Semantic Segmentation for Urban-Scene Images Midterm Report

Team DL Sailor Moon
Heinz College, Carnegie Mellon University
{ haijingf, haiyunta, yudiz, yihanq } @andrew.cmu.edu

Abstract

Urban-scene Image segmentation is an important and trending topic in computer vision with wide use cases. For this project, our team aims to develop an advanced deep learning algorithm improving the performance of the baseline models like FCN-8s and DeepLabv3+ and solve problems with imbalanced dataset. In our current experiment, we found out that the performance of the baseline models varies upon objects with different scales in the images. Therefore, in future experiments, we are considering incorporating the positional patterns and context prior knowledge of those images in our model design to address this issue. Besides, we also explored some other backbones we could use in our model, which generated better performance than ResNet50 and ResNet100, such as WiderResNet38.

https://github.com/thyt115n/11785-project

1 Introduction

1.1 Background

Semantic image segmentation, the task of labeling each pixel of an image with a corresponding class of what is being represented, has always been a challenging and crucial task in the field of computer vision [1]. Urban-scene Image segmentation is a particular type that falls into this topic. It has been widely developed in recent years, which expedites the development of applications like autonomous driving vehicles. Take autonomous driving as an example: the images and videos captured by car-mounted cameras generally form large scale datasets that are applicable for deep neural network training. Therefore, the use of advanced deep learning techniques plays a significant role in improving segmentation performance for the overall scene background and the individual objects moving in front of the cars. As a result, these well-developed deep learning algorithms become solid cornerstones for recognizing roads, pedestrians, cars, and sidewalks at a pixel-level accuracy to support further autonomous driving algorithms.

1.2 Problem Statement

Urban-scene images is a specific type of image in semantic image segmentation that has intrinsic features in regards to positional patterns and context prior knowledge. For example, since the urban-scene images used in autonomous driving usually are usually captured by the camera positioned at the front of the car, data points are mostly road-driving pictures with spatial positioning bias. In horizontally segmented sections, roads are usually centered, with side-walk and trees at the left and right-hand sides of the picture. The spatial applies to the vertical position as well: sky is at the top section, while cars are usually captured at the lower part of the image[2]. On the other hand, noises like car paint with a figure would confuse a model to distinguish between the paint and a real person. In such a case, a contextual dependency called context prior knowledge would help in developing a more advanced model for urban-scene image segmentation [4].

The distinct nature of urban-scene images yields the possibility of incorporating intrinsic prior and feature extractions to general semantic segmentation models, not limited to the structural priors and context priors as discussed above. In this project, we would like to incorporate multiple helpful prior knowledge that applies to urban-scene images. We aim to deploy an integrated and advanced deep learning algorithm that targets urban-scene image semantic segmentation to improve the performance of the baseline models and better solve the problem incurred by an imbalanced dataset.

2 Literature Review

Context Prior for Scene Segmentation Changqian Yu et al. [4] emphasizes the problem that most semantic segmentation approaches rarely distinguish different types of contextual dependencies, and thus generate noises in the scene understanding. They proposed a learned Context Prior Network (CPNet) that can work with conventional deep CNN to capture the difference between intra-class and inter-class contexts to boost up model performances. The research is highly valuable for our problem setting when we are at the first step of reducing image limitations and incorporating urban-scene characteristics.

Height-Driven Attention Net(HANet) Sungha C. et al. [2] observed that, when dividing the image horizontally with three different parts: low, middle, and high, the overall uncertainty of prediction represented by entropy is greatly reduced. Based on such observation, they developed a novel height-driven attention networks (HANet) as a general add-on module, which was able to greatly improve the performance of various baseline models of semantic segmentation for urban-scene images.

Hierarchical Multi-Scale Attention To address the limitations of averaging and max-pooling in multi-scale inference, Tao et al.[5] proposes a hierarchical attention mechanism to combine predictions through relative weighting between adjacent scales. This mechanism has proven to be both memory and computationally efficient and leads to better model accuracy.

