SOBEL HEURISTIC KERNEL FOR AERIAL SEMANTIC SEGMENTATION

Tao Hu[1], *Yao Wang*[1], *Yisong Chen*[1,2,3], *Peng Lu*[4], *Heng Wang*[1,3], *Guoping Wang*[1,2,3]

[1]Graphics and Interaction Laboratory, Peking University
[2]Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University
[3]Key Lab of Machine Perception and Intelligent, MOE;Department of Computer Sciences,Peking University
[4]School of Computer Science, Beijing University of Posts and Telecommunications
{taohu,yaowang95,chenyisong,hw,wgp}@pku.edu.cn, lupeng@bupt.edu.cn

ABSTRACT

Misclassification in semantic segmentation mostly occurs in the pixels around the semantic contour. In this work, we address the task of aerial image segmentation by borrowing the kernel prior from classical edge detecting operator. We propose a module called Sobel Heuristic Kernel(SHK). Our work makes several main contributions and experimentally shows good performance. To the best of our knowledge, we are the first to combine traditional edge detection method and deep learning method in semantic segmentation. Our SHK module reaches state of the art in the Inria Aerial Image Labeling dataset.

Index Terms— Semantic Segmentation, Edge Detection

1. INTRODUCTION

Semantic segmentation is a basic task in computer vision. As the Deep Learning technology prevails, most of the methods improve segmentation result based on the following aspects: (1). *larger receptive field*. Atrous Convolution[1, 2] is proposed to greatly enlarge the receptive field while it brings noteworthy memory overload by the large feature map. (2). *multi-scale context fusion*. Parallelized model is the mainstream method while high memory cost makes it less efficient. Another direction is Pyramid Pooling Module(PSP)[3] which utilizes the deeper level feature map to extract multi-scale information and significantly reduces the memory cost. Those aspects try to improve overall segmentation effect while the typical misclassification cases are often ignored.

Where misclassification often occurs? Misclassification mainly occurs in the semantic contour pixels(the pixels which construct the semantic concept of an image) as indicated in the uncertainty map in Fig 1. Its difficulty lies in the fact that as the actual receptive field in the neural network is totally different with the theoretical receptive field[4]. And the pixels in the internal area of object can be suitably perceived by surrounding pixels thus they can be rightly classified. While the pixels around the semantic contour cannot be rightly perceived by their surrounding pixels, their surrounding pixel labels can be entirely different. We attempt to solve this problem by borrowing the kernel prior from edge detectors.

Deep Learning Kernel Prior: Prior information is often applied into the neural network to realize some outside constraints. For instance, weight penalties of various L1 and L2 regularization and soft weight sharing[5] are frequently employed to prevent overfitting. On the other hand, low-rank kernel prior is employed to speed up the convolution[6]. In this paper, a sobel-filter like prior in the convolution kernel is utilized to implant the edge detection module in the neural network.

To solve this problem of misclassification surrounding semantic contours, we propose a sobel heuristic kernel, which originates from sobel edge detection method, to embed the edge-detection module into the neural network to establish an end-to-end architecture.

Our contributions are mainly two parts: (1). We put forward a module called Sobel Heuristic Kernel(SHK) to solve the misclassification around the semantic contour. (2). Our SHK module gives a pioneering and promising result in integrating traditional methods into deep neural networks and reaches state of the art result in the Inria Aerial Image Labeling dataset.

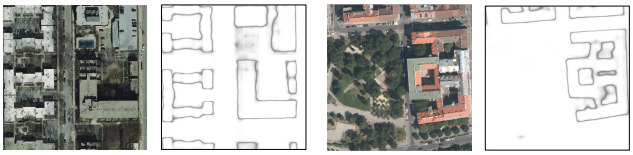

Fig. 1. Uncertainty Map. (b),(d) are the uncertainty maps of (a),(c) accordingly. White color means low uncertainty, black color means high uncertainty. The uncertainty map is obtained by max operation on the final prediction softmax probability map.

