling for each pixel. Inspired by the strong image transformation capacity of cGAN [4] and description capacity for small objects of DCN [13], a multi-step supervised learning method is proposed for image semantic segmentation in this paper, which contains image-level and pixel-level supervised learning steps. Specifically, the

image-level supervised learning step focus on segmenting while the pixel-level supervised learning step aims to consider the labeling precisely. The details of the proposed method can be described as follows:

- **Image-level supervised learning step.** Category label map is considered as a RGB image, and it is used as the ground truth of the image-level supervision to train the cGAN model for transforming original images to region-based ones. Furthermore, in order to remain the spatial structure information of region-based segmented mask, a novel information interaction (II) loss term is incorporated in the framework, which enhances the pixel-to-pixel interaction through considering the 2-order information of generated map. Actually, because the category label map has only few kinds of colors, the generated image from cGAN owns weak semantic information to a certain extent.
- **Pixel-level supervised learning step.** Based on the weak semantic image from cGAN, a DCN is followed to complete the final label recognition of each pixel or super-pixel in the image. As we introduced before, this part is pixel-level supervision. Similar to FCN, multi convolutional layers and a softmax layer are used in this subtask. However, because of the inherent limitation of FCN aforementioned before, in order to obtain a better segmentation result, the FCN architecture need to be modified. Inspired by the successful application of DCN for small object detection, the second sub-Network for pixel-level supervision learning is builded up with dilated kernels, which can remain the spatial information from original image to predicted category label map well.

Even though the proposed method is introduced as two separated parts, the network guided by our method is still builded as the popular end-to-end fashion. The loss functions of the two sub-Networks are summed as the final loss of the overview network, and the parameters of the two sub-Networks are optimized simultaneously. In summary, the contributions of this work are listed as follows:

1. A multi-step supervised learning based approach is proposed for image semantic segmentation, which divides this challenging task into two much simpler ones, image-level supervised learning (segmenting) and pixel-level supervised learning (labeling).
2. A cGAN is used to generated the weak semantic map of the original image. Additionally, a novel information interaction (II) loss term incorporated to the cGAN framework, which can enhance the global and detail structure information of the generated map.
3. A DCN is used to predict the final semantic category label, which has strong discriminative capacity and can remain the spatial information of the image well.

The rest of paper is organized as follows: In Sect. 1, we introduce the related works to image semantic segmentation. Section 2 describes the proposed method. We report the experiment results in Sect. 3 and conclude the paper in Sect. 4.

1 Related Work

This section details some related works for the task of image semantic segmentation. First of all, we introduce some previous works for image semantic segmentation. Additionally, because generative adversarial network and dilated network are used as the sub-Networks of the proposed method, they are also introduced in this section.

1.1 Previous Works

In the past years, image semantic segmentation has attract a lot attentions, because its wide applications. The recent fully convolutional network (FCN) has led to remarkable results in image semantic segmentation task. However, due to the operation and many pooling layers, the FCN typically suffers from low spatial resolution predictions, which causes inconsistent relationships between the neighboring pixels inside the deepest layers. Recently, there has many attempts to address these problems, all of this subsequent work can be divided into several groups. The works in [2,3] used FCN learned potentials, in the separated globalization framework to refine the original FCN results. The methods in [5,23] integrated a CRF-like inference procedure into their network, which allowed to train such models in an end-to-end fashion and achieve satisfactory performances. Even so, these methods did not fix the core problem, such as the lack of consistent mechanism in the deep layers inside the network.

1.2 Generative Adversarial Network

Generative adversarial networks (GAN) [4] is recently introduced as alternative frameworks for training generative models in order to sidestep the difficulty of approximating many intractable probabilistic computations. Adversarial networks have the advantages that Markov chains are never needed, only the backpropagation is used to obtain gradients, no inference is required during learning, and a wide variety of factors and interactions can easily be incorporated into the model. Furthermore, as demonstrated in [4], it can provide state-of-the-art log-likelihood estimated and realistic samples. In an unconditioned generative model, there is no control on models of the data being generated. However, by conditioning the model on additional information it is possible to direct the data generation process. Such conditioning could be based on class labels, on some part of data for inpainting, or data from different modality.