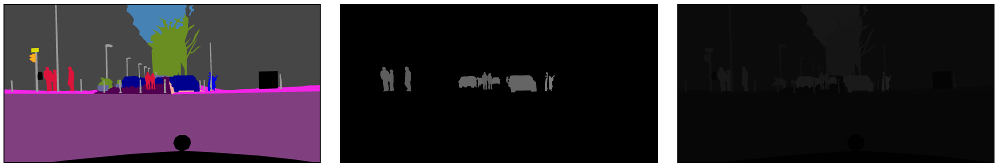
Multi-Scale Objects of Urban-Scene Images One major problem people are facing when working on urban-scene images is its variety of object scales. Many specific tricks are used to address this problem. For example, M. Yang et al. [18] proposes using a set of atrous densely connected convolutional layers to address large change of scales without decreasing resolution. Also, Li, X. et al. [19] proposes to conduct scale normalization with geometry property as prior information to cope with parsing failures on objects of heterogeneous scales.

Evaluation Metrics for Scene Segmentation Gabriela Csurka et al. [16] discusses various evaluation metrics for scene segmentation and their comparison based on empirical results. The paper discusses the strengths and limitations of the few existing measures like Pixel Accuracy, Jaccard Index (Intersection of Union), and Per-Class Accuracy. It also proposes a revised version of F1 Score call per-image F1 score (denoted BF) that provides a more comprehensive view of model performance by showing independence with the Jaccard Index, which can be considered when we evaluate our model.

3 Dataset Description

Cityscapes The dataset we will be primarily using is Cityscapes [6], a diverse large-scale dataset designed for urban scene semantic segmentation. It is derived from video sequences recorded in the streets of 50 cities. It contains 5K images with high-quality pixel-level annotations and 20K images with coarse annotations (Figure 1).

We only use the fine annotation set in our experiments. The fine annotation set with 5k data points is then split into a training set (2,975 images), a validation set (500 images), and a test set (1525 test images), and 19 semantic labels are defined. During our experiment, we used the toolkit **cityscapesScripts** for inspection, generating encode labels, and evaluation of the Cityscapes dataset. We also performed some data augmentations techniques, such as cropping into 512*1240, random horizontally flipping, random scaling, and Gaussian blur.



(a) Colored (b) Instance (c) Labeled

Figure 1: Cityscapes dataset demo

Mapillary Mapillary Vistas Dataset [7] has a larger dataset size but is more challenging due to multiple kinds of weather, times of day, and scene types, imaging devices, and differently experienced photographers. It includes a semantic image segmentation dataset, consisting of 25,000 high-resolution images and 100 instance-specifically annotated categories. Compared to **Cityscapes**, **Mapillary** is a larger dataset with more complicated environmental conditions (Table 1).

	Sequences	**Images**	**Multi Cities**	**Multi Weathers**	**Multi Times**	**Multi Scenes**
Cityscapes	50	5K	Yes	No	No	No
Mapillary	None	25K	Yes	Yes	Yes	Yes

Table 1: Comparison between Cityscapes and BDD100K

Currently, we have loaded **Cityscapes** and implemented baseline model evaluations on **Cityscapes**. Due to the limitation of CPU memory and GPU computing power in AWS EC2 Instance, we would use **Cityscapes** while developing the model modifications. If our model goes well with the **Cityscapes** dataset and we are able to require GPU cores with more computing power from AWS, we can apply our algorithms on **Mapillary**, which is a larger and more complicated dataset and could yield more generalized and unbiased results.

