Chinese Conference on Computer Vision (CCCV), October 2017

Deep Context Convolutional Neural Networks for Semantic Segmentation

Wenbin Yang[1,2], Quan Zhou[1,2,⋆], Yawen Fan[1], Guangwei Gao[3], Songsong Wu[4], Weihua Ou[5], Huimin Lu[6], Jie Cheng[7], and Longin Jan Latecki[8]

[1] National Engineering Research Center of Communications and Networking, Nanjing University of Posts & Telecommunications, Nanjing, China.
[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou 350121, China
[3] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing Univ. of Sci. & Tech., Nanjing, China.
[4] School of Automation, Nanjing Univ. of Posts & Telecom., Nanjing, China.
[5] School of Big Data and Computer Science, Guizhou Normal Univ., Guiyang, China.
[6] Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Kitakyushu, Japan.
[7] Huawei Technologies Co. Ltd., ShenZhen, China.
[8] Department of Computer and Information Sciences, Temple University, Philadelphia, USA.

Abstract. Recent years have witnessed the great progress for semantic segmentation using deep convolutional neural networks (DCNNs). This paper presents a novel fully convolutional network for semantic segmentation using multi-scale contextual convolutional features. Since objects in natural images tend to be with various scales and aspect ratios, capturing the rich contextual information is very critical for dense pixel prediction. On the other hand, with going deeper of the convolutional layers, the convolutional feature maps of traditional DCNNs gradually become coarser, which may be harmful for semantic segmentation. According to these observations, we attempt to design a deep context convolutional network (DCCNet), which combines the feature maps from different levels of network in a holistic manner for semantic segmentation. The proposed network allows us to fully exploit local and global contextual information, ranging from an entire scene, though sub-regions, to every single pixel, to perform pixel label estimation. The experimental results demonstrate that our DCCNet (without any postprocessing) outperforms state-of-the-art methods on PASCAL VOC 2012, which is the most popular and challenging semantic segmentation dataset.

1 Introduction

Image semantic segmentation is a classic and challenging visual task in the field of computer vision. It aims to assign a semantic label to each image pixel to

[⋆] Quan Zhou is the corresponding author.

achieve object recognition and segmentation tasks synchronously. Semantic segmentation is associated with many high-level vision tasks, including image classification [1–3], object recognition [4–6] and object detection [7–9].

Typically, the previous semantic segmentation models usually first extract local appearance features, such as intensities, colors, gradients and textures, to describe different object instances. Then these hand-craft features are fed into the well trained classifiers to identify the category label for each image pixel, including regression boosting [10, 11], random forests [12], or support vector machines [13]. Thereafter, great improvement have been achieved by incorporating contextual formulation [10, 14] and scene global structured predictions [15, 16] using conditional random field (CRF) [10, 15, 17] or markov random field (MRF) [16, 18]. However, the performance of these systems has always been compromised by the limited expressive power of the hand-craft features.

In recent years, DCNNs have gained a lot of attention and become very popular in the computer vision community. Due to its superiority to modeling high-level visual concepts, DCNNs substantially advance the state-of-the-art results for the task of semantic segmentation [17, 19, 20]. As the pioneer work, LeCun et. al. [21] employ the DCNNs at multiple image resolutions to compute image features, resulting in smooth predictions using a segmentation tree as postprocess. Another elegant work was proposed in [7], where the bounding box proposals and masked regions are used as inputs to train a DCNN. In the stage of classification, object shape information is taken into account within the trained DCNN. Similarly, the DCNN models can be also trained based on different image representation, such as superpixels [22]. In this work, the authors extracted the zoom-out spatial features, which are embedded into DCNNs to classify a superpixel. Although these methods can benefit from the delineated boundaries produced from a good segmentation, the object accurate shapes may be not always recovered well when there are some errors in segmentation results.

An alternative approach for dense pixel estimation relies on fully convolutional networks (FCN), where an end-to-end learning paradigm is adopted to train DCNNs [17, 19, 23]. Motivated from the DCNNs for image classification task, these methods directly target on estimating category-level per-pixel labels. The most representative work is [19], where the last fully connected layers of the DCNN are transformed into convolutional layers. In [17], Chen et al. proposed to learn a DeepLab model for semantic segmentation, where the receptive fields with different scales are employed using atrous convolution in deep convolutional networks. Then the final consistent segmentation results are enhanced using an independent fully connected CRF. The major limitation of FCN-based methods lies in the fact that the operation of max pooling and sub-sampling reduce feature map resolution, leading to the loss of spatial statistics for object instance. Therefore, Noh et al. [23] and Vijay et al. [24] proposed a coarse-to-fine structure with deconvolution network to learn the segmentation mask, yet with the cost of huge number of memory requirement and computing time.

