A Overview About Image Segmentation

Tao Hu
Department of Computer Science
Peking University
taohu@pku.edu.cn

Abstract

In this report, I summarize some traditional method in segmentation, and lay more emphasize on the neural network method, which focus on large receptive field, atrous convolution, multi scale fusion etc. finally two promising direction with respect to geometry information combination and graph structure combination is proposed.

1 Introduction

image segmentation have become an important task of current computer vision, which has been applied to many interesting fields: autonomous driving[9], medical imaging[17], to name a few. Convolutional Neural Networks(CNN) have pushed the performance of computer vision systems from coarse-grained to fined-grained task: classification[31][14][32][16], detection or localization[12][29][28], semantic segmentation[26], and instance segmentation[11][22]. most of them are designed as a end-to-end manner that delivered strikingly better result than those traditional hand-crafted feature-based method.

Several years ago, the CNN is only used in the classification task that suits for single value classification. Nowadays, the most important fundamental deep learning techniques for image segmentation derive from the Fully Convolutional Network(FCN)[26], which mainly use Deconvolution[27] to upsample the feature map to output a same size predict. besides the FCN architecture, a novel encoder-decoder framework named Segnet[1] is invented, it implements upsampling by remember the pooling indices to as it do downsampling by pooling. morever the Bayesian Segnet[18] utilize stochastics to enhance the accuracy. Conditional Random Field(CRF)[21] is often deployed to smooth a noise segmentation map, which serves as a post-process procedure for image segmentation, later another significant work unifies the FCN and CRF as a systematic framework is CRFasRNN by Shuai Zheng et al.[35]. recently, Liang-Chieh Chen et al. put forward a solid work named deeplab[3], it comes up a new convolution method named "Atrous" Convolution that greatly enlarge the receptive field of the CNN framework.

Contributions: Firstly, there appeared many methods about image segmentation, however, nobody organized the development of image segmentation field which would significantly facilitate our understanding and further improvement on this lively domain. Secondly, not only image segmentation but also instance segmentation which is the ultimate goal of the computer vision is included in this overview. Finally, several promising development directions are summarized.

In the following section 2, I will summary some methods before neural networks. several neural networks based method will be included in section 3. in section 4, I will demonstrate the mainstream dataset frequently utilized for segmentation accuracy and performance comparison. finally I will propose several promising development directions of the active field which will greatly convenient for the related research's advance.

Author Info: Taohu ,Peking University

2 Deep Neural Networks Method

Before the surge of Deep Learning, as a survey[33] said, the traditional method can be classified into Region Based Method and Edge Based Method.

Normalized Cuts[30] means a optimization function that minimizes the segmentation cost based on the Laplace Matrix.

Region Growing[8] examines neighboring pixels of initial seed points and determines whether the pixel neighbors should be added to the region. The process is iterated on, in the same manner as general data clustering algorithms.

Thresholding methods replace each pixel in an image with a black pixel if the image intensity $I_{i,j}$ is less than some fixed constant T (that is,$I_{i,j} < T$), or a white pixel if the image intensity is greater than that constant. a suitable threshold is critical for the segmentation result.

Edge-based segmentation represents a large group of methods based on information about edges in the image. many traditional edge detection method can be employed such as Roberts, Prewitt, Sobel. after the detection, the detection can be utilized to further obtain the segmentation result.

After the surge of Deep Learning, significant change happened everywhere in the computer vision. There are the following important things that greatly boost the advance of image segmentation field:

- the imagenet competition which fast boost the advance of deep learning.
- deeper network such as resnet[14], densenet[16] further improve the effect of deep learning.
- the come up of Fully Convolutional Network(fcn)[26] firstly make the pixel-level prediction possible, which also significant decrease the network parameters by removing the fully connected layer, what's more, fcn make the network insensitive to the input image size.
- atrous convolution[3] greatly enlarge the receptive field.
- Deformable Convolution Network.