4 Evaluation Metrics

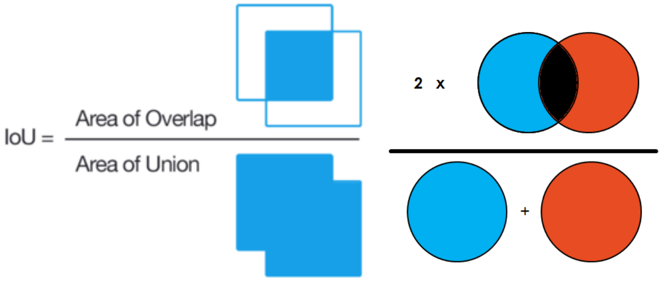
Pixel Accuracy Pixel accuracy is a straightforward metric that measures the proportion of correctly labeled pixels. There are two types of Pixel accuracy. **Overall Pixel accuracy** measures the proportion of correctly labeled pixels without considering the specific classes of label. **Overall Pixel accuracy** will highly bias the presence of very imbalanced classes. **Per-class Pixel accuracy** measures the proportion of correctly labeled pixels for each class and then averages over the classes and is thus less biased than **Overall Pixel accuracy**.

Intersection-Over-Union (Jaccard Index) The Intersection-Over-Union(**IoU**), as known as Jaccard Index, is calculated by the number of overlapping pixels between the predicted segmentation and the ground truth divided by the number of union pixels of predicted segmentation and the ground truth. Figure 2.a provides a visualized calculation of **IoU** scores. For multi-class segmentation in our project, we can calculate **per-class IoU** and also **mean IoU** (mIoU), which is taking the average of **per-class IoU**.

A **IoU** score is a range between 0 and 1, with 0 meaning totally wrong prediction and 1 meaning perfectly correct prediction. As **IoU** appreciated corrected labeled portion by accounting for overlap, it is a less biased measurement in general cases. One possible limitation is at **IoU** does not necessarily tell you how accurate the segmentation boundaries are[16].

Dice Coefficient(F1 Score) Dice Coefficient, which we called **F1** Score, is twice the area of overlap divided by the total number of pixels in both images (See Figure 2.b for visualized illustration). As for calculation, **F1** is a similar measurement as **IoU**. Looking closer, it is a harmonic mean of precision and recall by its definition and tends to provide more generalized information for unbalanced dataset.

Our project mainly uses per-class IoU and Mean Intersection-Over-Union (mIoU) as evaluation metrics. In the later work, when we are developing our modified models, we would also incorporate



(a) IoU Calculation (b) F1 Calculation

Figure 2: IoU Calculation vs F1 Calculation. Retrieved From Wikipedia

mean F-1 Score (mF1) to have more comprehensive views regarding model performance. Pixel accuracy in general is a highly limited metric that yields biased and uninformative impressions and is not considered in our project.

5 Baseline Models

5.1 *State-of-the-art* Models

After the breakthrough of implementing Deep Convolutional Neural Networks(CNN) in urban-scene segmentation, various advanced and modified methods have been developed to improve the scene parsing performance. Model architectures have been improved with techniques like Fully Convolutional Neural Networks(FCNs) [3], ACFNet [8], FastNet [9], etc. Other researches also highlight the limitation and specific characteristics in urban-scene datasets and develop algorithms like FoveaNet [10] to take care of the scale difference between images in urban-scene and general scene segmentation.

The **Cityscapes** dataset [6] we have selected is a widely-used dataset that has fruitful algorithms being developed and evaluated upon. As such, we introduce several *state-of-the-art* models that we have tried to run and demonstrate their prediction results. The *state-of-the-art* models that have been tested on Cityscapes Dataset includes Fully Convolutional Networks(FCNs) [3] , PSPNet[25], CGNet[20] and DeepLabv3+ [15] with various combination of backbones (ResNet-101, ResNet-50, ResNet-38, M3N21).

The table below shows their model performance using per-class IoU and mIoU accordingly.

State-of-the-art Model Performance in percentage										
Network	**Backbone**	**sky**	**building**	**road**	**fence**	**car**	**sign**	**person**	**cyclist**	**mIoU**
FCN-8s	ResNet-101	94.36	83.97	93.82	24.91	84.80	50.92	59.89	59.11	59.97
CGNet	M3N21	92.07	89.87	97.09	52.69	92.14	68.99	73.33	51.91	68.27
PSPNet	ResNet-101	77.57	86.81	95.27	33.31	88.99	53.34	63.25	63.87	69.28
DeepLabv3+	ResNet-101	92.82	89.02	96.74	41.00	91.02	64.48	66.52	66.98	75.21

Table 2: Comparison of mIoU and per-class IoU in percentage with *state-of-the-art* models on **Cityscapes** Dataset.