In order to further advance the performance of semantic segmentation, the context cues are widely employed for image semantic segmentation [4, 11, 14, 25].

However, we found that this major issue is not well addressed in current FCN-based models. Due to the repeated operation of max-pooling at successive layers of DCNNs, the spatial resolution is significantly reduced in feature maps when the DCNN is employed in a fully convolutional fashion [17, 19]. Motivated by spatial pyramid pooling [4, 25], in this paper, we present a *deep context convolutional network* (DCCNet) to explore contextual cues using the feature maps from different convolutional layers. Preciously, we exact all intermediate feature maps and fuse them together for pixel classification. Under this paradigm, the local and global contextual clues, ranging from en entire scene, through sub-regions, to every single pixel, are taken into account *jointly* to assign semantic label for each pixel. We compare the performance of our model with the mainstream models, which include contextual formulation [14], FCN-based approaches [17, 19, 21, 22, 26], and deconvolution network [24]. These are top-ranked models that previous studies have shown to significantly improve the results on PASCAL VOC 2012 dataset. In summary, the main contributions of this paper are twofolds:

- We propose a DCCNet to capture multi-scale context features in the manner of FCN for dense pixel prediction.
- We evaluate our DCCNet on PASCAL VOC 2012 dataset and achieve the state-of-the-art results for semantic segmentation.

2 The Architecture of DCCNet

In this section, we elaborate on architectural details of our DCCNet to investigate the multi-scale context information.

As shown in Fig. 1, we define our DCCNet based on VGG 16-layer network, which consists of four basic components, including convolution, rectified linear unit (ReLU), pooling (downsampling) and deconvolution (upsampling). In the main stream structure, we borrow the network architecture widely used in FCN-based model [19], where the final fully connected layers are transfered to convolution layers. It consists of the repeated convolution with 3×3 filter kernels, followed by a ReLU and a 2×2 max pooling operation with stride 2 for downsampling. The resolution of feature maps are reduced to $1/2^N$ of the original one after N max pooling operation. Here we set $N = 5$, leading to the 1/2, 1/4, 1/8, 1/16 and 1/32 of the original resolution, and denote the final three ones as DCCNet-8s, DCCNet-16s, DCCNet-32s, respectively.

Similar to previous FCN-based approaches [17, 19, 23], the major limitation of our main stream structure lies in the spatial statistics of pixels is gradually lost with going deeper of our DCCNet. On the other hand, compared with shallower layers, the deeper ones have large receptive fields and are able to see more pixels. Intuitively, the feature maps with different scales can provide sufficient spatial statistics and contextual cues, which complement each other to get more reliable predictions. To this end, we integrate the main stream with two additional streams from the max pooling layers of DCCNet-8s and DCCNet-16s, as shown in Fig. 1. Once augmented, DCCNet allows us to fuse predictions from three streams that are learned jointly in an end-to-end architecture. More specifically,

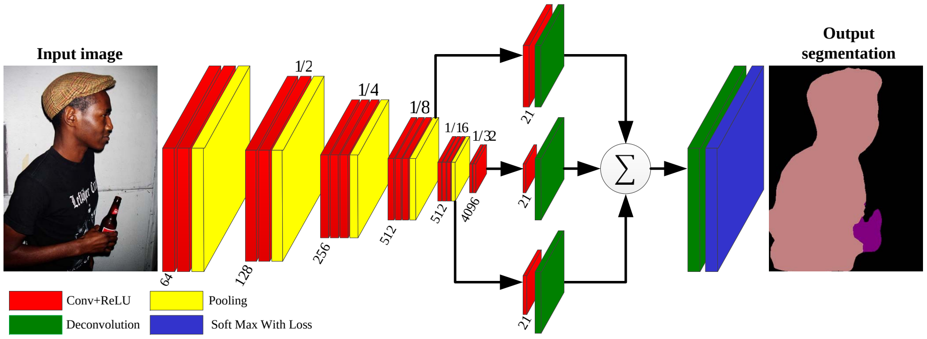

Fig. 1. Overview of our proposed DCCNet. Given an input image, we first use VGG-16 network to produce the hierarchical feature maps of intermediate pooling layers, which carry both local and global contextual cues and spatial statistics within different scales. Then the assocaited score maps are upsampled and concatenated to get the final per-pixel prediction. (best viewed in color)