2.1 Important Factors

2.1.1 Fully Convolutional Networks

The network structure is Fig1. FCN is the first framework that is trained end-to-end, pixel-to-pixel semantic segmentation. the network upsampling is realized by Deconvolution. we need to notice the deconvolution in FCN is not learnable, because it's initially set according to the bilinear interpolation. the FCN-8s need fine tuning twice based on FCN-32s and FCN-16s. the auther makes a padding of 100 to the input image so that the feature map is large enough after four times of downsampling(16X).

the more high layer combines with, the better final accuracy is. therefore, FCN-8s performs best compared with FCN-16s,FCN-32s which utilize less high layer than FCN-8s. The FCN network opens the wonderful gate of pixel-level prediction, and significantly boosts the later progress which make use of multi scale fusion, edge information, high receptive field.

2.1.2 Deeper Networks

Kaiming He et al. proposed the Resnet with identity mapping trick that remarkably decrease the gradient vanish. The model's description ability also increased as deeper networks. recently, G Huang et al. come up with a new style network that works better than resnet. the network structure is in Figure9. both of the resnet and densenet can be regarded as a basic model for classification. we can transfer them into fully convolutional network and append some upsampling layer such as deconvolution layer so that them can solve the image segmentation problem.

the classification model usually acts as the foundation of image segmentation model. once a better classification model appears, there must would emerge a better segmentation model based on them.

2.1.3 atrous convolution

Atrous convolution allows us to explicitly control the resolution at which feature responses are computed within Deep Convolutional Neural Networks. It also allows us to effectively enlarge the

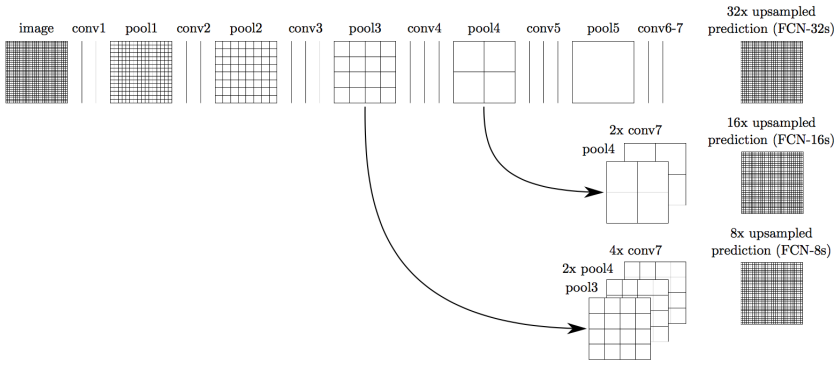


Figure 1: learn to combine coarse,high layer information with fine, low layer information. Pooling and prediction layers are shown as grids that reveal relative spatial coarseness, while intermediate layers are shown as vertical lines. First row (FCN-32s): upsamples stride 32 predictions back to pixels in a single step. Second row (FCN-16s): Combining predictions from both the final layer and the pool4 layer, at stride 16, lets our net predict finer details, while retaining high-level semantic information. Third row (FCN-8s): Additional predictions from pool3, at stride 8, provide further precision.

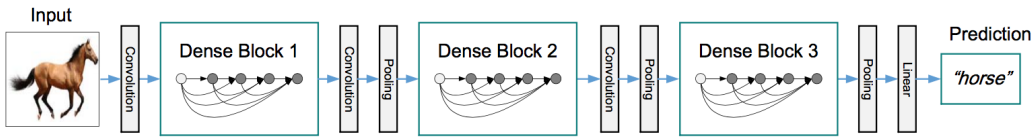


Figure 2: A deep DenseNet with three dense blocks. The layers between two adjacent blocks are referred to as transition layers and change feature map sizes via convolution and pooling.

field of view of filters to incorporate larger context without increasing the number of parameters or the amount of computation. the atrous convolution example is in Fig3.

2.1.4 deformable convolution networks

the deformable convolution networks(DCN)[6] is a special case of atrous convolution. atrous convolution hole is regular grid, however, DCN can learning the bias direction adaptively. the difference of DCN and simple convolution is Fig4. most importantly, DCN can be regarded as pluggable component of neural network, which can be applied in diverse computer vision task as long as it needs convolution operation.

as the traditional CNN are inherently limited to model geometric transformations due to the fixed geometric structures in its building modules. The DCN also is kind of attempt that try to make the network adapt to more complicated geometric transformations, which nearly bring no extra burden. therefore, it leads a possible direction that tries to make use of more geometry information. The image segmentation task will obtain many gains from DCN.

