International Journal of Multimedia Information Retrieval (2018) 7:87–93
https://doi.org/10.1007/s13735-017-0141-z

TRENDS AND SURVEYS

A review of semantic segmentation using deep neural networks

Yanming Guo[1] · Yu Liu[1] · Theodoros Georgiou[1] · Michael S. Lew[1]

Received: 9 October 2017 / Revised: 2 November 2017 / Accepted: 14 November 2017 / Published online: 24 November 2017

Abstract
During the long history of computer vision, one of the grand challenges has been semantic segmentation which is the ability to segment an unknown image into different parts and objects (e.g., beach, ocean, sun, dog, swimmer). Furthermore, segmentation is even deeper than object recognition because recognition is not necessary for segmentation. Specifically, humans can perform image segmentation without even knowing what the objects are (for example, in satellite imagery or medical X-ray scans, there may be several objects which are unknown, but they can still be segmented within the image typically for further investigation). Performing segmentation without knowing the exact identity of all objects in the scene is an important part of our visual understanding process which can give us a powerful model to understand the world and also be used to improve or augment existing computer vision techniques. Herein this work, we review the field of semantic segmentation as pertaining to deep convolutional neural networks. We provide comprehensive coverage of the top approaches and summarize the strengths, weaknesses and major challenges.

Keywords Image segmentation · Computer vision · Deep learning · Convolutional neural networks · Machine learning

1 Introduction

For the last three decades, one of the most difficult problems in computer vision has been image segmentation. Image segmentation is different from image classification or object recognition in that it is not necessary to know what the visual concepts or objects are beforehand. To be specific, an object classification will only classify objects that it has specific labels for such as horse, auto, house, dog. An ideal image segmentation algorithm will also segment unknown objects, that is, objects which are new or unknown. There are numerous applications [1–12] where image segmentations could be used to improve existing algorithms from cultural heritage preservation to image copy detection to satellite imagery analysis to on-the-fly visual search and human–computer interaction. In all of these applications, having access to segmentations would allow the problem to be approached at a semantic level. For example, in content-based image retrieval, each image could be segmented as it is added to the database. When a query is processed, it could be segmented and allow the user to query for similar segments in the database—e.g., *find all of the motorcycles in the database*. In human–computer interaction, every part of each video frame would be segmented so that the user could interact at a finer level with other humans and objects in the environment. In the context of an airport, for example, the security team is typically interested in any unattended baggage, some of which could hold dangerous materials. It would be beneficial to make queries for all objects which were left behind by a human.

Given a new image, an image segmentation algorithm should output which pixels of the image belong together semantically. For example, in Fig. 1, the input image consists of an audience watching two motorcyclists in a race. In Fig. 2, we see the ideal segmentation which clusters the pixels by the semantic objects—all of the pixels belonging to a motorcycle are colored green to show they belong together, similarly with the riders and audience who are colored pink.

It is currently unclear how the human brain finds the correct segmentation. Segmenting an image involves a deep

✉ Michael S. Lew
m.s.k.lew@liacs.leidenuniv.nl

Yanming Guo
y.guo@liacs.leidenuniv.nl

Yu Liu
y.liu@liacs.leidenuniv.nl

Theodoros Georgiou
t.georgiou@liacs.leidenuniv.nl

[1] LIACS, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

Fig. 1 Motorcycle racing image

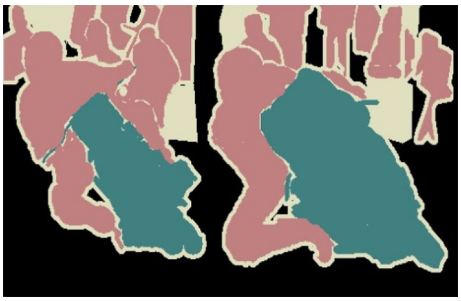

Fig. 2 Segmentation for motorcycle racing image

semantic understanding of the world and which things are parts of a whole.

Traditional image segmentation algorithms are typically based on clustering often with additional information from contours and edges [1,2,13]. For example, in the simplest case, satellite image segmentation can often successfully be performed by clustering pixels based on wavelength, that is, one would create clusters based on similar pixels which are also located spatially nearby.