Fully Convolutional Network(FCN) The key insight of Jonathan Long et al. [3] is to build "fully convolutional" networks that take input of the arbitrary size and produce correspondingly-sized output with efficient inference and learning. They adapted classification networks into fully convolutional networks and transferred representations to the segmentation task. Then they define a novel architecture that combines semantic information from a deep layer with appearance information from a shallow layer to produce accurate and detailed segmentation.

CGNet: A Light-weight Context Guided Network for Semantic Segmentation In semantic segmentation, one major problem is that models with small memory footprint get relatively low segmentation accuracy. To address this issue, Wu, T. et al.[20] proposed a novel Context Guided Network (CGNet), which is a lightweight and efficient network specially tailored for semantic segmentation. They present Context Guided (CG) block, which is the basic unit of CGNet, to model the spatial dependency and the semantic contextual information in segmentation task, and therefore improve accuracy.

PSPNet: Pyramid Scene Parsing Network For scene parsing, a big challenge lies in unrestricted open vocabulary and diverse scenes. In this paper, Zhao H. et al. [23] exploit the capability of global context information by different-region based context aggregation through the pyramid pooling module together with the proposed Pyramid Scene Parsing network (PSPNet). This PSPNet provides a superior framework for pixel-level prediction on scene parsing jobs, where a single PSPNet yields a mIoU accuracy of 80.2% on Cityscapes datasets.

DeepLabv3+: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs Chen, L. et al. [15] has figured that implementing DeepLab systems could help with current problems of Deep Convolutional Neural Networks (DCNN). The main advantages of DeepLab system are speed, accuracy as well as simplicity. To capture the contextual information at multiple scales, they were able to present an updated DeepLabv3+[24] system in 2018 that features several improvements compared to its DeepLabv3 [22] version in 2017. DeepLabv3 applies several parallel atrous convolutions with different rates (called Atrous Spatial Pyramid Pooling, or ASPP), while PSPNet [23] performs pooling operations at different grid scales. Specifically, DeepLabv3+ extends DeepLabv3 by adding a simple yet effective decoder module to refine the segmentation results especially along object boundaries, demonstrating the effectiveness of the proposed model on PASCAL VOC 2012 and Cityscapes datasets, achieving the test set performance of 89% and 82.1% without any post-processing.

5.2 Baseline Models

By comparing and evaluating the benefits and limitations of current state-of-the-art models we mentioned above, we select two baseline models to make further improvements: **FCN-8s** [3] and **DeepLabv3+** [15]. We select these two among the *state-of-the-art* models that demonstrate the most simple and most complex algorithms trained for this dataset. The baseline models can be based on different backbones. Therefore, we plan to implement our modification with these two baseline models with two backbones: ResNet-50 and ResNet-101, in order to illustrate the influence of backbones to our model modifications.

Table 3 below shows the baseline models with different backbones in the evaluation metrics of mIoU in percentage. The baseline model performance results that we ran is similar to the published results, in the sense that FCN is a simple model that has relatively low model performance, whereas DeepLabv3+ outperforms among the existing state-of-the-art models [8][15].

Baseline Model Performance in percentage										
Network	**Backbone**	**sky**	**building**	**road**	**fence**	**car**	**sign**	**person**	**cyclist**	**mIoU**
FCN-8s	ResNet50	94.86	91.97	97.87	56.53	93.89	79.43	82.30	61.49	72.35
FCN-8s	ResNet101	94.86	92.71	98.19	62.09	94.09	80.25	83.47	65.02	75.23
DeepLabv3+	ResNet50	95.4	92.6	98.2	57.9	95.2	77.2	85.9	68.1	76.8
DeepLabv3+	ResNet101	95.5	93	98.4	58.8	95.3	78.5	86	68.8	78.4

Table 3: mIoU in percentage of Wider Resnet38 on **Cityscapes** Dataset.