the final layers of these three streams are first convoluted with a 1 × 1 filter kernel to map them to three 21-component score maps, representing the confidence for each individual classes in PASCAL VOC dataset. However, these layers are with different resolution, they are thus required to be aligned by scaling and cropping. The deconvolution layers in our DCCNet utilizes the nonlinear upsampling to scale score maps to the resolution with respect to DCCNet-8s, where the upsampling filter kernels are learned with the initialized weights of bilinear interpolation [19]. Subsequently, a cropping operation is performed. Cropping removes any portion of the upsampled layer which extends beyond the other layer, resulting in layers of equal dimensions for exact fusion. Finally, the three score maps are fused to a single 21-component score map, which is further upsampled to obtain the semantic output with the same resolution of original image.

At first glance of our DCCNet, it might look similar to the skip version of FCN [19], but in fact their network architectures are quite different. Essentially, FCN relies on *gradually* learning finer-scale prediction from lower layers in a stage-by-stage manner. That is, in each stage, the net is trained using the initialization of the previous stage net, where the contextual features are explored *independently*. In contrast, our DCCNet employs a all-in-once fashion to fuse the computed intermediate feature maps, where the contextual cues are investigated *jointly* to make final estimation. In addition, compared with stage-by-stage scheme used in FCN model, our all-in-once fashion results in computational efficiency and is less tedious in training process.

3 Experience Results

To validate the effectiveness of our method, we conducted several experiments on PASCAL VOC 2012 test dataset [27], which is widely used to evaluate the performance for semantic segmentation.

3.1 Dataset and Evaluation Metrics

The PASCAL VOC 2012 dataset [27] contains 21 category classes, including 20 categories for foreground object classes and additional one class for background. The original dataset contains 1464, 1449, and 1456 images for training, validation, and testing, respectively, where each image in the training and validation subset has accurate pixel-level annotated ground truth. The dataset is augmented by the extra annotations [28], resulting in 10582 training images.

In our experience, we report our results and compare with other baselines using the metrics in terms of pixel-level mean intersection-over-union (mIoU). It is commonly used to penalizes both over- and under-segmentation for semantic segmentation, which is defined as the ratio of true positives to the sum of true positive, false positive and false negative, averaged over all 21 object classes:

$$\frac{1}{\mathcal{C}}\frac{\sum_m N_{mm}}{\sum_n N_{mn} + \sum_n N_{nm} - N_{mm}} \tag{1}$$

where $\mathcal{C} = 21$ denotes the total number of different classes, N_{mn} is the number of pixels of category m labeled as class n.

3.2 Baselines

We selected 9 state-of-the-art models as baselines for comparison, including flexible segmentation graph (FSG, [14]), deep hierarchical parsing(DHP, [21]), fully-connected network (FCN, [19]), DeepLab network (DLN, [17]) and its variants (DLN† as DeepLab-Msc and DLN‡ as DeepLab-Msc-CRF), deep zoom-out feature (DZF, [22]), dilated convolution model (DCM, [26]), and convolution and deconvolution network (CDN, [23]).

3.3 Implementation Details

Our implementation is based on the public platform Caffe [29]. We minimize the soft max loss averaged over all image positions with stochastic gradient descent algorithm. The parameters of initialized network are borrowed from the pre-trained VGG-16 model using ImageNet dataset [30]. Then we fine-tune our DCCNet model using 10582 training images of PASCAL VOC 2012 dataset. Inspired by [17], we use the “poly” learning rate policy where the learning rate γ in iteration T equals to the base B multiplied by a factor:

$$\gamma = B \cdot (1 - \frac{T}{M})^{power} \tag{2}$$

where M denotes the total number of training iterations. The base learning rate is set as $B = 0.01$ and power is set as 0.9.

The performance can be further improved by increasing the iteration number. We found that our best performance is achieved when total iteration number is set to $M = 20K$, and any refinement of this parameter will result in no more significant improvement of performance.