1.3 Dilated Convolutional Network

Dilated convolution [8,17] was original applied for wavelet decomposition in signal processing. It supports exponential expanding of receptive filed. Yu and Koltun developed a convolutional network for dense prediction in [22], in which dilated convolution was adopted to systematically combine multi-scale contextual information without sacrificing resolution or coverage. The method can be

mainly attributed to the expansion of receptive field by dilated convolution, and their work provides a simple yet effective way to enlarge receptive field for CNN. The dilated convolution operator has been referred to in the past as "convolution with a dilated filter". The convolution operator itself is modified to use the parameters in a different way. The dilated convolution operator can apply the same filter as different ranges using different dilation factors, and it enlarges the receptive field without reducing the size of feature map, so it can remain the spatial resolution of image even though many convolutional layers.

2 Proposed Method

This section details the proposed method with two important components, *conditional generative adversarial network* (cGAN) and *dilated convolutional network* (DCN). The cGAN is used to generate the weak semantic map of the image, which considers the category label map as another style of the input image, and this operation is called the image-level supervised learning step. Additionally, based on the generated map from cGAN, The DCN is used to recognize the real semantic category label map of the image while remaining the spatial resolution, which is called the pixel-level supervised learning step. The overview framework of the proposed method is shown in Fig. 2 and the procedures are described as follows:

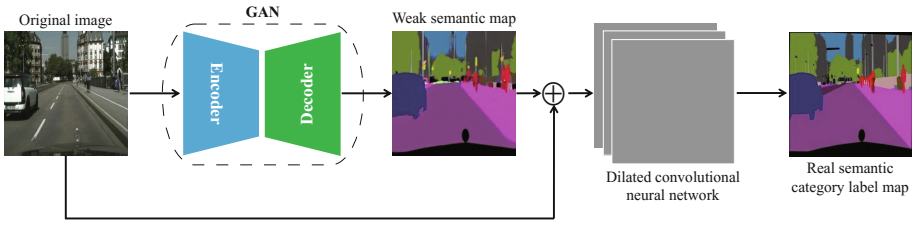

Fig. 2. The overview of the proposed method, including two important components, GAN and dilated convolutional neural network.

- Image-level supervised learning. cGAN equipped with a novel information interaction(II) loss term is used to generate the weak semantic (regional connected) map of original image.
- Pixel-level supervised learning. Combined with the generated weak semantic map from cGAN, a dilated convolutional network is used to recognize the semantic category label of each pixel in the image.
- The aforementioned two supervised learning steps are incorporated into a whole, and the two sub-Networks are optimized simultaneously in an end-to-end way.

2.1 Image-Level Supervised Learning Step

Image-level supervised learning step is mainly to finish the "segmentation" subtask of the image semantic segmentation. There are many unsupervised methods can complete this subtask, such as cluster and super-pixel-based methods. However, these unsupervised methods are depended on the initial conditions seriously, and the segmentation results are not stable. In this work, we transform segmentation to a generation problem in an appropriate way. Recently, GANs have achieved significant performances for image generation task, especially the successful application of *conditional generative adversarial network* (cGAN), it can transform the image from one style to another very well. Actually, the segmented map is another style of the original image, from a pixel-based one to a region-based one. In this case, cGAN can be used to generate the region based maps with weak semantic information. The loss function of the conditional cGAN used in this paper is shown as Eq. 1.

$$\ell_{cGAN} = \min_G \max_D \left\{ \mathrm{E}_{x\sim p_d(x)} \left[\log D\left(x \,|y\right) \right] + \mathrm{E}_{z\sim p_z(z)} \left[\log\left(1 - D\left(G\left(z \,|y\right)\right)\right) \right] \right\}, \tag{1}$$

where G is the generator and D is the discriminator. G tries to minimize this objective against an adversarial D that tries to maximize it. For instance, $G^* = \arg\min_G \max_D L_{cGAN}(G, D)$.