2.2 typical neural networks

2.2.1 SegNet

The most important feature of SegNet[1] is symmetry. in the unpooling layer, the network will remember pooling indices of the symmetric pooling layer. the detail architecture is in Fig5.

the author came up with a new SegNet with Bayesian Inference[19] which further boost the accuracy. Monte Carlo sampling with dropout at test time can generate a posterior distribution of pixel class labels, which will improve the segmentation performance by a margin of 2% as the author claimed.

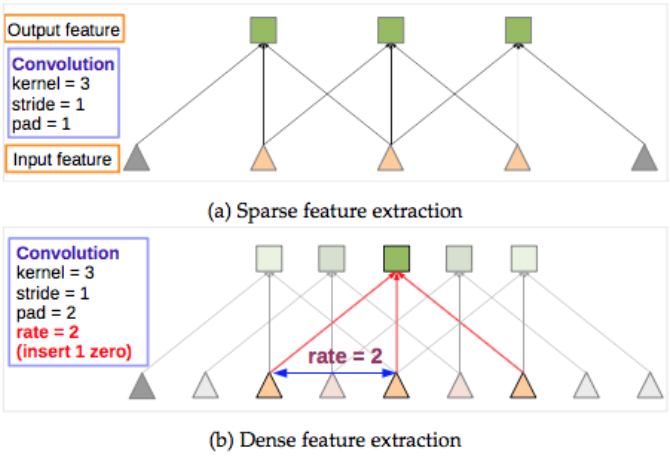


Figure 3: Illustration of atrous convolution in 1-D. (a) Sparse feature extraction with standard convolution on a low resoution input feature map. (b) Dense feature extraction with atrous convolution with rate r = 2, applied on a high resolution input feature map.

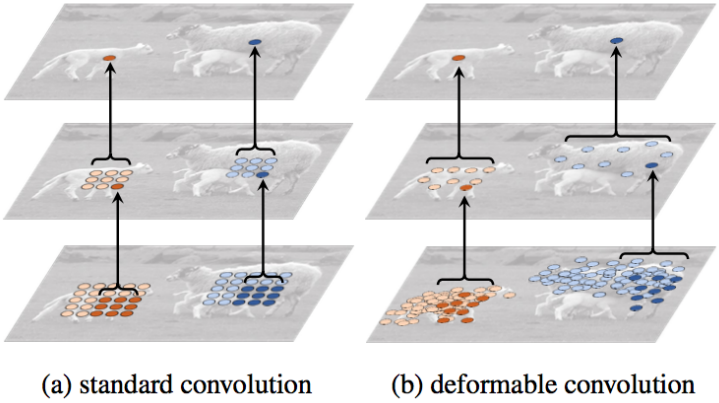


Figure 4: Illustration of atrous convolution in 1-D. (a) Sparse feature extraction with standard convolution on a low resoution input feature map. (b) Dense feature extraction with atrous convolution with rate r = 2, applied on a high resolution input feature map.

Obviously, if Monte Carlo sampling is used, the inference time will increase linearly with the sampling times.

2.2.2 Deeplab

Deeplab[3] is a very solid work which is the base model of many current models because that: **(1).** a novel atrous convolution is proposed. **(2).**, Deeplab further improve the network by padding, without the large padding of 100 on input image as FCN[26]. **(3).** a typical post process method called Fully Connected CRF is ingeniously applied to image segmentation.

