Technical Report of Transportation Project

Ruiqi Li
Dian Group
School of EIC
Huazhong University of
Science and Technology
rlee@hust.edu.cn

***Abstract*—Transportation Project of Dian Group launched in the spring of 2019, the goal of which is to meet engineering requirement of the transportation system of Pingxiang city, Jiangxi. Several different problems is to solve in the project. This technical report conclude progress that we made in some of the engineering problems as well as academic literature that we refers to.**

I. Introduction

The current public transportation surveillance platform of the city of Pingxiang underwent its construction and upgrade mostly 6 years ago, when deep learning remains developing and the majority of technologies adopted were machine learning and traditional image process algorithms. Deep learning has transcend original method drastically, reflecting necessity of substitution of the system. Practically, we have deployed our deep learning algorithms by upgrading and rebuilding the obsolete traffic monitoring system of a city and achieved improvement on the accuracy by 1000%, hence spared police department from heavy manual work.

The engineering problems we face mainly include crossing lane line detection and intruding forbidden area. According to traffic regulations, crossing full lane line is restricted on road since unexpected sheer off of vehicles may easily lead to traffic accident. The regulations also state clearly that engineering cars and other vehicle with potential safety threats should be restricted in terms of travel time and region. Automatic detection these illegal act and managing corresponding information will contributes to the surveillance of the city and therefore cut down the scale of them. Chapter II and chapter III in the rest of this report conclude works that have down in solving the two problem respectively.

In solving cross lane line problem, we learned that lane line detection is a vital task for the settlement of the problem. In the meantime, lane line detection task, which is considered as a instance segmentation problem, remaining active in computer vision research field. Thus we compose a literature review of instance and semantic segmentation problem in chapter IV.

(a) original data (b) mask

Fig. 2: Typical Cross Line Image and Generated Mask

II. Cross Line Detection

A picture from original data is shown in Fig.2 (a). To detect cross lane line vehicle picture from thousands of pictures captured by surveillance camera, we design a detection pipeline