Additionally, specific to the image segmenting subtask, besides the probability distribution information, spatial structure information should be considered in the proposed model as well. To address this issue, we mixed the cGAN objective with another two loss terms: L_1 distance term and the proposed information interaction (II) loss term. The L_1 and II loss terms are shown as Eqs. 2 and 3 respectively,

$$L_1(G) = \mathrm{E}_{x,y,z} \left[\|y - G\left(x, z\right)\|_1 \right], \tag{2}$$

$$II(G) = \mathrm{E}_{x,y,z} \left[\left\|y^2 - G^2\left(x, z\right)\right\|_1 \right], \tag{3}$$

where L_1 is to ensure the 1st-order information of the generated weak semantic map. In other words, it aims to estimate the label of each single pixel accurately as much as possible. II is to ensure the 2nd-order information of the generated weak semantic map. Specifically, as is shown in Fig. 4, corresponding elements in two matrices with small L_1 distance are closer. As for small II distance, besides the accurate single pixel information, the similar relationships among different elements are needed. That is to say, if we incorporate L_1 and II distances in the cGAN loss, the generated map will have the similar intensity and spatial structure information with the groundtruth. Therefore, the final objective the cGAN here is shown as Eq. 4,

$$G^* = \arg\min_G \max_D \mathrm{E}_{x,y,z} L_{cGAN} + \lambda_1 \cdot L_1 + \lambda_2 \cdot II, \tag{4}$$

where λ_1 and λ_2 are two regulation parameters, which are used to balance the relationships of three loss terms.

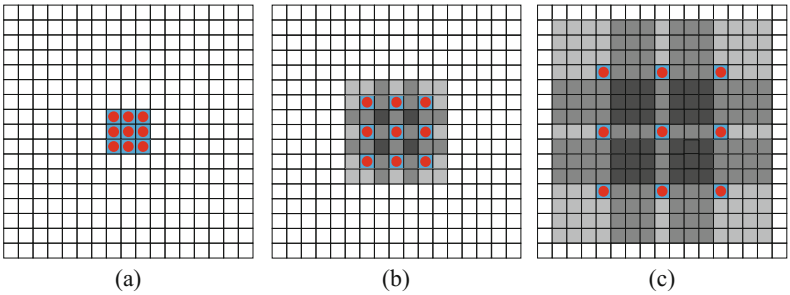

Fig. 3. Different receptive fields with different factors in dilated convolutional neural networks. (a), (b), (c) are 1-dilated, 2-dilated and 3-dilated, respectively, and the receptive field sizes of them are 3×3, 7×7 and 15×15, respectively.

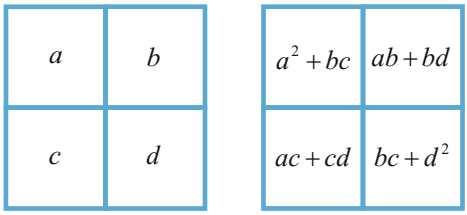

Fig. 4. *The details of* L_1 *and* II. Two matrices with small L_1 distance means that corresponding element-pair is close to each other. For instance, the first element in another matrix is close to 'a' if this matrix has small L_1 distance with the left one. Additionally, II is the 2nd-order of the matrix. Two matrices with small II distance means that, their structure interaction information is similar. For example, the first element in another matrix is close to $a^2 + bc$ if this matrix has small II distance with the left one. As can be seen, $a^2 + bc$ not only contains the information of the first element 'a' in the left matrix, but also contains the interaction information of 'b' and 'c'.