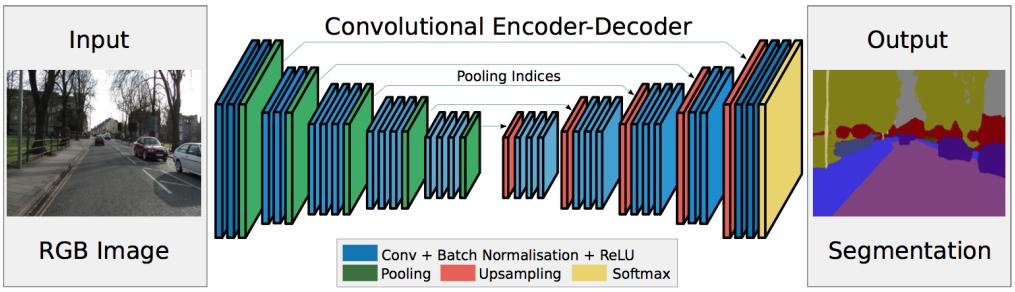


Figure 5: An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

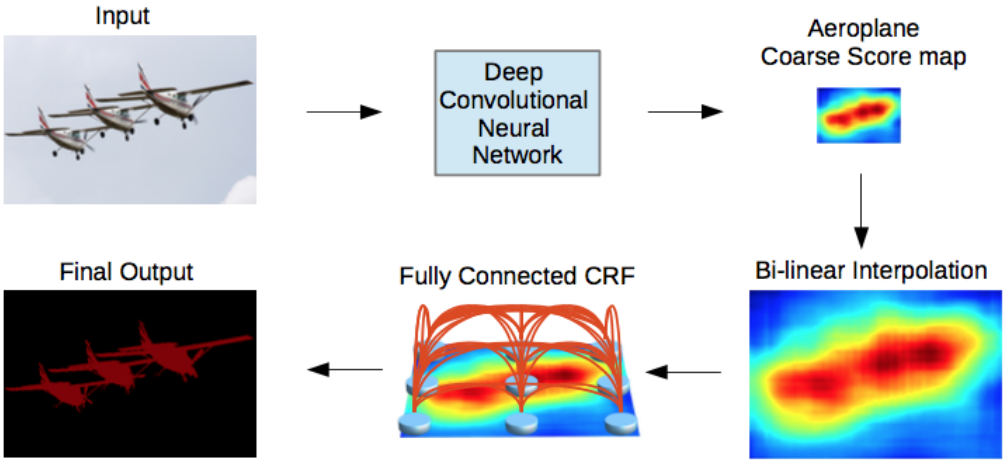


Figure 6: Model Illustration. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries.

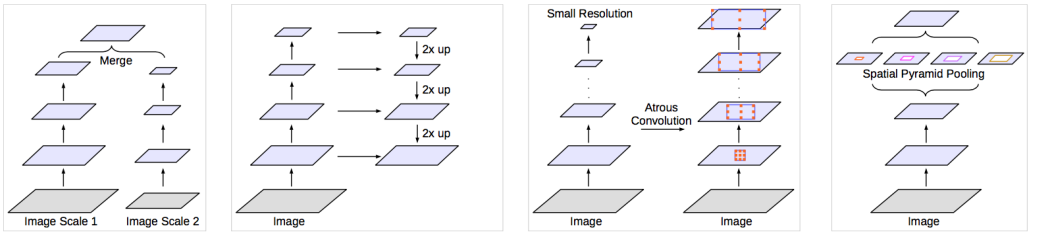
2.2.3 PSPNet

3 Frequently Used Dataset

3.1 PASCAL VOC 2012

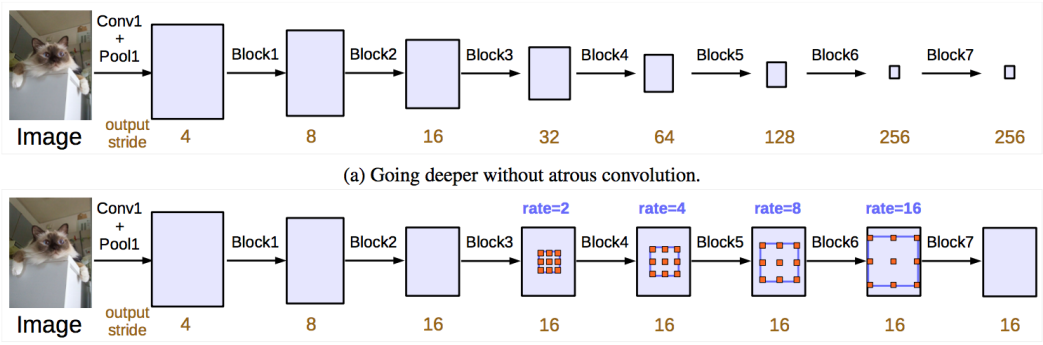
The Pascal VOC segmentation benchmark[10] includes 20 foreground object class and one background class. the original data comprises 1,464 training data, 1,449 validation data, and 1,456 test data,which are all pixel-level labeled. Usually, the data is augmented by the method[13], resulting in 10,582 training data.