There have been numerous enhancements and evolutions to the clustering approach. One of the most well-known and significant approaches is modeling using a Markov process [14]. Another notable method was combining contour detection in a hierarchical approach [15]. In SAR imagery, region growing with unsupervised learning was explored [8]. For good overviews of the older pre-deep learning approaches, we refer the reader to several surveys [9,16–20] which cover the works spanning color and edge image segmentations to medical image understanding. However, recent advances have made many of the older methods obsolete. Therefore, we turn to the current approaches which are considered to be the state of the art and have achieved the top benchmark performance across the well-known international datasets.

According to the main component of recent semantic segmentation methods, we divide them into three categories: region-based semantic segmentation, FCN-based semantic segmentation and weakly supervised segmentation. In the next part, we will talk about their main ideas.

2 Region-based semantic segmentation

The region-based methods generally follow the "segmentation using recognition" pipeline, which first extracts free-form regions from an image and describes them, followed by region-based classification. At test time, the region-based predictions are transformed to pixel predictions, usually by labeling a pixel according to the highest scoring region that contains it [21].

Regions with CNN feature (RCNN) [22] is one representative work for the region-based methods. It performs the semantic segmentation based on the object detection results. To be specific, RCNN first utilizes selective search [23] to extract a large quantity of object proposals and then computes CNN features for each of them. Finally, it classifies each region using the class-specific linear SVMs. Compared with traditional CNN structures which are mainly intended for image classification, RCNN can address more complicated tasks, such as object detection and image segmentation, and it even becomes one important basis for both fields. Moreover, RCNN can be built on top of any CNN structures, such as AlexNet [24], VGG [25], GoogLeNet [26] and ResNet [27].

For the image segmentation task, RCNN extracted two types of features for each region: full region feature and foreground feature, and found that it could lead to better performance when concatenating them together as the region feature. RCNN achieved significant performance improvements due to using the highly discriminative CNN features. However, it also suffers from three main drawbacks for the segmentation task, which motivated significant research:

1. *The feature is not compatible with the segmentation task.* Although the CNN feature has been repeatedly shown to give higher performance as compared to conventional hand-crafted features like SIFT [28] and HOG [29], it is not specifically designed for the image segmentation task. Hariharan et al. [30] argued that the network RCNN utilized was actually fine-tuned to classify bounding boxes (i.e., to extract full region features), making it suboptimal to extract foreground features. To address this issue, they introduced one additional network which was specifically fine-tuned on the region foreground and proposed to jointly train the two networks. For the proposal generation, SDS [30] replaced selective search with MCG [31] and reported better results. Given pre-computed proposals [21], aimed to combine the region classification and semantic segmentation together. It introduced a differentiable region-to-pixel layer which could map image regions to image pixels, making the whole network specifically fine-tuned for the image segmentation task.
2. *The feature does not contain enough spatial information for precise boundary generation.* RCNN employed the activations from the fully connected layer, which have

been verified to be more semantically meaningful than the features from intermediate layers. However, the intermediate layer activations contain more spatial information and thus are more precise in localization. To get the best of both worlds [32], utilized hypercolumns as pixel descriptors, which consist of activations of all CNN units above that pixel. Intuitively, the core idea was to treat the stages in the CNN in a similar way as a coarse-to-fine image pyramid where the coarse layer information typically led to higher accuracy, but poor spatial precision and the fine level information led to high spatial precision, but poor accuracy. By combining the coarse and fine layers, prior research had found that fusing the information could result in higher accuracy and precision. So, by connecting and using the information across CNN stages as a hypercolumn, the authors were also able to produce significant improvements. Likewise [33], utilized convolutional feature masking (CFM) to extract segment features directly from the last convolutional feature map, followed by a spatial pyramid pooling (SPP) layer [34]. As a consequence, CFM can determine the segmentation accurately and efficiently.

3. *Generating segment-based proposals takes time and would greatly affect the final performance*. In contrast to prior approaches which only formulated segmentation masks inside the pre-generated proposals, recent works tend to make the whole process end-to-end trainable. This can not only eliminate the side effect of object proposals, but also improve the efficiency. For instance [35], proposed a proposal-free framework, which segmented objects via mid-level patches. As it integrated region generation (i.e., image patches) into the network and modeled the segmentation branch as a pixel-wise classifier, the entire process of segmenting image patches was end-to-end trainable. The final object segmentation was achieved by merging the information from multi-scale patches. One more recent work appeared in [36], which extended Faster RCNN [37] by introducing an additional branch for predicting an object mask. Likewise, the whole network can also be trained end-to-end.