2.2 Pixel-Level Supervised Learning Step

Pixel-level supervised learning step is mainly to complete the semantic category labeling subtask of the image semantic segmentation. Most recent methods consider the semantic segmentation task as a direct classification one and they have achieved competitive performance. In order to obtain the global information of the image, traditional CNNs pooling layers to enlarge the receptive field, even though this strategy have gained satisfactory performance for classification and recognition tasks, the experiment results are not in accordance with our expectations when we apply this architecture to semantic segmentation task. Investigate its reasons, the pooling layers result in severe lose spatial information of the image when it enlarge the receptive field by narrowing down the size of the feature map. Even though some FCN-based methods adopt the strategy of pooling-unpooling to make the output of the network has the same size with

the original image, the visualized prediction results are still coarse due to the loss of detailed spatial information.

In this paper, to address the problem aforementioned, we use dilated convolutional kernel (Fig. 3) to replace the traditional kernel, which can enlarge the receptive field without narrowing down the size of feature map. Additionally, a softmax layer is followed by several dilated convolutional layers to recognize the category label of each pixel in the image. The loss function is shown as Eq. 5,

$$L_c = -\frac{1}{n^2}\left[\sum_{i=1}^{n^2}\sum_{j=1}^{k} \mathrm{I}\left\{y^i = j\right\} \log \frac{e^{\theta_j^T x^i}}{\sum_{l=1}^{k} e^{\theta_l^T x^i}}\right], \tag{5}$$

where n is the size of the input image, k is the category number of the dataset. y^i and j are the predicted category label and the real category label of i_{th} image respectively. $\mathrm{I}\{\cdot\}$ is the indicator function and θ represents the parameters of the network.

2.3 Network Architecture and Optimization Strategy

The cGAN used in this paper is a traditional encoder-decoder architecture, which has 10 convolutional layers, 5 layers for encoder and the others for decoder. The dilated network contains five dilation convolution layers and a softmax classifier layer. Additionally, the overall loss function for the proposed network is shown as Eq. 6,

$$\ell = L_{cGAN} + \lambda_1 \cdot L_1 + \lambda_2 \cdot II + \lambda_3 L_c, \tag{6}$$

where L_{cGAN} is the loss function of the conditional GAN. L_c is the loss function of the dilated convolutional network. L_1 is the 1-norm distance. II is the proposed interaction information loss term. And λ_1, λ_2 and λ_3 are three formulation parameters to balance these four loss terms. Two sub-Networks used in this paper are optimized simultaneously in an end-to-end fashion.

Implementation: In order to speed up the convergence, we adopted the "separated & combined" training strategy. Specifically, we obtained the initial parameters of cGAN and DCN by training them separately. Then, combined these two sub-Networks together and obtained the final parameters of the model through joint training in an end-to-end way.

3 Experiments

The section details of the experiment, including datasets, experiment settings, contrasting methods, evaluation metrics and results & analysis.

3.1 Datasets

The proposed method is tested on three public and challenging datasets: SIFT Flow [19], NYUDv2 [18] and SIFT Flow [19] and PASCAL VOC 2012 [6].

NYUDv2 is a RGB-D dataset collected using the Microsoft Kinect, and it has 1449 RGB-D images with pixel-wise labels. This dataset is challenging, because the additional depth information increases the structure complexity of the image.

SIFT Flow is a dataset of 2,688 images with pixel labels for 33 semantic classes.

PASCAL VOC 2012 dataset for semantic segmentation includes 2913 label images for 21 semantic categories. Some samples of the dataset is shown in Fig. 5.

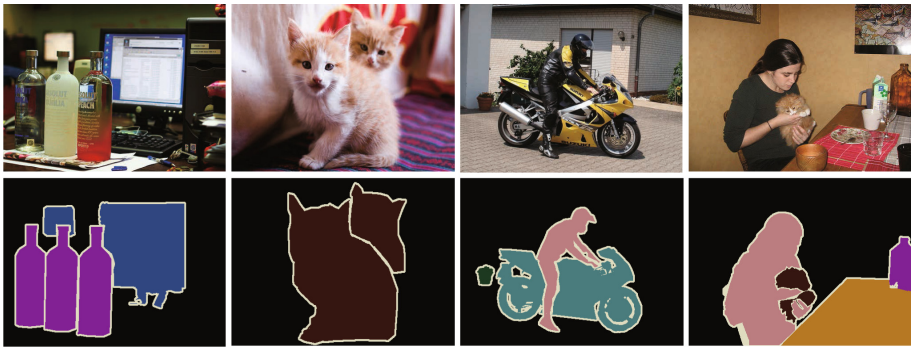

Fig. 5. *Samples of the PASCAL VOC2012 dataset.* The dataset consists of 2913 images from 20 classes.