5.2.1 FCN-8s

FCN is one of the earliest breakthroughs in the field of semantic segmentation. We plan to use it as one of the pipelines as it is one of the simplest models that can be used with multiple backbones. By modifying backbones with latest research discoveries, we can have different model performances. As mentioned before, a context prior layer [4] can be applied on different ResNet to achieve better contextual information, which in turn has the possibility of improving model performance. Therefore, a simple model like FCN is helpful for us to investigate the influence of backbone modification in our future experiment. Specifically, we use FCN-8s [8] with ResNet-50 backbone and ResNet-101 backbone as baseline models. Figure 3 illustrates the pipeline of FCN for image semantic segmentation.

Before FCN-8s, we do data augmentation on our dataset by cropping images into 512*1240, random horizontally flipping, random scaling, and Gaussian blur. We train FCN-8s using SGD, with learning rate of 0.01, momentum of 0.9 and weight decay of 0.0005. The loss function we use is CrossEntropyLoss.

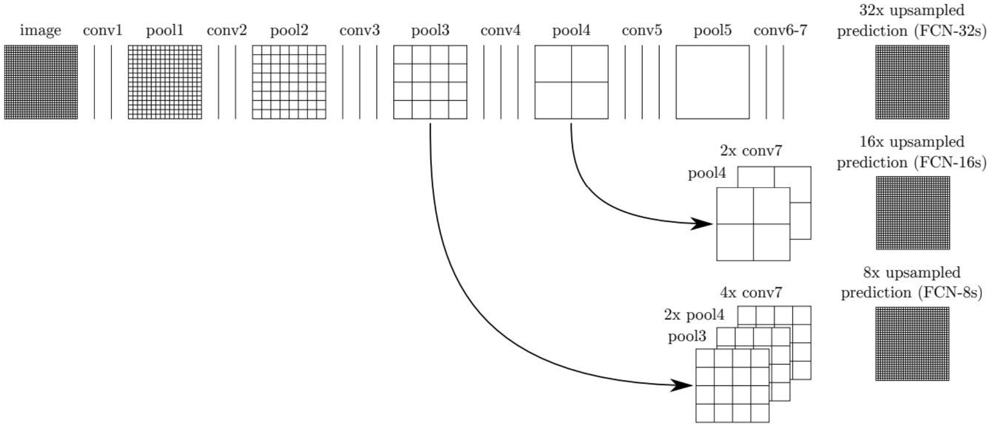


image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
32x upsampled prediction (FCN-32s)
2x conv7
pool4
16x upsampled prediction (FCN-16s)
4x conv7
2x pool4
pool3
8x upsampled prediction (FCN-8s)

Figure 3: Baseline Model: FCN Pipeline Illustration Retrieved From Paper[8]. Our baseline chooses Row 3 FCN-8s.

5.2.2 DeepLabv3+

We found out that DeepLabv3+ outperforms among the state-of-the-art models in mIoU and per-class IoU scores. Therefore, we view DeepLabv3+ as an outstanding achievement in urban-scene image semantic segmentation and plan to use it as a stronger baseline model in future experiments. As mentioned before, modification like positional prior[2] and context prior [4] can be generalized on top of various models. So in the future, we can incorporate and experiment with the influence of fruitful revision on one of the top models DeepLabv3+ to see whether model modifications have add-up influences on model performance results. Figure 4 illustrates the pipeline of DeepLabv3+.

Before DeepLabv3+, we do data augmentation on our dataset by cropping images into 512*1240, random horizontally flipping, random scaling, and Gaussian blur. We train DeepLabv3+ using SGD, with learning rate of 0.01, momentum of 0.9 and weight decay of 0.0005. The loss function we use is CrossEntropyLoss.

