CSE 509: Digital Video Processing Road Surface Segmentation and Classification Group 8

Krutikkumar Parmar
1225491150
Arizona State University
Tempe, USA
kmparmar@asu.edu

Vedant Parikh
1225655171
Arizona State University
Tempe, USA
vparikh5@asu.edu

Janam Vaidya
1222377325
Arizona State University
Tempe, USA
jvaidya4@asu.edu

Rishabh Pandat
1222158509
Arizona State University
Tempe, USA
rpandat@asu.edu

***Abstract*—Our goal with this work is to perform road detection with the differentiation of surface variations, in addition to concomitant surface damage detection. We also aim to show that it is possible to use passive vision (dashboard cameras) to detect road damage. We used two methods for our problem and compared their results. We believe that the contributions and differentials of our approach are detecting and recognizing the different types of surfaces (asphalt, other pavement, and unpaved), detecting potholes and water puddles on the road, even on different types of surfaces, and in conjunction with the previous points, detect other damage and patterns on the road, all with the same approach**

***Index Terms*—Road surface detection, Road segmentation, Road surface obstacles detection**

I. Introduction

In the area of vehicle and robotic navigation, a significant unresolved challenge is the detection of obstacles and finding a reliable path. The research has been mostly focused on detecting the road and the datasets used in the research were mostly from developed countries that contained roads with very good conditions [1]. In this project, our main focus was on detecting a path but we also focused on finding the surface type of the road and whether it contains any obstacles or not as well as markings on the road. Detecting the road surface type is very useful in Autonomous Vehicles. Depending on the type of road surface, it is important to change the vehicle speed and the autonomous system should adapt its way of driving. The safety of the passenger and the conservation of the vehicle both depend on how the autonomous system drives on different types of roads. For a similar reason, it is also very important to detect potholes and water-puddles on the road. We used two different approaches to solve the problem. Our first approach used Convolutional Neural Network(CNN) and our second approach used Vision Transformer (ViT). We used a subset of the Road Traversing Knowledge (RTK) dataset with instance segmentation ground truth masks. In the end, we compared the results of these two approaches.

II. Related Work

Our project was inspired by the work of T. Rateke et al. [2]. The dataset that we have used was created by them. For semantic segmentation, they used U-NET architecture with ResNet34. They used the transfer learning method to get better performance. In some of the semantic segmentation applications, it has been shown that Mask-RCNN has performed better than U-Net [3], [4]. The Vision Transformer has also been proven to be a state-of-the-art model which gives us similar results as Convolutional Neural Networks while using very few resources. This led us to the idea of using Mask-RCNN and Vision Transformer to solve our problem.

III. Materials and Methods

A. Dataset

A subset of the Road Traversing Knowledge (RTK) dataset with instance segmentation ground truth masks was used in this project [2]. The dataset consists of 710 images of 12 different classes with their ground truth segmented masks. Fig. 1 and Fig. 2 show two original images and their ground truth masks. The classes are as follows :

- *Asphalt*
- *Paved*
- *Unpaved*
- *Markings*
- *Speed-Bump*
- *Cats-Eye*
- *Storm-Drain*
- *Hole-Cover*
- *Patch*
- *Water-Puddle*
- *Pothole*
- *Cracks*

We have split the dataset into 9:1 ratios which gave us 639 images in the training dataset and 71 images in the testing dataset.