For both types of segmentation image, index 0 corresponds to background and index 255 corresponds to 'void' or unlabeled. the segmentation is categorized into class segmentation and instance segmentation, which both are included in Pascal VOC 2012. a labeled image is demonstrated in Fig10. the white border indicates that these pixel's label is ambiguous so that you should not make utilize of them.



(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

Figure 7: Alternative architectures to capture multi-scale context.



(a) Going deeper without atrous convolution.

(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.

Figure 8: Cascaded modules without and with atrous convolution.

3.2 Cityscapes

Cityscapes[5] is recently released large-scale dataset based on road scenes, which contains pixel-level annotations of 5000 images from 50 different cities and several different seasons. the size of labeled image is 1024-by-2048. Following the evaluation protocols[5]. 19 semantic labels (belonging to 7 super categories: ground, construction, object, nature, sky, human, and vehicle) are used for evaluation (the void label is not considered for evaluation). the detailed class definitions is in Table 1. the evaluation includes Pixel-level semantic labeling and Instance-level semantic labeling.

class segmentation and instance segmentation annotation both are provided. 5000 fine annotation and coarser polygonal annotations for a set of 20000 images are given.

Table 1: Class Definitions

Group	Classes
flat	$road, sidewalk, parking^{+}, railtrack^{+}$
human	$person^{*}, rider^{*}$
vehicle	$car^{*}, truck^{*}, bus^{*}, onrails^{*}, motorcycle^{*}, bicycle^{*}, caravan^{*+}, trailer^{*+}$
construction	$building, wall, fence, guardrail^{+}, bridge^{+}, tunnel^{+}$
object	$pole, polegroup^{+}, trafficsign, trafficlight$
nature	$vegetation, terrain$
sky	sky
void	$ground^{+}, dynamic^{+}, static^{+}$

* Single instance annotations are available. However, if the boundary between such instances cannot be clearly seen, the whole crowd/group is labeled together and annotated as group, e.g. car group.
+ This label is not included in any evaluation and treated as void (or in the case of license plate as the vehicle mounted on).

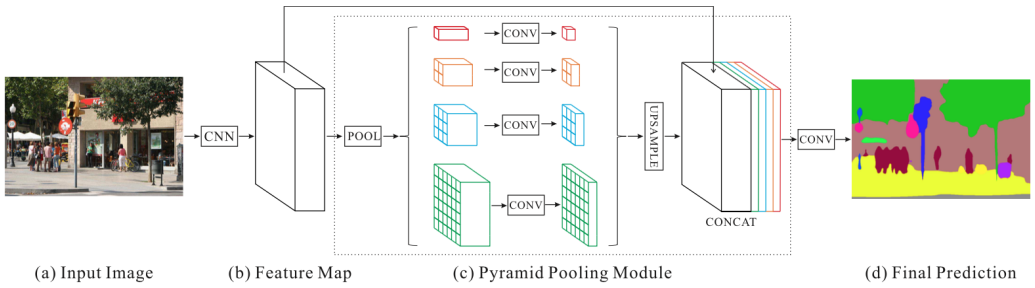


Figure 9: Overview of proposed PSPNet. Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatena- tion layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d).