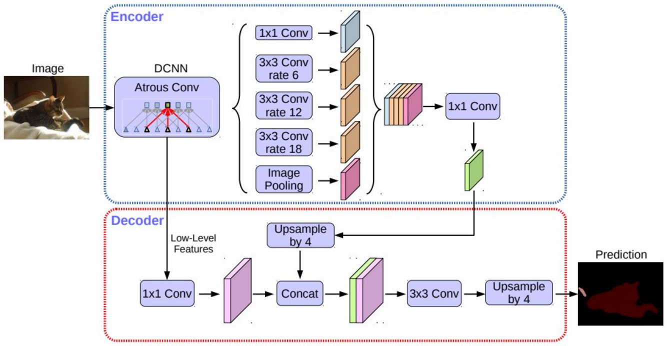


Figure 4: Baseline Model: DeepLabv3+ Pipeline Illustration Retrieved From Paper [15].

6 Current Experiment

In general, people believe that increasing depth increases the performance of a network. However, according to Wider or Deeper: Revisiting the ResNet Model for Visual Recognition, we found that simply increasing depth may not be the best way to increase performance, particularly given other limitations. Investigations into deep residual networks have also suggested that they may not be operating as a single deep network in fact, but rather as an ensemble of many relatively shallow networks. Thus, We experiment on a shallower architecture of residual networks Wider Resnet 38, which significantly outperforms much deeper models such as ResNet-101 on semantic image segmentation. The performance of Wider Resnet 38 is as follows.

Network	**sky**	**building**	**road**	**fence**	**car**	**sign**	**person**	**cyclist**	**mIoU**
Wider Resnet38	95.5	93.1	98.5	59.1	95.7	78.7	86.6	69.0	78.4

Table 3:mIoU and per-class IoU in percentage of Wider ResNet38 on **Cityscapes** Dataset.

7 Planned Experiment & Intermediate Conclusion

In our current experiment, we found out that the performance of the baseline models varies on different objects in the images. For example, the per-class intersection over Union (per-class IoU) is around 95 for sky and road objects; whereas, the per-class IoU for fence and road sign objects is less than 40. The current model is showing a biased performance on part of the classes and for sure has the space to be improved using various context information and spatial information.

As mentioned before, current research studies have fruitful modifications that are generalizable to different models and backbones [2][4]. In the next two weeks, we plan to continue our experiments by implementing the CPNet[2] and HANet[4] on top of our baseline models. We also plan to have another week in aggregating both modifications to see the results. At the same time, we as a group will keep researching possible modifications to add in the future experiment.

References

[1] Jordan, J. (2018). An overview of semantic image segmentation. Jeremy Jordan. Retrieved October 11, 2020, from https://www.jeremyjordan.me/semantic-segmentation/

[2] Choi, S., Kim, J. T., Choo, J. (2020). Cars Can't Fly Up in the Sky: Improving Urban-Scene Segmentation via Height-Driven Attention Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr42600.2020.00939

[3] Long, J., Shelhamer, E., & Darrell, T. (2015, March 08). Fully Convolutional Networks for Semantic Segmentation. Retrieved October 11, 2020, from https://arxiv.org/abs/1411.4038

[4] Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N. (2020). Context Prior for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12416-12425).

[5] Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical Multi-Scale Attention for Semantic Segmentation. arXiv preprint arXiv:2005.10821.

[6] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223). Retrieved October 11, 2020, From https://www.cityscapes-dataset.com/downloads/

[7] G. Neuhold, T. Ollmann, S. R. Bulò and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 5000-5009, doi: 10.1109/ICCV.2017.534.

[8] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In Proc. of the International Conference on Computer Vision (ICCV), 2019

[9] Olafenwa, J., Olafenwa, M. (2018). FastNet. arXiv preprint arXiv:1802.02186.