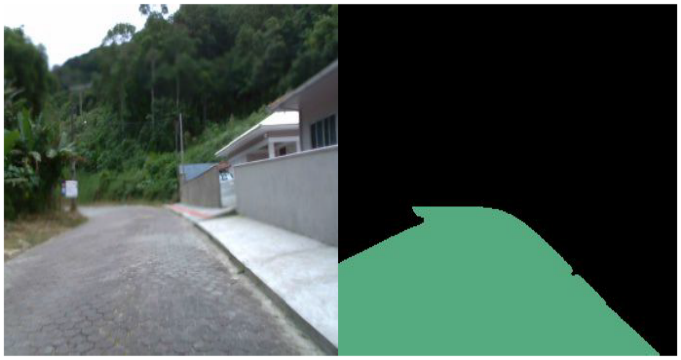


Fig. 1: Original Image with Paved Road and its ground truth mask

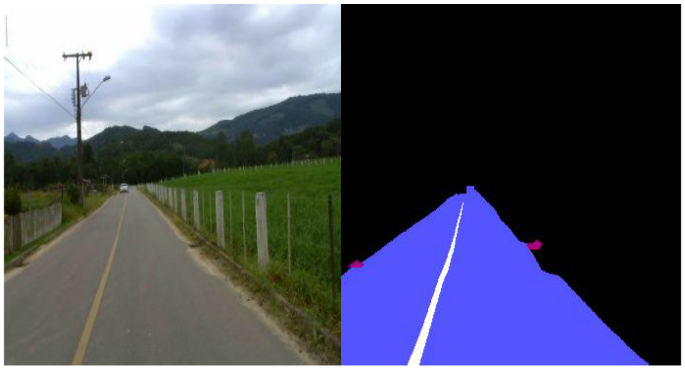


Fig. 2: Original Image with Asphalt Road, Road markings,water puddle and its ground truth mask

B. Approach 1

In the first approach, we used Mask Region-Based Convolutional Neural Network (Mask-RCNN) [6] since we had to do semantic segmentation. So the output should also give us the mask of the detected region. The original image and its ground truth mask are given as inputs while training the Mask-RCNN. We have used weights from a network that was pre-trained on the MS COCO dataset. The output of Mask-RCNN is the predicted mask of the region and the predicted label of that particular region.

C. Approach 2

In the second approach, we used Vision Transformer(ViT) [7]. As shown in Fig.4, the original image is split into patches, in our case patch size is 16x16. These image patches are flattened and a lower-dimensional linear embedding is created from them. Then the patches are fed as an input to a state-of-the-art transformer encoder structure consisting of self-attention and multi-layered perceptron layers. The output of the transformer is also the same as Mask-RCNN which is predicted masks of regions and predicted labels of those regions.

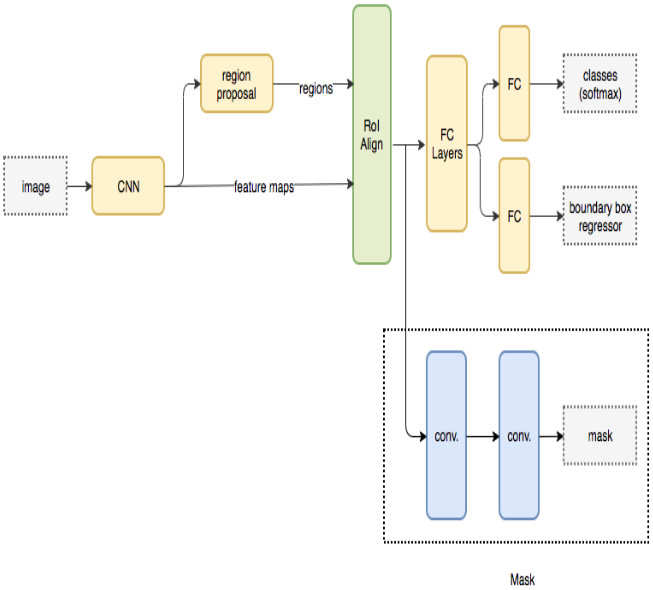


Fig. 3: Architecture of Mask-RCNN [5]