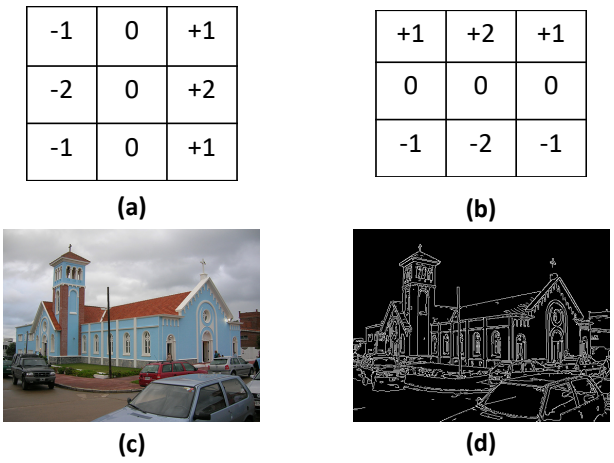

Fig. 2. An overview of Sobel Filter. (a). sobel vertical kernel Gx. (b). sobel horizontal kernel Gy. (c). an image demo. (d). the sobel detection result of image(c).

2. APPROACH

2.1. Sobel Detector Recap

The Sobel operator[7] performs a 2-D spatial gradient measurement on an image to emphasize regions of high spatial frequency that corresponds to edges. Typically it is used to find the approximate absolute gradient magnitude at each pixel in an input grayscale image[8].

The operator consists of a pair of 3×3 convolution kernels as shown in Fig 2. These kernels are designed to maximally respond to edges running vertically and horizontally relative to the pixel grid, one kernel for each of the two perpendicular orientations. The kernels can be applied separately to the input image to produce separate measurements of the gradient component in each orientation(call them Gx and Gy). These can then be combined together to find the absolute magnitude of the gradient at each point and the orientation of that gradient. The gradient magnitude is given by:

$$|G| = \sqrt{Gx^2 + Gy^2} \tag{1}$$

Typically, an approximate magnitude is computed using:

$$|G| = |Gx| + |Gy| \tag{2}$$

which is much faster to compute. The angle of the edge can be calculated by the spatial gradient:

$$\theta = arctan(\frac{Gy}{Gx}) \tag{3}$$

2.2. Sobel Heuristic Kernel

We restrict the convolution kernel in neural network according to Sobel Detector. As shown in Fig 4, we incorporate Sobel-shaped Mask which makes the neural network only learn the vertically or horizontally bordered filter weights.

In the following, analysis of the SHK module from the perspective of back propagation will be given. We can denote the current feature map as F_i, feature map after convolution and nonlinearity as F_{i+1}, the final loss as L, kernel variable as X(just makes it as 3×3 shape for convenience), an equal shape mask as M, f as nonlinear function, the masked kernel variable as $\bar{X} = M \circ X$,where $\circ$ means Hadamard Product. Therefore, we have:

$$f(F_i\bar{X}) = F_{i+1} \tag{4}$$

where

$$\bar{X} = M \circ X \tag{5}$$

According to Back Propagation, we could easily obtain $\frac{\partial L}{\partial F_{i+1}}$.

We can get the gradient of kernel by chain rule:

$$\frac{\partial L}{\partial X} = \frac{F_i M}{f'(F_i\bar{(X)})}\frac{\partial L}{\partial F_{i+1}} \tag{6}$$

From the Mask indicated in Fig 4, it is obvious that where the mask value equals zero will get zero gradient. SHK could realize an end-to-end learnable kernel, which can be plugged anywhere in the neural network.

2.3. Overall Framework

In image segmentation, there are three main neural network paradigms. (1). U-shape[9], which is composed of an encoder network and a corresponding decoder network. (2). Cone-shape[1]. It is formed by a classification model with final fully connected layers removed. Upsampling to the original scale of the image size is appended in the tail. (3). Pyramid shape[10], it accepts arbitrary size as input and outputs proportionally sized feature maps at multiple levels in a fully convolutional fashion. The construction of the pyramid involves a bottom-up pathway, a top-down pathway, and lateral connections.