3.2 Experiment Settings

In this paper, we choose 80% images of the dataset to train the network, and use the others to be the testing set. We train the network using mini-batch SGD with patch size 224×224 and batch size 10. The initial learning rate is set to 2.5×10^{-4}, weight decay is set to 5×10^{-4}, momentum is 0.9 and the network is trained for 200 epoches. Additionally, the formulation parameters λ_1, λ_2 and λ_3 are set to 1 in our experiments.

3.3 Contrasting Methods

To verify the effectiveness of the proposed method, four state-of-the-art methods are used as the contrasting methods: FCN8s, Deconv, cGAN and cGAN+DCN.

FCN8s [15] is one version of the original FCN methods, which achieves the best performance in this series because it uses more information from lower layers of the network.

Deconv [16] is a method based on the encoder-decoder architecture, which achieves the satisfactory performance for image semantic segmentation with double parameters compared to FCN8s.

cGAN [10] is a method which uses conditional generative adversarial network to generate a continuous fake "label map", and then discretizes the fake "label map" as the semantic category label of the image.

cGAN + DCN (Ours1) is a method that uses traditional conditional generate adversarial network to generate a weak semantic "label map", then integrates the original image and weak semantic "label" map information by a dilated convolutional network to finalize the category recognition of each pixel in the image. Additionally, this method trains two sub-Networks separately and the II loss term is not used in the image-level supervised step.

3.4 Evaluation Metrics

Pixel classification accuracy (Pix.acc) and mean intersection over union (Mean IoU) are used to verify the methods.

3.5 Results and Analysis

This section details the experiments on NYUDv2 [18], SIFT Flow [19] and PASCAL VOC 2012 [6]. The experiment results are shown in Table 1.

Table 1. Experiment results on three datasets (%).

Dataset	NYUDv2		SIFT flow		PASCAL VOC 2012	
Method	Pix.acc	Mean IoU	Pix.acc	Mean IoU	Pix.acc	Mean IoU
FCN8s [15]	66.84	40.91	85.82	44.75	92.56	66.48
Deconv [16]	68.30	42.78	87.43	45.59	92.84	69.37
cGAN [10]	55.76	37.35	74.82	39.40	67.36	43.55
Ours1	69.52	44.84	88.07	47.28	93.43	72.69
Ours2	**71.35**	**47.06**	**90.53**	**50.21**	**94.18**	**75.82**

From Table 1 we can see that, Deconv method achieves better performance than FCN8s. Specifically, it obtains 1.46% and 1.87% improvement in terms of pixel.acc and mean IoU on NYUDv2 dataset, respectively. The reason is that, compared to FCN8s, Deconv uses a strategy by enlarging the prediction map gradually by a stride of 2 in each step, this avoids the loss of spatial structure information in a certain extent.

Additionally, the results of cGAN method are not satisfactory as we expected. The reason is that, even though the result through descretizing the output of the traditional cGAN have semantic information in a certain extent, it is coarse

since no classification mechanism is used in this framework. Compared to the cGAN method, Ours1 (cGAN + DCN) method achieves better performance. For example, Ours1 method obtains surprising 13.25% Pix.acc improvement and 7.88% Mean IoU improvement on SIFT Flow dataset. This is because Ours1 method divides the complex image semantic segmentation into two simpler ones with a multi-step supervised learning ones, segmentation supervised by image-level and category labeling by pixel-level. Specifically, cGAN is used to generate a region-based weak semantic map of the original image, and based on this weak-semantic map, a dilated convolutional network is used to predict the precise category label of each pixel while remaining the spatial information of the original image well. This also demonstrates the importance of classification mechanism for image semantic segmentation task.

