# Template for Preparation of Manuscripts for *Tsinghua Science and Technology*

This template is to be used for preparing manuscripts for submission to *Tsinghua Science and Technology*. Use of this template will save time in the review and production processes and will expedite publication. However, use of the template is not a requirement of submission. Do not modify the template in any way (delete spaces, modify font size/line height, etc.).

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 0?/?? pp???-???
Volume 18, Number 3, June 2013

# Attention to Hard Examples: A Cascade Coarse to Fine Network Architecture for Semantic Segmentation

Wang Zhenyang, Deng Zhidong*, and Wang Shiyao

**Abstract:** Scene labeling, based on semantic segmentation, is a fundamental topic in computer vision. The goal is to assign each pixel in the image a category label. Convolutional neural networks, especially the fully convolutional neural networks, have attracted increasing attention on semantic segmentation due to the powerful capabilities of extracting hierarchical features. Since it is required to learn to make dense predictions for each pixel, a simple network is hardly to obtain considerable performances on different scenes. In this paper, we propose a novel semantic segmentation network called HMNet, which aims to pay more attention to the hard examples. The network has three branches, where the first branch produces coarse output predictions, and the second branch selects the hard examples which will be fed to the last branch. All above branches focus on their own objectives and collaboratively learn to predict from coarse to fine inference. Since the semantic segmentation dataset contains a large number of relatively easy samples and some hard ones, HMNet is encouraged to select these hard examples to make further predictions which is help to improve the final performance. In order to evaluate predicting performance of the proposed HMNet, we conduct experiments on two public datasets including Sift Flow and Stanford Background Dataset. We show that the three branches can be trained in an end-to-end manner and the experimental results show that compared to all existing models, our HMNet consistently yields the best performance, with accuracy of 91.6% and 89.7%, respectively.

**Key words:** semantic segmentation; online hard example mining; convolutional neural network

## 1 Introduction

Semantic segmentation, also known as scene labeling, is one of the fundamental research topics in computer vision. The goal of semantic segmentation is to identify and assign each pixel in the image with a category. This requires a complete understanding of the semantic information of the entire image. That means, for the testing image, it needs to predict the label of each object, and it is also very important to determine the boundary of each object in pixel level. Semantic segmentation has a strong application requirement in the field of environment perception and autonomous self driving car.

• Wang Zhenyang, Deng Zhidong, Wang Shiyao are with the Department of Computer Science, Tsinghua University, Beijing 100084, China. E-mail: crazycry2010@gmail.com, michael@tsinghua.edu.cn, sy-wang14@mails.tsinghua.edu.cn
* To whom correspondence should be addressed.
Manuscript received: 2017-09-20; revised: year-month-day; accepted: year-month-day

## 2 Related Works

## 3 Method

Inspired by online hard example mining(OHEM) algorithm, we propose a cascade coarse to fine network architecture CCFNet. Section 3.1 illustrates how to turn a ImageNet pre-trained model to a cascade coarse to fine semantic segmentation network CCFNet. Taking the characteristics of semantic segmentation task into consideration, section 4 introduces a hard instance mining method to learn the attention about the difficulty of each instance. Section 4 detailes the training and testing process of CCFNet.

### 3.1 Cascade Coarse to Fine Network Architecture

We choose ResNet-50 pre-trained on ImageNet as our baseline model. ResNet originally is designed for image classification which win ILSRVC 2015 competation and surpass the human performance on ImageNet dataset. ResNet-50 is one of the version provided in experiments, which is faster than VGG-16 and more accurate than VGG-19. Compared with ResNet-101, ResNet-50 is cheaper in computation resources and memory consumption, but can achieve a comparable accuracy. Figure **??**a visualizes the network architecture of ResNet-50, which is composed by five stages with different configurations of layers and a classification stage. We treat the ResNet-50 as a common feature extraction part of CCFNet by discarding the classification stage.

The architecture of CCFNet is shown as Fig. **??**b, which is composed by three cascade network branches: a coarse segmentation branch as a baseline result, an attention branch to predict the difficulty of labeling each pixel instance, and a refine segmentation branch to refine the final segmentation results. These three branches share a common feature extraction network.