3 FCN-based semantic segmentation

The key idea in FCN-based methods [38–40] is that they learn a mapping from pixels to pixels, without extracting the region proposals. The FCN network pipeline is an extension of the classical CNN. The main idea is to make the classical CNN take as input arbitrary-sized images. The restriction of CNNs to accept and produce labels only for specific sized inputs comes from the fully connected layers which are, by definition, fixed. Contrary to them, FCNs only have convolutional and pooling layers which give them the ability to make predictions on arbitrary-sized inputs. Although this is the case, the size of the output of FCNs depends on the input size rather than always producing a fixed-size output. Thus, these kinds of networks are commonly used for local rather than global tasks (i.e., semantic segmentation [38] or object detection [41] instead of object classification [37]).

Since FCNs are composed of convolutional, pooling and upsampling layers, depending on the definition of a loss function, they can be end-to-end trainable. The networks of [38] produce a pixel-dense output with 21 channels, each one corresponding to one PASCAL VOC-2012 class, including background. They typically use the per-pixel softmax loss function. Using the above configuration, they tried two different learning schemes. The first approach used a batch size of 20 images and accumulated the gradients from all images, and the second method was with batch size one, or online learning. Their experiments showed that online learning with higher momentum produced better FCN models in less wall-clock training time.

One issue in FCN approaches is that by propagating through several alternated convolutional and pooling layers, the resolution of the output feature maps is down-sampled. Therefore, the direct predictions of FCN are typically in low resolution, resulting in relatively fuzzy object boundaries. A variety of FCN-based approaches are proposed very recently to address this issue. For example [39], proposed a multi-scale convolutional network which consists of multiple scale sub-networks with different resolution outputs to progressively refine the coarse prediction. Long et al. [38] learned to combine coarse, high layer information with fine, low layer information. The multilayer outputs were followed by deconvolutional layers for bilinear upsampling to pixel-dense outputs. To accurately reconstruct highly nonlinear structures of object boundaries [42], replaced the simple deconvolutional procedure in [38] with a deep deconvolutional network for identifying pixel-wise class labels and predicting segmentation masks. Apart from the deconvolutional layers, DeepLab-CRF [43,44] offered an alternative to raise the output resolution. It first applied the atrous convolution to increase the feature resolution and then employed bilinear interpolation to upsample the score map to reach the original image resolution. Afterward, the CRF method [45] was adopted to refine the object boundary. Instead of applying CRF inference as a post-processing step disconnected from the CNN training [46], extended [43,44] and introduced an end-to-end trainable network by interpreting the dense CRFs as a recurrent neural network (RNN).

In addition to producing high-resolution prediction for better segmentation, some works attempted to improve the segmentation precision through exploiting the contextual information. For example [47], utilized global average pooling to obtain global context and added the global context into fully convolutional networks for semantic segmenta-

tion, bringing consistent increase for the accuracy. As its extension [48], raised a more representative global context information by different-region-based context aggregation via the pyramid scene parsing network. As the alternatives of global context [49], utilized exponentially expanded dilated convolutions to aggregate multi-scale contextual information. Lin et al. [50,51] explored two types of spatial context to improve the segmentation performance: patch-patch context and patch-background context, and utilized CRFs to explicitly model the contextual relations.

4 Weakly supervised semantic segmentation

Most of the relevant methods in semantic segmentation rely on a large number of images with pixel-wise segmentation masks. However, manually annotating these masks is quite time-consuming, frustrating and commercially expensive. Therefore, some weakly supervised methods have recently been proposed, which are dedicated to fulfilling the semantic segmentation by utilizing annotated bounding boxes, or even image-level labels.

For example [52], employed the bounding box annotations as a supervision to train the network and iteratively improved the estimated masks for semantic segmentation. Papandreou et al. [53] proposed an expectation–maximization (EM) method for training semantic segmentation models with weakly annotated data, i.e., image-level or bounding box annotation, and found solely using image-level annotation was insufficient to train a high-quality segmentation model, while using bounding box annotation could obtain a competitive model with pixel-level annotation. Nevertheless, it was generally beneficial to combine them together. In order to adapt well to address the weakly supervised semantic segmentation task, the aforementioned approaches utilized slightly different networks and training procedures with fully supervised semantic segmentation. More recently, Khoreva et al. [54] viewed the weak supervision problem as an issue of input label noise and explored recursive training as a de-nosing strategy. By carefully designing the input labels from given bounding boxes, they reached $\sim$ 95% of the quality of the fully supervised model with the same training procedure.