[10] Li, X., Jie, Z., Wang, W., Liu, C., Yang, J., Shen, X., . . . Feng, J. (2017). FoveaNet: Perspective-Aware Urban Scene Parsing. 2017 IEEE International Conference on Computer Vision (ICCV). doi:10.1109/iccv.2017.91

[11] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Anal- ysis and Machine Intelligence (TPAMI), 39(12):2481–2495, 2017. 1, 3

[12] Liu, W., Rabinovich, A., Berg, A. C. ParseNet: Looking wider to see better. arXiv 2015. arXiv preprint arXiv:1506.04579.

[13] Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1520-1528).

[14] Valada, A., Vertens, J., Dhall, A., Burgard, W. (2017, May). Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4644-4651). IEEE.

[15] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[16] Csurka, G., Larlus, D., Perronnin, F. (2013). What is a good evaluation measure for semantic segmentation? BMVC.

[17] cityscapesScripts: https://github.com/mcordts/cityscapesScripts

[18] M. Yang, K. Yu, C. Zhang, Z. Li and K. Yang, "DenseASPP for Semantic Segmentation in Street Scenes," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 3684-3692, doi: 10.1109/CVPR.2018.00388.

[19] Li, X., Jie, Z., Wang, W., Liu, C., Yang, J., Shen, X., ... Feng, J. (2017). Foveanet: Perspective-aware urban scene parsing. In Proceedings of the IEEE International Conference on Computer Vision (pp. 784-792).

[20] ZWu, Tianyi, et al. "Cgnet: A light-weight context guided network for semantic segmentation." arXiv preprint arXiv:1811.08201 (2018).

[21] Zifeng Wu, Chunhua Shen, and Anton van den Hengel, Wider or Deeper: Revisiting the ResNet Model for Visual Recognition, Pattern Recognition(119-133).

[22] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

[23] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In:CVPR. (2017)

[24] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.

Appendix: Timeline & Division of Work

Project Timeline				
Week	**Date**	**Week Debrief**	**Division of Work**	**Important Event**
12	Oct.12-Oct.18	**[DONE]**1.Literature Review **[DONE]**2. Define Work Scope	1. Each member does at least 3 researches on related work	
13	Oct.19-Oct.25	**[DONE]**1. Exploratory Data Analysis **[DONE]**2. Data Loading	1. Each member load data and share exploratory data analysis result	
14	Oct.26-Nov.1	**[DONE]**1. Organize data loading code **[DONE]**2. Data Pre-processing **[DONE]**3. Evaluation metrics research	1. Each member research evaluation metrics 2. Data pre-processing: haiyunta & yudiz 3. Code organization: haijingf & yihanq	
15	Nov.2-Nov.8	**[DONE]**1. Draft midterm report **[DONE]**2. Run baseline FCN-8s ResNet-50/101 **[DONE]**3. Run baseline DeepLabv3+ ResNet-50/101	1. FCN ResNet101: yudiz 2. FCN ResNet50 haiyunta 3. DeepLabv3+ResNet50: haijingf 4.Deeplabv3+ ResNet101: yihanq	
16	**Nov.9-Nov.15 (Current Week)**	**[DONE]**1. Finalize midterm report **[DONE]**2. Continue to run baseline DeepLabv3+ **[DONE]** **[IP]**3. Experiment modifications with Wider ResNet38	1. DeepLabv3+ResNet50: haijingf 2.Deeplabv3+ ResNet101: yihanq 3. Wider ResNet38 modification: yudiz & haiyunta	Midterm Report DUE Nov.10
17	Nov.16-Nov.22	1. Continue experiment HANet & CPNet	TBD	
18	Nov.23-Nov.29	1. Start drafting final report 2.Continue experiment HANet & CPNet	TBD	
19	Nov.20-Dec.6	1. Continue drafting final report 2. Aggregate HANet & CPNet and Finalize experiment result	TBD	Draft Final Report DUE Nov.27
20	Dec.7-Dec.13	1.Organize and document code and results 2. Finalize project deck 3. Finalize project video 4. Finalize final report	TBD	Project Video DUE Dec.9 Peer Review DUE Dec.12 Final Report DUE Dec.13

Table 3: Timeline Table for Project Schedule