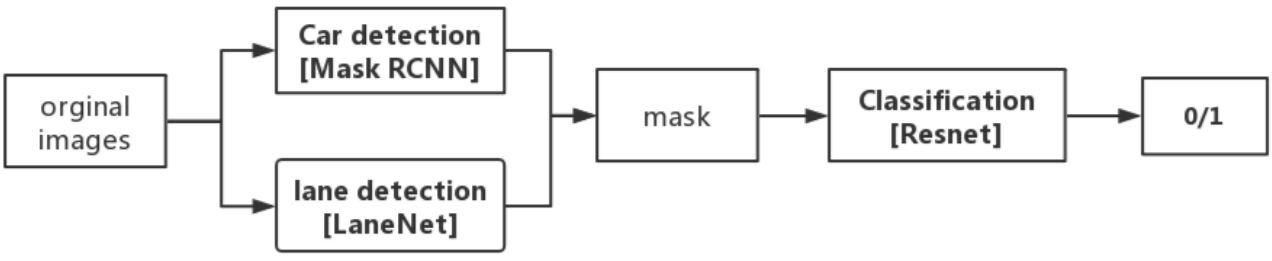

Fig. 1: Cross Line Detection Pipeline

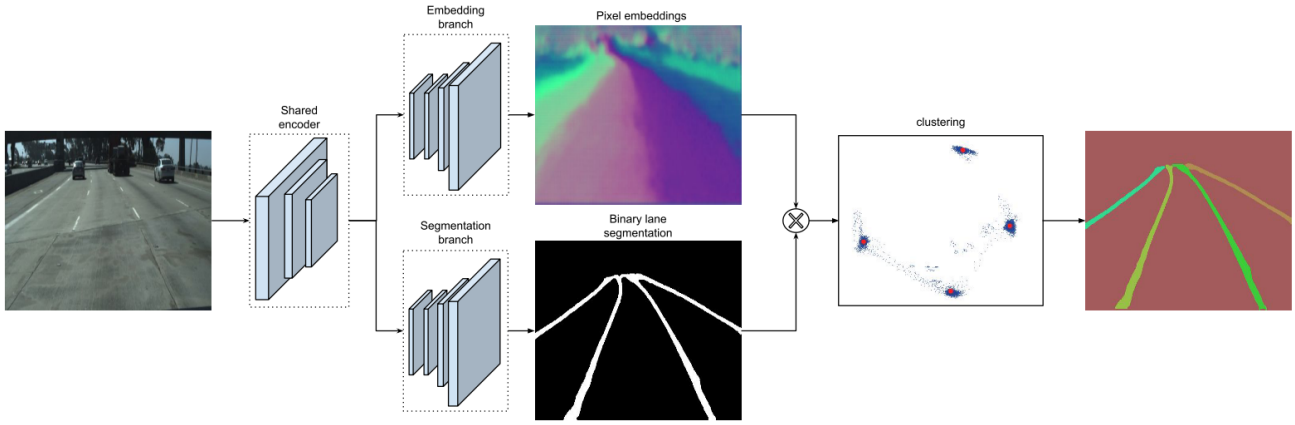

Fig. 3: Cross Line Detection Pipeline

, shown in Fig.1, consist of three major steps. To determine whether a picture contains cross line vehicle, we intuitively consider that the vehicles and lane in the image should be located firstly. Hence Mask R-CNN [4] and Lane Net [2] are deployed in our detection pipeline, the example result of which are shown in Fig.1 (b).

The generated mask contains two channel, each channel labeled pixels with binary value, where numerical 1 stand for pixel that compose vehicle or lane line. After mask generating process, the two channel mask is send to an classifier which formed by Resnet [3] to determine whether a picture contains cross line vehicle or not.

A. Lane Line Detection

For lane line detection branch, we adopted LaneNet [2] to determine vehicle mask. LaneNet is a branched, multi-task network, consisting of a lane segmentation branch and a lane embedding branch that can be trained end-to-end. The lane segmentation branch has two output classes, background or lane, while the lane embedding branch, which is trained using a clustering loss function, assigns a lane id to each pixel from the lane segmentation branch while ignoring the background pixels, thus disentangles the segmented lane pixels into different lane instances.

Having estimated the lane instances, i.e. which pixels belong to which lane, as a final step we would like to convert each one of them into a parametric description. LanNet apply a perspective transformation onto the image before fitting a curve, but in contrast to existing methods that rely on a fixed transformation matrix for doing the perspective transformation, we train a neural network to output the transformation coefficients. In particular, the neural network takes as input the image and is optimized with a loss function that is tailored to the lane fitting problem. An inherent advantage of the proposed method is that the lane fitting is robust against road plane changes and is specifically optimized for better fitting the lanes. An overview of our full pipeline can be seen in Fig.3.

B. Vehicle Detection

For vehicle detection branch, we adopt Mask R-CNN [4]. In respect of engineering, we implement Mask R-CNN by MMDetection [5] which is a toolbox developed for object detection research. We adopted pre-trained model using dataset. To determine the mask of vehicles, we filter out the output with label car, truck, bus of Mask R-CNN. An typical output is shown in Fig.2 (b) with green color.

C. Binary Classification

To draw the conclusion that whether a image contains violator vehicle or not, we take the combined mask of vehicle and lane line as the input of Resnet-50 and output a binary prediction.

III. Literature Review of Segmentation Problem

The section is organized that subsection A describe a typical pipeline of segmentation model; subsection B and subsection C respectively conclude usual encoder schema and decoder schema. Subsection D summarize common strategy for improving performance on multi-scale scene.