Our overall segmentation model is shown in Fig 3. For simplicity, we set the input image size as 512×512. We use ResNet101[11] as our backbone of the feature network. Multi-scale feature maps are extracted from different stages in the feature network. 1×1 convolution is applied to generate multi-scale semantic score map for each class. Score map of low resolution will be upsampled with a deconvolution layer, then added up with higher ones to generate new score maps.

We also propose Sobel Heuristic Block(SHB) as indicated in Fig 3. We put the SHB in the deepest part of the FCN architecture because the deeper part shows more semantic feature for SHB to extract and fuse the edge information. In our SHB module, there is no nonlinearity after the SHK and a skip layer appended from the origin. The kernel size of SHB

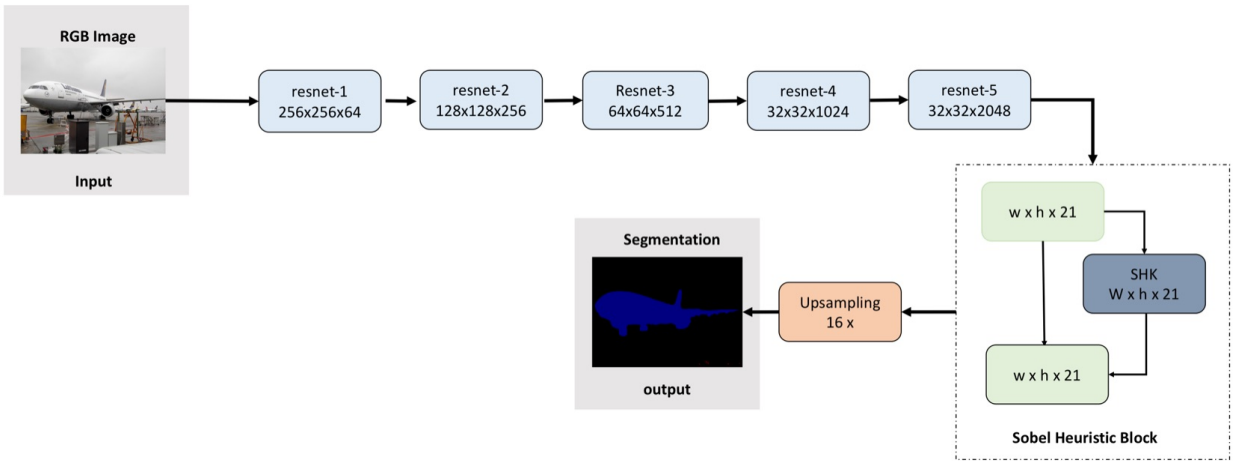

Fig. 3. Our network structure. Sobel Heuristic Block(SHB) is appended in the deepest part of our architecture. A SHB module is comprised of two convolutions and a SHK module. The detail of SHK module is illustrated in Fig 4.

is 3×3. Before upsampling, SHB module is employed to refine the context result. The final semantic score map will be generated after the last upsampling, which is used for the final prediction. Each SHB is comprised of two convolutions and SHK module. The detailed structure of SHK is demonstrated in Fig 4.

In our SHB, the feature map size is $w \times h \times c$, where c is the class number. If we adopt the channelwise SHK indicated in the later experiment 3.1, the number of kernel parameter is $3 \times 3 \times 2 \times c$ per channel. Thus our SHB module has negligible parameter burden.

3. EXPERIMENT

We evaluate our approach on the Inria Aerial Image labeling dataset[12]. The dataset owns images with coverage of 810 km^2 (405 km^2 for training and 405 km^2 for testing). Our backbone network ResNet 101 is pretrained on ImageNet[13]. Its ground truth contains two semantic classes: building and not building. The images cover dissimilar urban settlements, ranging from densely populated areas (e.g., San Franciscos financial district) to alpine towns (e.g,. Lienz in Austrian Tyrol). Instead of splitting adjacent portions of the same images into the training and test subsets, different cities are included in each of the subsets. The original dataset image size is 5000×5000, we crop it into 473×473 randomly.