**The feature extraction network is a fundamental convolutional neural network. A modified version of ResNet-50 is adopted by this paper. For semantic segmentation task, the context is important to predict the correct label of each pixel instance. But it is difficult to determine the boundary of each pixel's context, since different objects may have different sizes. And the problem gets more complicated when considering the various perspective of each images. A simple yet effective method to solve this problem is to integrate multi-scale features for label predicting. The residual error model itself has the property of extracting and integrating multi-scale features, which can be seen from Fig. ??. From the unravelled view by Veit et al. ??, a two-unit ResNet is equivalent to an ensemble of four sub-networks with different receptive fields , as illustrated in Fig. ??. So the whole ResNet-50 can be expanded as a linearly growing ensemble of sub-networks, which can extract and integrate multi-scale features.**

**Besides, there are two improvements adopted by ResNet-50 to make it more suitable for segmentation task. First, we only keep the first three pooling layers to preserve the resolution. So the final resolution of prediction is 1/8 of the original input image. Secondly, we replace the convolutional layer in the last two stages with the dilated convolutions. It can help to enlarge the reception field of predicted feature maps. The modified ResNet-50 is used as feature extraction network in CCFNet.**

**The coarse segmentation branch is a baseline model for image semantic segmentation. This branch is shown as the red part in Fig. ?? b, we adopt a fully convolutional network(FCN) with two convolutional layers to predict the semantic classes for their regions. Since the resolution is 1/8 of the original input image, the feature maps are up-sampled by bilinear interpolation. Finally, a pixel-wise softmax loss is adopted to predict the probabilities of each pixel.**

**The attention branch is proposed to learn a soft-attention, which is a one-channel feature map with the same resolution as the input image. It is mainly used to indicate the segment difficulty of each pixel. The idea behind is simple yet effective. The segmentation datasets contain a large number of easy pixel instances and a small number of difficulty pixel instances. Paying more attention on these difficulty pixel instances can make the training process converges faster and efficiently. The attention branch shares the same feature extraction network as the coarse segmentation branch, and has a similar network structure. The major different is that the attention branch is a two-category semantic segmentation network which is only used to predict the segment difficulty. Just the same as the coarse segmentation branch, softmax layer is adopted again to generate a soft-attention weighting coefficient.**

**The refine segmentation branch refines the segmentation results as the final network output. This branch is more complicated compared with the first two branches. The coarse segmentation branch is hard to segment all the pixels correctly. So the pixel instances can be divided into two groups by the coarse segmentation branch. The pixels which can be segment correctly by the coarse segmentation branch is denote as easy pixel instances, while the others are difficult ones. A fine segmentation network is introduced to reclassification the difficult pixel instances. Inspired by the PSPNet ??, pyramid pooling is adopted by the fine segmentation**

network to extract multi-scale features. And the refine segmentation branch is a weighted summation of the coarse segmentation branch and the fine segmentation network, with the weighting coefficient predicted from the attention branch. So an end-to-end learning branch is proposed to learn the final segmentation result directly.

The three branches are cascaded one by one, and constitute an end-to-end learning network for semantic segmentation.

### 3.2 Learn the Attention about Hard Instance

In the training process, the rough semantic segmentation network can give a semantic prediction result, combined with the real semantic label of the training image, we can get the correct result of the classification of the two categories of labels. The label can be used as a supervisory label for the difficult sample prediction network so that the difficult sample prediction network can predict the correctness and difficulty of each pixel classification of the training image.

### 3.3 Training and Testing

For example: The parallelization of cutoff pair interactions is mature on CPUs, and typically employs a voxel-based method.

## 4 Experimental results

### 4.1 Network Configuration

### 4.2 Data Sets

### 4.3 Comparison Results

### 4.4 Results Visualization

## 5 Conclusions

For example: The parallelization of cutoff pair interactions is mature on CPUs, and typically employs a voxel-based method.

References to the literature are cited by *number in square brackets* at appropriate locations (*before* a period, comma, etc.) in the text.

Examples:

- Negotiation research spans many disciplines [3].
- This result was later contradicted by Becker and Seligman [5, 6], who .......
- This effect has been widely studied [1-3, 7].
- ......achieved until rather recently [11, 21, 22], with......
- ......stage of cap formation (see Fig. 5 in Ref. [14]).

### Acknowledgements

This work was supported in part by the National Science Foundation of China (NSFC) under Grant Nos. 91420106, 90820305, and 60775040, and by the National High-Tech R&D Program of China under Grant No. 2012AA041402.