Finally, Ours2 method gains better performance than Ours1 method. Specifically, our method achieves 0.75% pix.acc improvement and 3.13% mean IoU improvement on PASCAL VOC 2012 dataset, compared to the Ours1 method. The reason is that, compared to Ours1 method which uses two sub-Networks separately, Ours2 method optimizes two sub-Networks simultaneously, this enhances the joint representation capability of the model. Besides, the novel information interaction (II) loss term is vital for the image-level supervised step, which considers the interaction information among different pixels and makes the structure information more precise.

In general, the proposed method achieves the satisfactory performances, and this demonstrates the effectiveness of the proposed framework for the image semantic segmentation task.

4 Conclusion

In this work, a cGAN and DCN based multi-step supervised learning method is proposed for image semantic segmentation task. Specifically, the cGAN used in image-level supervised learning step is to generate initial weak semantic map, and the DCN used in pixel-level supervised learning step is to finalize the category label recognition of each pixel in the image. Additionally, a novel information interaction (II) loss term is proposed to obtain a segmentation map with more precise spatial structure in image-level supervised learning step. Finally, the experiment results on three public and challenging datasets have verified the rationality and effectiveness of the proposed method.

References

1. Dai, J., He, K., Sun, J.: BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation, pp. 1635–1643 (2015)
2. Donahue, J., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning, p. I-647 (2014)
3. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network, pp. 2366–2374 (2014)

4. Goodfellow, I.J., et al.: Generative adversarial nets. In: International Conference on Neural Information Processing Systems, pp. 2672–2680 (2014)
5. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 345–360. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_23
6. Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: International Conference on Computer Vision, pp. 991–998 (2011)
7. He, Y., Chiu, W.C., Keuper, M., Fritz, M.: STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling, pp. 7158–7167 (2016)
8. Holschneider, M., Kronland-Martinet, R., Morlet, J., Tchamitchian, P.: A real-time algorithm for signal analysis with the help of the wavelet transform. In: Combes, J.M., Grossmann, A., Tchamitchian, P. (eds.) Wavelets. IPTI, pp. 286–297. Springer, Heidelberg (1990). https://doi.org/10.1007/978-3-642-75988-8_28
9. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network, pp. 597–606 (2015)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5967–5976 (2017)
11. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221 (2013)
12. Kemker, R., Salvaggio, C., Kanan, C.: High-resolution multispectral dataset for semantic segmentation (2017)
13. Li, J., Wu, Y., Zhao, J., Guan, L., Ye, C., Yang, T.: Pedestrian detection with dilated convolution, region proposal network and boosted decision trees. In: International Joint Conference on Neural Networks, pp. 4052–4057 (2017)
14. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation, pp. 3194–3203 (2015)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
16. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: IEEE International Conference on Computer Vision, pp. 1520–1528 (2015)
17. Shensa, M.J.: The discrete wavelet transform: wedding the a trous and Mallat algorithms. IEEE Trans. Sig. Process. **40**(10), 2464–2482 (1992)
18. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
19. Tighe, J., Lazebnik, S.: SuperParsing: scalable nonparametric image parsing with superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_26
20. Wang, P., et al.: Understanding convolution for semantic segmentation (2017)
21. Yasrab, R.: DCSeg: decoupled CNN for classification and semantic segmentation. In: IEEE Sponsored International Conference on Knowledge and Smart Technologies (2017)
22. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions (2015)

23. Zheng, S., et al.: Conditional random fields as recurrent neural networks, pp. 1529–1537 (2015)
24. Zhou, H., Zhang, J., Lei, J., Li, S., Tu, D.: Image semantic segmentation based on FCN-CRF model. In: International Conference on Image, Vision and Computing, pp. 9–14 (2016)