During the training time, we use standard Adam[14] with initial learning rate 2.5e-4, momentum 0.99 and weight decay 5e-4. Random resize and flip are used as data augmentations. The performance is measured by standard mean Intersection-over-Union(IoU). All the experiments are running with tensorflow[15]. All our results were obtained by computing median over 5 runs.

In the ablation study, we mainly discuss the choice of sobel order, SHK connection type and boundary mIoU statistics.

3.1. Ablation Study

Sobel Order: As indicated in Fig 4, horizontal SHK and vertical SHK are combined by an operator. We experimentally evaluate three kinds of SHK operator: (1). half-order, following strict definition of Sobel Detector. (2). one-order, following the approximate format of Sobel Detector. (3). second-order, meaning high-order approximation of Sobel Detector. The result is shown in Table 2.

If we strictly follow the definition of Sobel detector[7], the setting of SHK should be half-order, however, we found that the half-order causes non-convergence, the reason of which we speculate is that the square root operator causes optimization problem about gradient. We also try the one-order, second-order choice for ablation study. From the result, we can know that two-order makes the activations inconsistent while one-order avoids that problem. From the above observation, the one-order SHK will be chosen as the default setting in our later experiment.

Table 2. SHK Order

Method	mean IoU	mean acc.	pixel acc.
half-order	non-convergence		
one-order	68.3	79.5	93.2
two-order	67.4	79.0	92.5

SHK Connection Type: We explore two different types of SHK connection, one is channelwise SHK, which is same as the Depthwise Convolution[16]. The other one is channel-cross SHK, namely the standard convolution with SHK mask. The related result is displayed in Table 3.

Table 1. Test result on Inria Aerial Image Labeling Dataset. The details of some methods are not revealed in the leaderboard.

method	Bellingham		Bloomington		Innsbruck		San Francisco		East Tyrol		Overall	
	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
Inria1[12]	52.91	95.14	46.08	94.95	58.12	95.16	57.84	86.05	59.03	96.40	55.82	93.54
DukeAMLL	66.90	96.69	58.48	96.15	69.92	96.37	75.54	91.87	72.34	97.42	70.91	95.70
NUS	**70.74**	97.00	66.06	96.74	73.17	96.75	73.57	91.19	76.06	97.81	72.45	95.90
ENPC Singh2	64.28	96.00	65.84	96.52	**77.11**	**97.31**	75.86	92.01	78.68	98.12	73.30	95.99
Our Method	70.73	**97.09**	**69.98**	**97.22**	76.74	97.29	**76.73**	**92.34**	**79.09**	**98.17**	**75.33**	**96.42**

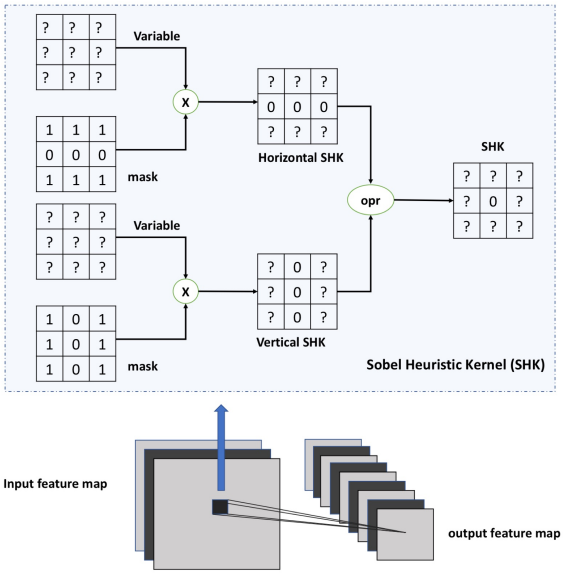

Fig. 4. Sobel Heuristic Kernel. SHK includes horizontal SHK and vertical SHK, they are combined by an operator, the choice of the operator will be discussed in the later Experiment section 3.1.