Table 1. Per-class results on PASCAL VOC 2012 test set.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
DHP [21]	-	-	-	-	-	-	-	-	-	-	-
FSG [14]	-	-	-	-	-	-	-	-	-	-	-
DLN [17]	87.5	72.0	31.0	71.2	53.7	60.5	77.0	71.9	73.1	25.2	62.6
DLN† [17]	88.3	79.4	34.1	72.6	52.9	61.0	77.9	73.0	73.7	26.4	62.2
DZF [22]	89.8	81.9	35.1	78.2	57.4	56.5	80.5	74.0	79.8	22.4	69.6
DLN‡ [17]	92.6	80.4	36.8	77.4	55.2	66.4	81.5	77.5	78.9	27.1	68.2
FCN [19]	91.2	76.8	34.4	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5
DCM [26]	90.7	82.2	37.4	72.7	57.1	62.7	82.8	77.8	78.9	28.0	70.0
CDN [23]	92.7	85.9	42.6	78.9	62.6	66.6	87.4	77.8	79.5	26.3	73.4
DCCNet	93.2	84.1	39.0	82.1	67.7	78.4	87.4	83.4	85.8	38.2	77.2
Method	table	dog	horse	mbike	person	planet	sheep	sofa	train	tv	mIoU
DHP [21]	-	-	-	-	-	-	-	-	-	-	60.6
FSG [14]	-	-	-	-	-	-	-	-	-	-	64.4
DLN [17]	49.1	68.7	63.3	73.9	73.6	50.8	72.3	42.1	67.9	52.6	62.1
DLN† [17]	49.3	68.4	64.1	74.0	75.0	51.7	72.2	42.5	67.2	55.7	62.9
DZF [22]	53.7	74.0	76.0	76.6	68.8	44.3	70.2	40.2	68.9	55.3	64.4
DLN‡ [17]	52.7	74.3	69.6	79.4	79.0	56.9	78.8	45.2	72.7	59.3	67.1
FCN [19]	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	67.2
DCM [26]	51.6	73.1	72.8	81.5	79.1	56.6	77.1	49.9	75.3	60.9	67.6
CDN [23]	60.2	70.8	76.5	79.6	77.7	58.2	77.4	52.9	75.2	59.8	69.6
DCCNet	45.6	80.9	75.2	77.3	82.8	56.7	81.5	51.8	82.4	70.5	71.4

3.4 Overall Results

We report the results in Table 1, and compare with the baselines in terms of mIoU. The results clearly demonstrate that our DCCNet outperforms prior state-of-the-art methods, including the approaches using contextual formulation [14, 26], FCN-based models [17, 19, 21, 22], and deconvolution network [23]. Trained with only PASCAL VOC 2012 data, our DCCNnet achieves 71.4% mIoU among all 21 categories, and gets the highest accuracy on the classes of “bird”, “boat”, “bottle”, “bus”, “car”, “cat”, “chair”, “cow”, “dog”, “person”, “sheep”, “train” and “TV”. For fair comparison, we construct our DCCNet without any postprocessing to enhance the consistent semantic segmentation outputs. Even so, it is intriguing that our DCCNet is superior to the existing methods [17] that employ CRF for further improving performance. Another interesting result is our approach outperforms deconvolution network [23]. This is probably because the proposed DCCNet has more powerful generalization than [23] due to its simple network architecture.

Some visual pleasing results of simultaneous recognition and segmentation are shown in Fig. 2. Each example shows both the original image and the color coded output. Except the boundary pixels that exhibit relative higher confusion, nearly all pixels are correctly classified. It is evident that our DCCNet can handle