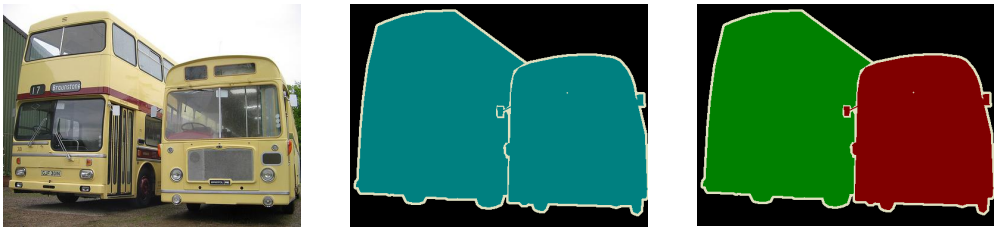


Figure 10: an annotation example of Pascal VOC 2012. The left image is original image, the middle is class-level segmentation, the right is object segmentation.

3.3 COCO

COCO[25] is a new image recognition, segmentation, and captioning dataset which contains 123,287 images, 886,284 instances. in the segmentation task, we mainly utilize the object segmentation labels.

3.4 Measure of Performance

To assess the performance, we rely on the standard Jaccard Index, commonly known as the Pascal VOC intersection-over-union metric IoU = $\frac{TP}{TP+FP+FN}$,where TP, FP and FN are the numbers of true positive, false positive, and false negative pixels, respectively, determined over the whole test set.

In Cityscapes[], it has two semantic granularities, i.e. classes and categories, we report two separate mean performance scores:$IoU_{category}$ and IoU_{class}. In either case, pixels labeled as void do not contribute to the score.

It is well-known that the global IoU measure is biased toward object instances that cover a large image area. In street scenes with their strong scale variation this can be problematic. Specifically for traffic participants, which their strong scale variation this can be problematic. Specifically for traffic participants, which are the key classes in our scenario, we aim to evaluate how well the individual instances in the scene are represented in the labeling. To address this, we additionally evaluate the semantic labeling using an instance-level intersection-over-union metric $iIoU = \frac{iTP}{iTP+FP+iFN}$. Again iTP, FP, and iFN denote the numbers of true positive, false positive, and false negative pixels, respectively. However, in contrast to the standard IoU measure, iTP and iFN are computed by weighting the contribution of each pixel by the ratio of the class' average instance size to the size of the respective ground truth instance. It is important to note here that unlike the instance-level task below, we assume that the methods only yield a standard per-pixel semantic class labeling as output. Therefore, the false positive pixels are not associated with any instance and thus do not

require normalization. The final scores, $iIoUcategory$ and $iIoU_{class}$ are obtained as the means for the two semantic granularities.

4 Several Promising development directions

After the outburst of Deep Learning application in lots of computer vision task. and the novel pixel-level prediction framework fcn[26] is proposed in 2015. the segmentation benchmark advanced dramatically. the following Table 44 shows the current best result of the mainstream benchmark. Recently several promising directions appeared which might accelerate the advance of image segmentation. graph data formulation[20][7][23] transfer the image to to high-dimensional irregular domains, such as social networks, brain connectomes or words' embedding, represented by graphs. the graph data application currently only us applied in classification task. a graph network example is demonstrated in Fig11. an generalization to pixel-level segmentation need to be taken great effort.

On the other side, Liang et al.[24] proposed a graph LSTM[15] which ingeniously combine the graph and LSTM via super-pixel. the author come up with an adaptive graph topography. the Graph LSTM is more naturally aligned with visual patterns in the image. and provide a more economic information propagation route. a confidence-driven scheme is proposed, which update the hidden and memory states of nodes progressively till all nodes are updated.

Table 2: Pascal VOC Best Result

Method	mIoU
Deep Layer Cascade(LC)	82.7
TuSimple	83.1
Large_Kernel_Matters	83.6
Multipath-RefineNet	84.2
ResNet-38_MS_COCO	84.9
PSPNet	85.4
DeepLabv3	85.7
CASIA_IVA_SDN	86
IDW_CNN	86.3

Table 3: Cityscapes Best Result

Method	mIoU
TuSimple	77.6
SAC-multiple	78.1
depthAwareSeg_RNN_ff	78.2
SegModel	79.2
TuSimple_Coarse	80.1
NetWarp	80.5
ResNet-38	80.6
PSPNet	81.2
motovis	81.3