Aside from employing box annotations as weak supervision signal, there are also some works established based on image-level labels. For instance [55], interpreted the segmentation task within the multiple-instance learning (MIL) framework and added an extra layer to constrain the model to assign more weight to important pixels for image-level classification. During test, the constraining layer is removed and the label of each image pixel is inferred by taking the maximum probability for this pixel. Similar work was proposed in [56], which also cast each image as a bag of pixel-level instances and defined a pixel-level loss for adapting to MIL. On the other hand [57], proposed a self-training framework, i.e., constrained CNN, and utilized a novel loss function to enforce the consistency between the per-image annotation and the predicted segmentation masks.

One main limitation of employing the image-level supervision is the ignorance of the object localization. To improve the localization performance, some approaches [58–61] have proposed to exploit the notion of objectness, either by incorporating it in the loss function [58,59], or by employing pre-trained network as external objectness module [60,61]. Another promising way to improve the segmentation performance is to utilize additional weakly supervised images, such as web images, to train CNNs, such as [62,63].

5 Discussion

5.1 Strengths and benefits

If we are able to perform automatic image annotation, then this can have both practical and theoretical benefits. In classic object recognition, we design an algorithm which can analyze a sub-window within the image to detect a particular object. For example, if one has a classic object detector and a ten megapixel image, then one would try to use the detector at all ten million locations in the image which could easily require minutes to weeks depending upon the complexity of the object detector and the number of image transformations being considered such as rotation and scale.

In the case of automatic image segmentation, instead of having to try using the object detector at all pixel locations, we now only have to try it for the number of segments in the image which is typically between 10 and 100 and certainly orders of magnitude less than the number of locations in an image. Furthermore, one might also try using the object detector at different orientations which can also be alleviated by the image segmentation

The benefits are not limited to merely computational speed, but also to enhancing accuracy. When one does perform window-based object detection, one often also has to deal with background noise and distractors. If the automatic image segmentation algorithm works well, then it will have automatically removed the background noise which will significantly increase the accuracy of the object recognition.

Furthermore, automatic image segmentation can give us insights into how the human visual system is able to perform the same task. It can provide theoretical justifications for the strengths and weaknesses of visual information systems; it can give us deep insight into the conditions when visual information systems will not be able to correctly understand visual concepts or objects in the world.

Automated segmentation can go beyond object recognition and detection in that it is not required to know the object

or visual concepts beforehand. This can lead to major breakthroughs in general computer vision because it allows new objects to be learned by the system. When an unknown object is found and is not classified by the existing database, then a new entry can be made for the new unknown object and this can lead to a truly general computer vision system.
So the main benefits of automatic image segmentation are as follows:

1. It can improve computational efficiency.
2. It can improve accuracy by eliminating background noise.
3. It can give both theoretical and deep insights into both how visual systems work and what the limitations are.
4. It can be more general than object detection and recognition.

However, there are also challenges and pitfalls to be considered. Currently, these can be described as follows:

5.2 Major challenges and weaknesses

1. *How general are the methods?* Realistically, it is unclear how well the top algorithms work on general imagery. It often occurs that the best methods for a dataset are fine-tuned for only the imagery of a specific situation, place or context, so the generality is unclear. Therefore, this is clearly one of the major future challenges for the research community.
2. *How much data are necessary to train the algorithm?* Some of the best approaches require enormous amounts of labeled data. This means that in some situations, those algorithms will be unsuitable because the labeled datasets are unavailable. For scene classification, the credible datasets typically contain millions to hundreds of millions of training images; however, for most applications the training set size is more likely to be in the thousands. If the domain experts find it difficult or impossible to create very large training sets, then is it possible to design deep learning algorithms which require fewer examples?
3. *How much computational resources are required?* Some of the top methods require rather heavy usage of near-supercomputers for the training phase which may not be available in all contexts. Many researchers are therefore considering the question: For a specific number of parameters, what is the best accuracy that can be achieved?
4. *When will the methods fail?* Achieving higher accuracy is beneficial, but it is important to have an understanding of the ramifications of incorrect segmentations. In some situations such as driving an automobile in a city, it is not difficult to encounter segmentation problems that were not covered by the training dataset. Having an extremely accurate image segmentation would be very beneficial. However, it is not clear if we are yet at that point. For example, consider Fig. 3 which shows the output from the well-known FCN approach in the lower image.