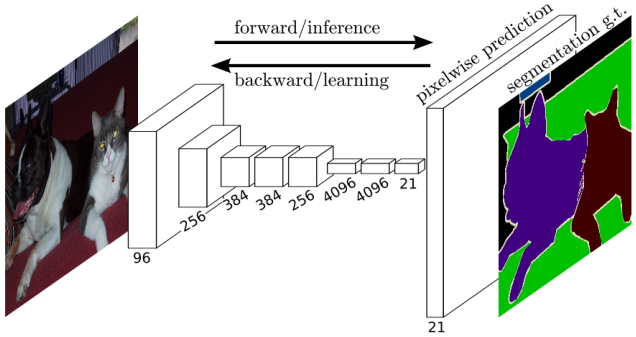

Fig. 4: Encoder-Decoder Pipeline of FCN

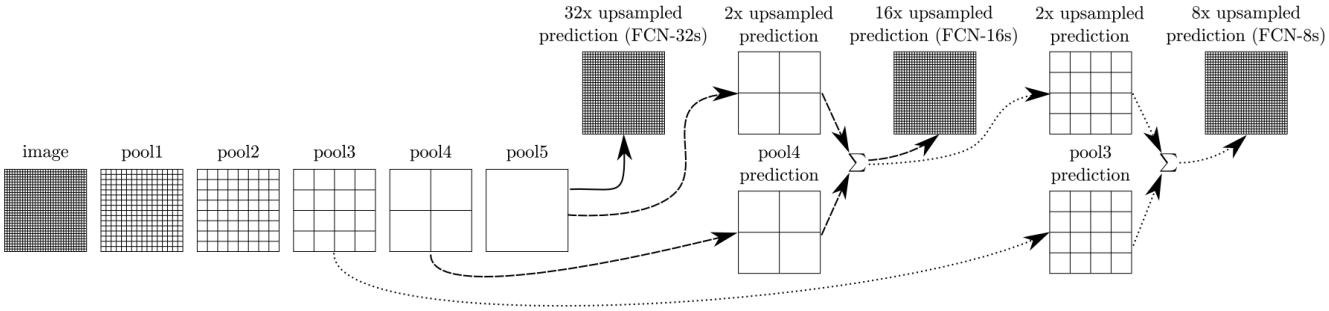

Fig. 5: Feature Fusion Strategy of FCN: DAG net

A. FCN: A Typical Pipeline

Fully Convolution Network [6] proposed in 2014 is a typical case of adapting deep neural network in completing semantic segmentation task. The pipeline of FCN, shown as Fig.4, is usually illustrated as a encoder-decoder structure. The structure is widely employed by a majority of semantic segmentation models.

An encoder refers to a forward procedure, consist of a stack of convolution-pooling layers. The FCN obtain its finest result by adapting deconvolution along with the skip layer fusion strategy as the decoder.

Two major strategy of decoder is deployed by Fully Convolutional Network to gain finer precision. First off, Deconvolution outperform bilinear interpolation since this type of computation is available for back propagation algorithm and can learn its parameter to fit the data while upsampled the feature map by a certain factor.

To combine coarse, high layer information with fine, low layer information, FCN designed a information fusion strategy denoted as DAG net shown in Fig.5. Another benefit of applying such framework is that the integrated model is capable of learning to handle multi scale object.

B. Decoder Schema

To our knowledge, the decoder in a semantic segmentation model is vital and could influence greatly on the performance of segmentation in respect of accuracy and computational cost. A good decoder should reserve more high level semantic information while enlarge the resolution of the feature map.

1) Deconvolution: Deconvolution is an computation method which is widely used in the design of decoder. To derive the calculation method and an interpretation for it, we firstly denote convolution computation as matrix operation form:

$$\mathbf{Y} = \mathbf{C}\mathbf{X} \tag{1}$$

while X, Y, C respectively denote the matrix of input feature map, output feature map and convolution computation.

Considering the inverse operation of convolution. Since the inverse matrix of C do not exist, heuristically design a operation shown as follow:

$$\mathbf{X}' = \mathbf{C}^{\mathrm{T}}\mathbf{Y} \tag{2}$$

Obviously, a computation in this form is learnable while satisfying the needs of upsample.

2) Novel Unsample Method of SegNet: SegNet [7] contributes to the decoding technique in a novel method of upsample. As Fig.6 illustrated, during max-pooling of encoding process the max-pooling indices are recorded, and while decoding the coarse feature map the pixel value returns to the pooling indices position, leaving others pixels equals to zeros.

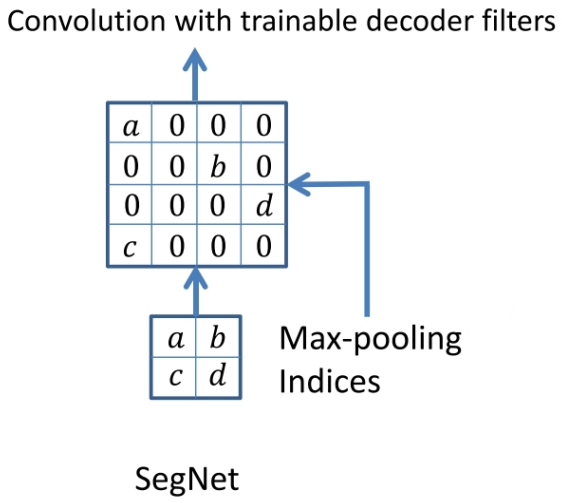

Fig. 6: Pooling Strategy of SegNet

Although this type of upsample fashion retain spatial information to some extent, the pooling computation leaving the output feature map too many zeros that it is hard for the following convolution layer to learn.

C. Encoder Schema

1) ENet and Dilated Convolution Bottleneck: It is very important for the network to have a wide receptive field, so it can perform classification by taking a wider context into account. Simply performing pooling to the feature maps might be a rough way to downsample and therefore enlarge the receptive filed. Therefore ENet [9] decided to use dilated convolutions initially presented in [8] to improve encoding performance model. Moreover, ENet improve dilated convolutional layers by designing bottleneck blocks as shown in Fig.7. A bottleneck