### References

The font is Times New Roman \zihao{5-}. This part is placed at the end of the manuscript. References should be numbered sequentially as they appear throughout the text. Only one publication should be given for each number. The list of references should only include papers that have been published or accepted by a named publication or recognized preprint server. Authors should ensure the accuracy and completeness of all references before submission. Please ensure references are given in the correct style as shown below in order to avoid delays in typesetting your article

### References

[1] M. Abrahams and M. Kattenfeld, The role of turbidity as a constraint on predator-prey interactions in aquatic environments, *Behav. Ecol. Sociobiol.,* vol. 40, no. 3, pp.169-174, Mar. 1997. (Journal style)

[2] R. C. Calfee and R. R. Valencia, *APA Guide to Preparing Manuscripts for Journal Publication*. Washington, DC: American Psychological Association, 1991. (Authored book style)

[3] J. M. ONeill and J. Egan, Mens and womens gender role journeys: metaphor for healing, transition, and transformation, in *Gender Issues Across the Life Cycle*, B. R. Wainrib, Ed. New York: Springer, 1992, pp. 107-123. (Book chapter style)

[4] J. C. Phelan, B. G. Link, A. Stueve, and B. A. Pescosolido, Have public conceptions of mental health changed in the past half century? Does it matter? presented at the 124th Annu. Meeting American Public Health Association, New York, USA, 1969. (Paper presented at conferences style)

[5] A. Mahdavi and B. Spasojevic, Incorporating simulation into building systems control logic, In *Proc. 10th Int. Building Performance Simulation Association Conf.*, Beijing, China, 2007, pp. 1175-1181. (Paper in proceedings style)

[6] X. Yang, Study of building material emissions and indoor air quality, Ph.D. dissertation, Dept. Arch., MIT, MA, USA, 1999. (Dissertation style)

[7] J. Dong, S. Martin, and P. Waldo, Method and system for dynamically presenting cluster analysis results, US Patent US6380937B1, March 30, 2002. (Patent style)

[8] Y. Jiang, Liquid desiccant air-conditioning system and its applications, (in Chinese), *Heating Ventilating & Air Conditioning*, vol. 34, no. 11, pp. 88-97, 2004. (Non-English publication style)

[9] M. K. Slifka and J. L. Whitton, Clinical implications of dysregulated cytokine production, *J. Mol. Med.*, doi:10.1007/s001090000086. (Article by DOI style)

[10] J. Doe. The dictionary of substances and their effects, http://www.rsc.org/ dose/title, 1999, Jan. 15. (Online document style)

First A. Author Photo. Biographies should be limited to one paragraph consisting of the following: sequentially ordered list of degrees, including years achieved; sequentially ordered places of employ concluding with current employment; associa-tion with any official journals or conferences; major profes-sional and/or academic achievements, i.e., best paper awards, research grants, etc.; any publication information (number of papers and titles of books published); current research interests; association with any professional associations.

Second B. Author Photo. Biographies should be limited to one paragraph consisting of the following: sequentially ordered list of degrees, including years achieved; sequentially ordered places of employ concluding with current employment; associa-tion with any official journals or conferences; major profes-sional and/or academic achievements, i.e., best paper awards, research grants, etc.; any publication information (number of papers and titles of books published); current research interests; association with any professional associations.

Third C. Author Photo. Biographies should be limited to one paragraph consisting of the following: sequentially ordered list of degrees, including years achieved; sequentially ordered places of employ concluding with current employment; associa-tion with any official journals or conferences; major profes-sional and/or academic achievements, i.e., best paper awards, research grants, etc.; any publication information (number of papers and titles of books published); current research interests; association with any professional associations.

**FIGURES. Below each figure, there should be a figure legend. The font is Times New Roman \zihao{5-}. A figure legend normally begins with a brief title describing its whole contents, and continues with a short description of each panel. The format of figure legends is "Fig. 1 Figure legend". The preferred figure formats is TIFF or JPEG with a preferred resolution of 600dpi relative to the final figure size. When figures are divided into parts, each part should be labeled with a lower-case (a), (b), and so on, in the same font size as used elsewhere in the figure. All lettering in figures should be in lower-case, except for the first letter of each label which should be capitalized. Use a single space to separate a number and its units. Ensure that the labels are sufficiently large and clear to be readable when the figure is reproduced in the print version of the journal. Figures are referred to in the manuscript as "Fig. 1" and "Fig. 1a", or "Figure 1" at the beginning of a sen-tence. After a manuscript is accepted, we may ask the authors to provide high resolution figures. Failure to supply the file can significantly delay the publication of your work. If possible, please provide a set of high resolution figures along with your initial manuscript. The width of figures should better be 8, 12, or 17 cm, and the height should be less than 19 cm.**

**TABLES. The font is Times New Roman \zihao{5-}. Above each table, there should be a brief title describing its contents. Title format should follow "Table 1 Table title". All details, such as nonstandard abbreviations, and a description of standards of error analysis, should be included in footnotes. Tables are referred to in the manuscript as "Table 1". The concise type tables are recommended.**

**EQUATIONS. The font is Times New Roman \zihao{5}. Equations and mathematical expressions are identified by parenthetical numbers, such as (1), and are referred to in the manuscript as "Eq. (1)", or "Equation (1)" at the beginning of a sentence.**