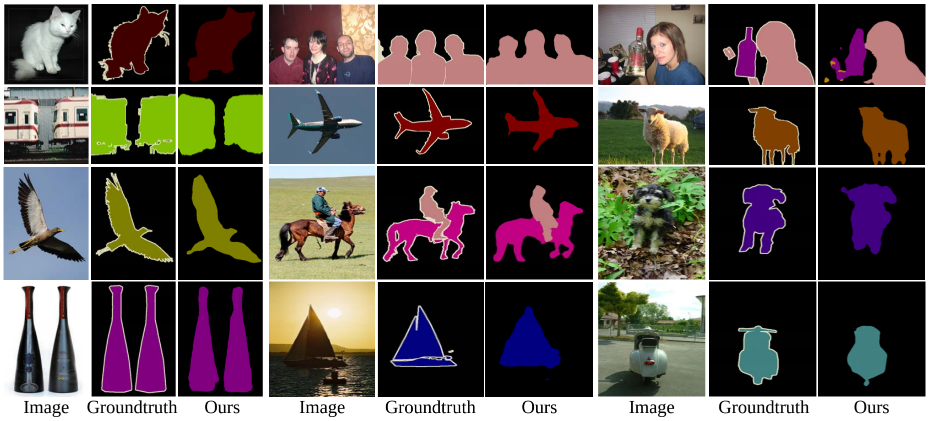

Fig. 2. Some visual examples of our semantic segmentation outputs. In each column, we show the original image, corresponding ground truth and our segmentation results. For clarity, we use different colors to denote different categories. (best viewed in color)

large appearance variations of object classes ("person", "dog", and "boat", etc.) and efficiently prohibit the clutter background (see the last four visual examples in the right column).

4 Conclusion and Future Work

This paper describes a DCCNet to explore the multi-scale context information for semantic segmentation problem. Combining fine layers and coarse layers provides a more powerful representation with different receptive fields, allowing us to produce semantically accurate predictions and detailed segmentation maps. Our experimental results show that the proposed method advances the state-of-the-art results in the PASCAL VOC 2012 image semantic segmentation task.

Despite obtaining impressive results, we believe that even better results can be achieved by employing CRF models as well as [17, 31] does. We also hope to demonstrate the generality of our DCCNet for other visual tasks.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their valuable comments and suggestions. This work was partly supported by the National Science Foundation (Grant No. IIS-1302164), the National Natural Science Foundation of China (Grant No. 61401228, 61402238, 61402122, 61571240, 61501247, 61501259, 61671253), China Postdoctoral Science Foundation (Grant No. 2015M581841), Natural Science Foundation of Jiangsu Province (Grant No. BK20150849, BK20160908), Postdoctoral Science Foundation of Jiangsu

Province (Grant No. 1501019A), Open Research Fund of Key Lab of Ministry of Education for Broadband Wireless Communication and Sensor Network Technology (Nanjing University of Posts and Telecommunications) (Grant No. NYKL2015012), Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (Grant No. MJUKF201710), Open Fund Project of Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education (Nanjing University of Science and Technology) (Grant No. JYB201709, JYB201710), and Natural Science Foundation of Guizhou Province (Grant No.[2017]1130).

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778
3. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. (2015) 1–9
4. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE TPAMI **37** (2015) 1904–1916
5. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV **104** (2013) 154–171
6. Girshick, R.: Fast r-cnn. In: ICCV. (2015) 1440–1448
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014) 580–587
8. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: CVPR. (2014) 2147–2154
9. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. (2015) 91–99
10. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. IJCV **81** (2009) 2–23
11. Tu, Z., Bai, X.: Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE TPAMI **32** (2010) 1744–1757
12. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: CVPR. (2008) 1–8
13. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: ICCV. (2009) 670–677
14. Zhou, Q., Zheng, B., Zhu, W., Latecki, L.J.: Multi-scale context for scene labeling via flexible segmentation graph. PR **59** (2016) 312–324
15. Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS. (2011) 4–10
16. Carreira, J., Sminchisescu, C.: Cpmc: Automatic object segmentation using constrained parametric min-cuts. IEEE TPAMI **34** (2012) 1312–1328
17. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)

18. Zhou, Q., Zhu, J., Liu, W.: Learning dynamic hybrid markov random field for image labeling. IEEE TIP **22** (2013) 2219–2232
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE TPAMI **39** (2017) 640–651
20. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR. (2016) 3640–3649
21. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE TPAMI **35** (2013) 1915–1929
22. Mostajabi, M., Yadollahpour, P., Shakhnarovich, G.: Feedforward semantic segmentation with zoom-out features. In: CVPR. (2015) 3376–3385
23. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV. (2015) 1520–1528
24. Badrinarayanan, V., Alex, K., Roberto, C.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 (2015)
25. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178
26. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
27. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111** (2015) 98–136
28. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV. (2011) 991–998
29. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACMMM. (2014) 675–678
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
31. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: ICCV. (2015) 1529–1537