CS 766 Project Proposal

Yun-Shiuan Chuang
ychuang26@wisc.edu

Varun Sreenivasan
vsreenivasan@wisc.edu

Jacob Lorenz
jlorenz2@wisc.edu

February 2021

1 Introduction

1.1 Briefly explain what problem you are trying to solve.

In this project, we aim to implement the instance segmentation model, Mask R-CNN [4], in the domain of autonomous vehicles. We also seek to make improvements to the model through modifications that were not explored in the original paper.

1.2 Why is this problem important? Why are you interested in it?

The task of instance segmentation is extremely important for critical computer vision tasks such as those involved with autonomous vehicles. For instance, if we solely use object detection in the autonomous vehicles, there is a chance that the bounding boxes of multiple cars may overlap and this will confuse the self-driving vehicle. Instance segmentation helps overcome this flaw. The ability to detect the spatial boundaries of objects down to pixel level detail instead of broadly sensing their location could mean the difference between the vehicle safely navigating its way through and the vehicle striking other cars or objects. In a world of high velocity traffic and unpredictability, the smallest details can have some of the most serious consequences, good or bad.

Instance segmentation will continue to play a crucial role in various computer vision tasks long into the future. Autonomous vehicles, medical imaging, facial recognition, robotic procedures; all of these fields rely on being able to accurately differentiate object instances, and we are fascinated by its long term potential. For that reason, we want to spend the next several months exposing ourselves to various flavors of the Mask R-CNN algorithm and developing a strong understanding of its capabilities and current limitations as it relates to instance segmentation.

1.3 What is the current state-of-the-art?

[3] reviewed the current state-of-the-art models for instance segmentation. Briefly summarized here, there are 4 main families of instance segmentation models: 1) classification of mask proposals, 2) detection followed by segmentation, 3) labelling pixels followed by clustering, and 4) dense sliding window methods. The second family detection followed by segmentation) was deemed the only one that suits for real time applications due to its fast run-time, which we consider an essential requirement for autonomous cars. Within this family, **Mask R-CNN** [4] and **Mask Scoring R-CNN** [5] are the two models with the highest performance as evaluated by the average precision (AP) on the COCO dataset [6] (39.6% for Mask Scoring R-CNN and 37.1% for Mask R-CNN). Mask R-CNN is composed of a CNN backbone (e.g, ResNet), followed by a RoIAlign layer, and three prediction heads: 1)instance classification, 2) bounding box prediction, and 3) mask prediction. The mask prediction head generates the binary mask for instance segmentation. Scoring R-CNN is an extension to Mask R-CNN with an additional MaskIoU head module. Although having a slight increase in AP, Scoring R-CNN is more complicated than Mask R-CNN due to its more complicated model architecture. Given the computational resource and time constraint we have, we aim to based our work on the light-weighted Mask R-CNN.

1.4 Are you planning on re-implementing an existing solution, or propose a new approach?

Our plan is to do both. First, we are going to implement the Mask R-CNN as in the original paper [4] and replicate the results on the **COCO Dataset** [6], which contains 91 generic objects types (e.g., cat, pizza, hat). After replicating the results on this data set, we then plan to implement this model on autonomous vehicle data sets such as **Cityscapes Dataset** [1] (urban street scenes) and the **Indian Driving Dataset** [2], which is presumably more challenging because the road scenes are more diverse and and unstructured.

The next stage of our project involves improving the model performance. Possible modifications includes tweaking the model architecture (e.g., the R-CNN backbones) and exploring different training techniques (e.g., multi-scale train/test, horizontal flip test) in an attempt to determine which combinations work best in different scenarios and improve the overall performance.

1.5 If you are proposing your own approach, why do you think existing approaches cannot adequately solve this problem? Why do you think your solution will work better?

The authors of the Mask R-CNN [4] have acknowledged that they haven't explored many possible ways to optimize the model performance. To our knowledge, since this is an emerging topic in the field, many of the optimization works can be done but haven't been done. We don't claim certainty that our solution and modifications will work better than the current state of the art; however, for the sake of curiosity and exploration we think that at least attempting to change various aspects of the implementation ought to give interesting results and enable a discussion as to why certain modifications outperform others in one or many different domains.

1.6 How will you evaluate the performance of your solution? What results and comparisons are you eventually planning to show? Include a time-line that you would like to follow.

The creative part of our project revolves around the various modifications we aim to make to the model architecture, as well as the domain of data (i.e., autonomous cars) we are using the models on. We plan to deliver results in the form of a NxM matrix where N represents the number of different versions of the Mask R-CNN algorithm we use, and we will test each on M different data sets. To evaluate the model performance, we plan to make use of a common metric for instance segmentation called average precision (AP). Please see the time table below for a tentative plan of our expected progress.

2 Time table

See Table 1 for a tentative time table.

Date	Task	Milestone
02/22	• Write the proposal	Project proposal due (02/24)
03/01	• Replicate the paper results – download the COCO data set – set up the ML environment (e.g., PyTorch, GPU config) – read and understand the scripts – train the model with COCO data set – evaluate the model with COCO test set	
03/08	• Implement model and train on Cityscapes Dataset using default parameters • Evaluate the model	
03/15	• Implement model and train on Indian Driving Dataset using default parameters • Evaluate the model • Write mid-term report	
03/22	• Write mid-term report • Create a list of modifications to explore.	Project mid-term report due (03/24)
03/29	• Implement and re-train model with new parameters	
04/05	• Implement and re-train model with new parameters	
04/12	• Implement and re-train model with new parameters	
04/19	• Tabulate results to assess the impact of modifications. • Use the best model to perform Instance Segmentation on video stream.	
04/26	• Project webpage	
05/03	• Project webpage	Project webpage due (05/05)

Table 1: The Time Table

References

[1] 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.

[2] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar. IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751, 7.

[3] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International Journal of Multimedia Information Retrieval*, 9(3):171–189, Jul 2020.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[5] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.