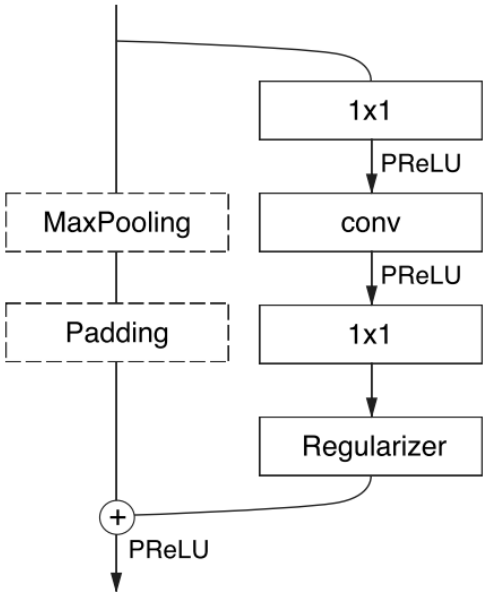

Fig. 7: ENet Bottleneck Block. *conv* is either a regular, dilated, or full convolution.

block, which refers to the basic block of the ENet, consisting of 1×1 *conv* and regular *conv* and sum up with the output of pooling and padding computation. These gave a significant accuracy boost in corresponding experiment.

2) Hybrid Dilated Convolution: One theoretical issue exists in the dilated convolution framework is known as "gridding-effect". A illustration of the effect is that one actual pixel in the output of a stack of convolution layer only relative with some of the input pixels, while a gap of pixel is skipped during computation.

Wang *et al.* proposed a Hybrid Dilated Convolution method in order to improve the performance of downsample. The solution to reducing gridding-effect is to chose special dilation rate. By doing this, the top layer can access information from a broader range of pixels. Similar to ENet, Wang *et al.* also designed an basic block and connecting these blocks together. This process is repeated through all layers, thus making the receptive field unchanged at the top layer.

D. Cope with Multi-Scale Object

A major challenge of object detection problem is that the scale of objects might varies greatly from each other, thus various of method in detecting multi-scale object have been developed. Likewise, different scene patches in an input picture challenge the precision of semantic segmentation result.

Pooling to downsample has one big advantage. Filters operating on downsampled images have a bigger receptive field, that allows them to gather more context. This is especially important when trying to differentiate between classes like, for example, rider and pedestrian in a road scene. It is not enough that the network learns how people look, the context in which they appear is equally important.

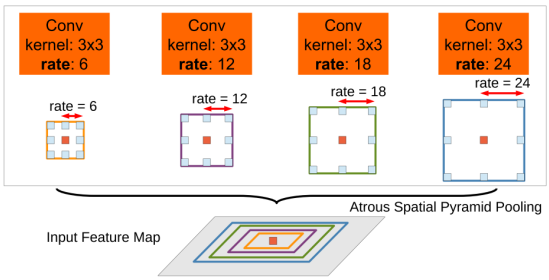

Fig. 8: Atrous Spatial Pyramid Pooling

1) DeepLab: In resolving the situation, Chen *et al.* proposed DeepLab [12] which deploy Atrous Pyramid Pooling Module to improve the performance of semantic segmentation module. Pyramid Pooling Module firstly appeared in the paper of He *et al.* [11] , which take a certain size of input feature map and concatenate the output of the pooling layers with different strides as the final output. This process is illustrated in Fig.8.

2) Pyramid Scene Parsing Network: Zhao *et al.* proposed PSPNet [10] in promoting the result of multi-scale feature fusion. As explained in [11], traditional Spatial Pyramid Pooling flatten the output of different pooling layers and concatenate them together. Nevertheless, PSPNet upsample the output of different pooling layers and concatenate the 2-dimensional output feature map together in channel. Thus the concatenated feature map is more explainable and is able to perform convolution computation.

REFERENCES

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár and Ross Girshick. Mask R-CNN, 2017; arXiv:1703.06870. II, II-B

[2] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans and Luc Van Gool. Towards End-to-End Lane Detection: an Instance Segmentation Approach, 2018; arXiv:1802.05591. II, II-A

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition, 2015; arXiv:1512.03385. II

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár and Ross Girshick. Mask R-CNN, 2017; arXiv:1703.06870. II, II-B

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy and Dahua Lin. MMDetection: Open MMLab Detection Toolbox and Benchmark, 2019; arXiv:1906.07155. II-B

[6] Jonathan Long, Evan Shelhamer and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation, 2014; arXiv:1411.4038. III-A

[7] Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, 2015; arXiv:1511.00561. III-B2

[8] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions, 2015; arXiv:1511.07122. III-C1

[9] Adam Paszke, Abhishek Chaurasia, Sangpil Kim and Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation, 2016; arXiv:1606.02147. III-C1

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang and Jiaya Jia. Pyramid Scene Parsing Network, 2016; arXiv:1612.01105. III-D2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, 2014; arXiv:1406.4729. DOI: 10.1007/978-3-319-10578-9_23. III-D1, III-D2

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, 2014; arXiv:1412.7062. III-D1