Channelwise has 1% higher mIoU than channelcross connection type.We can interpret Channelwise loss as the effects of Sobel Detector on feature maps.

Table 3. SHK connection type

Method	mean IoU	mean acc.	pixel acc.
Channelwise	68.3	79.5	93.2
Channelcross	67.3	78.5	92.4

Boundary Statistics: Detail experiment about improvement about the SHK is conducted in boundary and non-boundary regions following [10]: a) boundary region, whose pixels locate close to objects boundary (distance ≤ 7). b). internal region as other pixels. The segmentation results of both regions are in Table 4. We find that our SHK module mainly improves the result in Boundary region(1% higher than 0.2% higher), which strongly supports our argument.

Table 4. SHK Boundary Statistics

Method	Boundary(mIoU)	Internal(mIoU)	Overall(mIoU)
Baseline	60.85	66.14	67.3
SHK	61.87	66.32	68.4

3.2. Final Result

Our final algorithm is applied in the Inria Aerial Image Labeling dataset, which performs the best in the city of Bloomington, San Francisco, East Tyrol in both mIoU and accuracy. Our overall mIoU is 75.33%, which reaches state of the art compared with other methods. Some visualization results are shown in Fig 5.

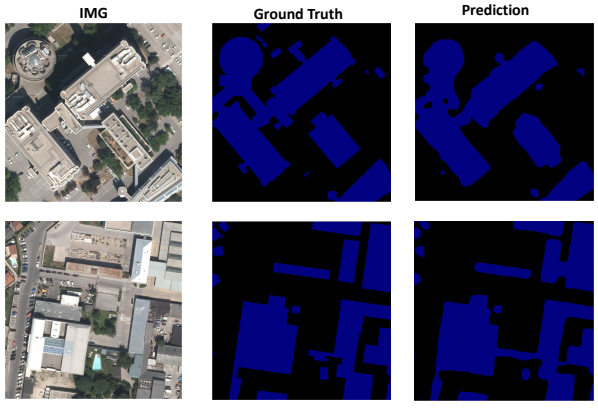

Fig. 5. Inria Aerial Image labeling Dataset Validation set Visualization Result

4. CONCLUSION

In this paper, we combine Sobel Detector[7] with Fully Convolutional Network to improve the segmentation result around the semantic contour. Our algorithm reaches state of the art in the Inria Aerial Image Labeling Dataset[12]. It demonstrates that traditional method doesn't die out in the prevailing trend of Deep Learning, they just live in other patterns that coordinate with Deep Learning framework.

In our future work, we will try to combine other traditional method such as super-pixel method to further improve the semantic segmentation.

Acknowledgements

We acknowledge the anonymous reviewers for their comments and suggestions. This work is supported by The National Key Technology Research and Development Program of China under Grants 2017YFB1002705 and 2017YFB1002601, and by National Natural Science Foundation of China (NSFC) under Grants 61472010, 61632003, 61631001 and 61661146002.

5. REFERENCES

[1] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[3] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[5] Steven J Nowlan and Geoffrey E Hinton, "Simplifying neural networks by soft weight-sharing," *Neural computation*, vol. 4, no. 4, pp. 473–493, 1992.

[6] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Advances in Neural Information Processing Systems*, 2014, pp. 1269–1277.

[7] Irwin Sobel, "History and definition of the sobel operator," *Retrieved from the World Wide Web*, 2014.

[8] A. Walker R. Fisher, S. Perkins and E. Wolfart., "Sobel edge detector," `http://homepages.inf.ed.ac.uk/rbf/HIPR2/sobel.htm`, 2003.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 234–241.

[10] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun, "Large kernel matters improve semantic segmentation by global convolutional network," 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez, "Can semantic labeling

methods generalize to any city? the inria aerial image labeling benchmark," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[15] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: A system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[16] François Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint arXiv:1610.02357*, 2016.