Performance on Pascal VOC 2012 and Cityscapes.(Until 2017.7.1)

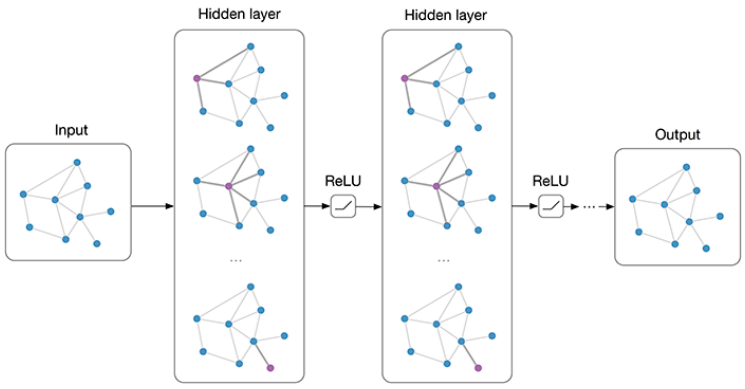


Figure 11: Multi-layer Graph Convolutional Network (GCN)[20] with first-order filters. This figure is best viewed in colour.

The current mainstream method of image segmentation includes: multi scale fusion[3][4], large receptive field[3], combine with the edge information[2], deeper networks[34]. hardly any paper utilize the geometry shape information of object. since there existed a study that summarized shape is the most important information when a child start to obtain the power of recognize diverse object. a child spend the first 9 months of his(her) lives learning to coordinate his(her) eyes to focus and perceive depth, color and geometry. It is not until 12 months when he(she) learn how to recognize

objects and semantics. This illustrates that a grounding in geometry is important to learn the basics in human vision.

To utilize the geometry information, the most direct method is to create diverse geometry data(by computer graph, simulation etc.) and feed them into the neural networks, however, this will lead to a extremely slow training process. according to nowadays knowledge, it's very difficult to strictly combine the geometry information, nevertheless, Dai et al.[6] recently raised a promising works including deformable convolution and pooling structure that is pluggable in most computer vision task.

5 Conclusion

In this report, I conclude various type of neural network method in image segmentation with regard to large receptive field, atrous convolution, multi scale fusion etc. at the end, two promising direction with respect to geometry information combination and graph structure combination is proposed. as the time is limited, in the future, I will attempt to design some novel network architecture based on graph structure.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2015.

[2] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4545–4554, 2016.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2016.

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks, 2017.

[7] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3844–3852, 2016.

[8] GM Espindola, Gilberto Câmara, IA Reis, LS Bins, and AM Monteiro. Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation. *International Journal of Remote Sensing*, 27(14):3035–3040, 2006.

[9] Andreas Ess, Tobias Mueller, Helmut Grabner, and Luc J. Van Gool. Segmentation-based urban traffic scene understanding. In *BMVC*. British Machine Vision Association, 2009.

[10] Mark Everingham, Ali Eslami, Luc Van Gool, K Christopher, I Williams, John Winn, Andrew Zisserman, et al. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98, 2015.

[11] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P. Murphy. Semantic instance segmentation via deep metric learning, 2017.

[12] Ross Girshick. Fast r-cnn, 2015.

[13] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 991–998. IEEE, 2011.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Gao Huang, Zhuang Liu, Kilian Q. Weinberger, and Laurens van der Maaten. Densely connected convolutional networks, 2016.

[17] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. Cnn-based segmentation of medical imaging data, 2017.

[18] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, 2015.

[19] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, 2015.

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[21] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. 2012.

[22] Ke Li, Bharath Hariharan, and Jitendra Malik. Iterative instance segmentation, 2015.

[23] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[24] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. *arXiv preprint arXiv:1603.07063*, 2016.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[27] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation, 2015.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2015.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[30] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[32] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2016.

[33] Nida M Zaitoun and Musbah J Aqel. Survey on image segmentation techniques. *Procedia Computer Science*, 65:797–806, 2015.

[34] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016.

[35] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. 2015.