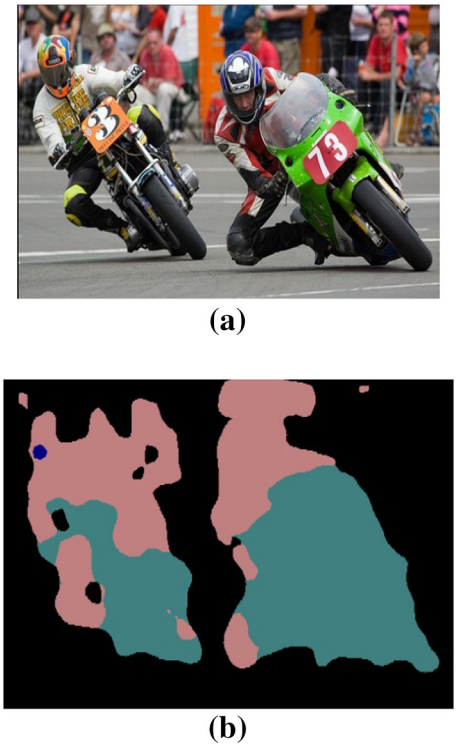

Fig. 3 **a** Original image. **b** Example of automatic image segmentation

Note that the segmentation has difficulties with the audience members and also the objects in the foreground. In some cases, the semantic segmentation extends beyond the motorcycle to the leg of the rider. In the general case, this means that using segmentations also requires understanding the effect that the errors will have on the entire system.

6 Conclusions

Image segmentation has made significant advances in recent years. Recent work based largely on deep learning techniques which has resulted in groundbreaking improvements in the accuracy of the segmentations (e.g., currently reported over 79% (mIOU) on the PASCAL VOC-2012 test set [44]). Because image segmentations are a mid-level representation, they have the potential to make major contributions across the wide field of visual understanding from image classification to image synthesis; from object recognition to object modeling; from high-performance indexing to relevance feedback and interactive search.

References

1. Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation and recognition. Comput Vis Image Underst 115(2):224–241
2. Sonka M, Hlavac V, Boyle R (2014) Image processing, analysis, and machine vision. Cengage Learning, Boston, USA. https://en.wikipedia.org/wiki/Cengage
3. Thomee B, Huiskes MJ, Bakker E, Lew MS (2008) Large scale image copy detection evaluation. In: MIR
4. Chatfield K, Arandjelović R, Parkhi O, Zisserman A (2015) On-the-fly learning for visual search of large-scale image and video datasets. Int J Multimed Inform Retr 4(2):75–93
5. Mallik A, Chaudhury S (2012) Acquisition of multimedia ontology: an application in preservation of cultural heritage. Int J Multimed Inform Retr 1(4):249–262
6. Atmosukarto I, Shapiro LG (2013) 3D object retrieval using salient views. Int J Multimed Inform Retr 2(2):103–115
7. Sebe N, Lew MS, Huang TS (2004) The state-of-the-art in human-computer interaction. In: HCI Workshop
8. Yu P, Qin AK, Clausi DA (2012) Unsupervised polarimetric SAR image segmentation and classification using region growing with edge penalty. IEEE Trans Geosci Remote Sens 50(4):1302–1317
9. Patil DD, Deore SG (2013) Medical image segmentation: a review. Int J Comput Sci Mobile Comput 2(1):22–27
10. Carson C, Belongie S, Greenspan H, Malik J (2002) Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans Pattern Anal Mach Intell 24(8):1026–1038
11. Tu Z, Chen X, Yuille AL, Zhu SC (2005) Image parsing: unifying segmentation, detection, and recognition. Int J Comput Vision 63(2):113–140
12. Lew M, Bakker E, Sebe N, Huang T (2007) Human-computer intelligent interaction: a survey, In: HCI 2007, LNCS 4796, Springer, Berlin
13. Ilea DE, Whelan PF (2011) Image segmentation based on the integration of colour-texture descriptors—a review. Pattern Recogn 44(10–11):2479–2501
14. Geman S (1984) Stochastic relaxation, Gibbs distribution, and Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741
15. Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell 33(5):898–916
16. Aly AA, Deris SB, Zaki N (2011) Research review for digital image segmentation techniques. Int J Comput Sci Inform Technol 3(5):99
17. Khan MW (2014) A survey: image segmentation techniques. Int J Future Comput Commun 3(2):89
18. Vantaram SR, Saber E (2012) Survey of contemporary trends in color image segmentation. J Electron Imaging 21(4):040901-1-040901-28
19. Zuva T, Olugbara OO, Ojo SO, Ngwira SM (2011) Image segmentation, available techniques, developments and open issues. Can J Image Process Comput Vis 2(3):20–29
20. Muthukrishnan R, Radha M (2011) Edge detection techniques for image segmentation. Int J Comput Sci Inform Technol 3(6):259
21. Caesar H, Uijlings J, Ferrari V (2016) Region-based semantic segmentation with end-to-end training. In: ECCV
22. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR
23. Uijlings JRR, Van De Sande KEA, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
24. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: NIPS
25. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: ICLR
26. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: CVPR
27. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR
28. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60(2):91–110
29. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: CVPR
30. Hariharan B, Arbeláez P, Girshick R, Malik J (2014) Simultaneous detection and segmentation. In: ECCV
31. Arbeláez P, Pont-Tuset J, Barron J T, Marques F, Malik J (2014) Multiscale combinatorial grouping. In: CVPR
32. Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: CVPR
33. Dai J, He K, Sun J (2015) Convolutional feature masking for joint object and stuff segmentation. In: CVPR
34. He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: ECCV
35. Liu S, Qi X, Shi J, Zhang H, Jia J (2016) Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation. In: CVPR
36. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. https://arxiv.org/abs/1703.06870
37. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS
38. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: CVPR
39. Eigen D, Fergus R (2015) Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV
40. Liu Y, Guo Y, Lew MS (2017) On the exploration of convolutional fusion networks for visual recognition. In: MMM
41. Dai J, Li Y, He K, Sun J (2016) R-FCN: Object Detection via region-based fully convolutional networks. In: NIPS
42. Noh H, Hong S, Han B (2015) Learning deconvolution network for semantic segmentation. In: ICCV
43. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2015) Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR
44. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille A L (2017) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In: IEEE transactions on pattern analysis and machine intelligence
45. Krähenbühl P, Koltun V (2011) Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS
46. Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr P H (2015) Conditional random fields as recurrent neural networks. In: ICCV
47. Liu W, Rabinovich A, Berg AC (2016) Parsenet: Looking wider to see better. In: ICLR Workshop
48. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: CVPR

49. Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In: ICLR
50. Lin G, Shen C, van den Hengel A, Reid I (2016) Efficient piecewise training of deep structured models for semantic segmentation. In: CVPR
51. Lin G, Shen C, Van Den Hengel A, Reid I (2017) Exploring context with deep structured models for semantic segmentation. In: IEEE transactions on pattern analysis and machine intelligence
52. Dai J, He K, Sun J (2015) Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV
53. Papandreou G, Chen LC, Murphy KP, Yuille AL (2015) Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: ICCV
54. Khoreva A, Benenson R, Hosang J, Hein M, Schiele B (2017) Simple does it: weakly supervised instance and semantic segmentation. In: CVPR
55. Pinheiro PO, Collobert R (2015) From image-level to pixel-level labeling with convolutional networks. In: CVPR
56. Pathak D, Shelhamer E, Long J, Darrell T (2015) Fully convolutional multi-class multiple instance learning. In: ICLR Workshop
57. Pathak D, Krahenbuhl P, Darrell T (2015) Constrained convolutional neural networks for weakly supervised segmentation. In: ICCV
58. Bearman A, Russakovsky O, Ferrari V, Fei-Fei L (2016) What's the point: Semantic segmentation with point supervision. In: ECCV
59. Wei Y, Liang X, Chen Y, Jie Z, Xiao Y, Zhao Y, Yan S (2016) Learning to segment with image-level annotations. Pattern Recogn 59:234–244
60. Saleh F, Akbarian MSA, Salzmann M, Petersson L, Gould S, Alvarez JM (2016) Built-in foreground/ background prior for weakly-supervised semantic segmentation. In: ECCV
61. Shimoda W, Yanai K (2016) Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: ECCV
62. Wei Y, Liang X, Chen Y, Shen X, Cheng MM, Feng J, Zhao Y, Yan S (2016) STC: A simple to complex framework for weakly-supervised semantic segmentation. In: IEEE transactions on pattern analysis and machine intelligence
63. Jin B, Ortiz-Segovia MV, Süsstrunk S (2017) Webly supervised semantic segmentation. In: CVPR