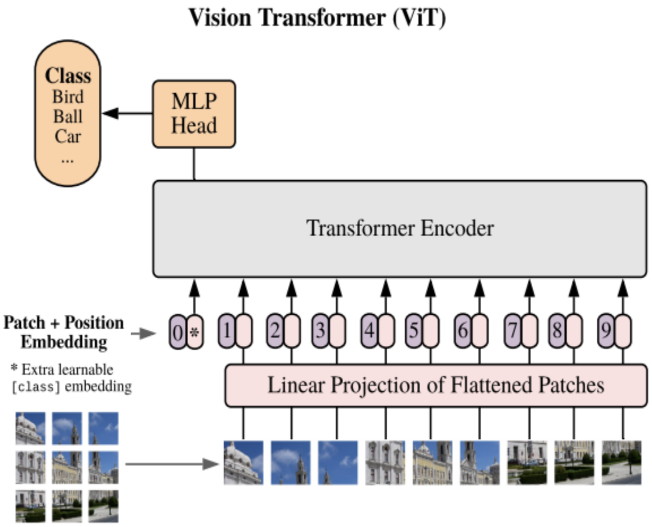

Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Fig. 4: Architecture of ViT [8]

IV. Results

A. Mean IoU

We have used Mean Intersection over Union(IoU) metrics to compare the results of our approaches. In order to calculate the IoU of the predicted mask, we first calculate the intersection of the predicted mask and ground truth mask and then calculate the union of the predicted mask and ground truth mask, and then take the division of these two. This gives an IoU for the predicted mask. Finally, we will take mean of all the IoUs of all the classes.

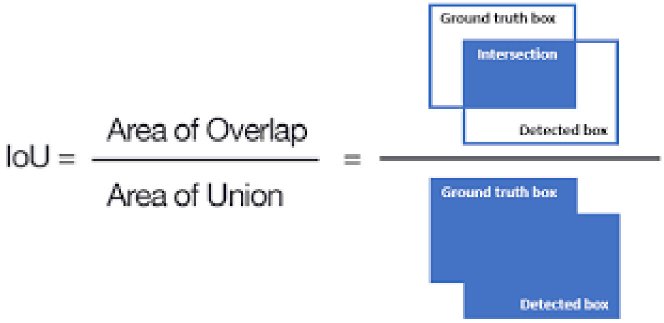


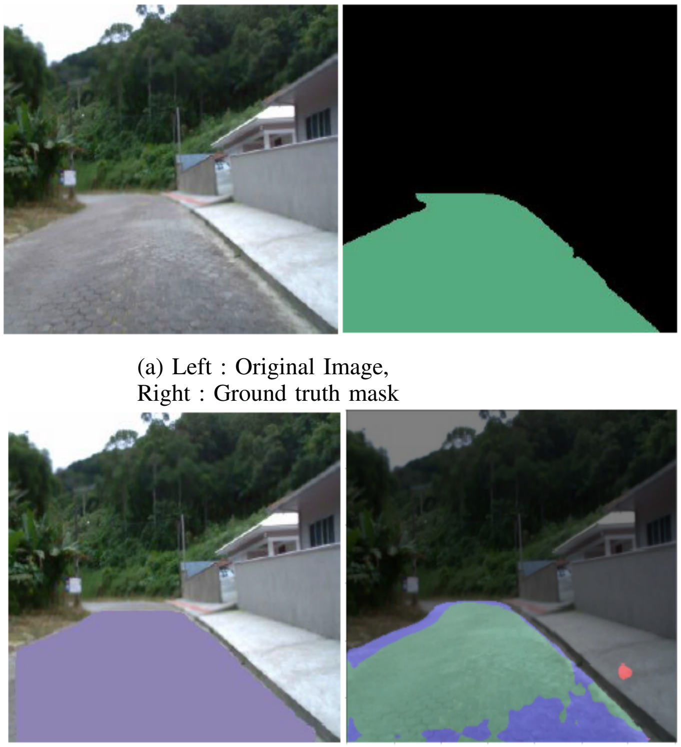
Fig. 5: Intersection over Union (IoU)

B. *Comparison*

TABLE I: Comparision of Mean IoU of all the classes between Approach 1 and Approach 2

Class	***Mask RCNN***	***ViT***
roadAsphalt	0.471	0.638
roadPaved	0.573	0.545
roadUnpaved	0.537	0.657
roadMarking	0.34	0.12
speedBump	0	0
catsEye	0.108	0
stormDrain	0.67	0
manholeCover	0	0
patchs	0.036	0.08
waterPuddle	0.061	0
pothole	0.004	0.016
cracks	0.038	0.094

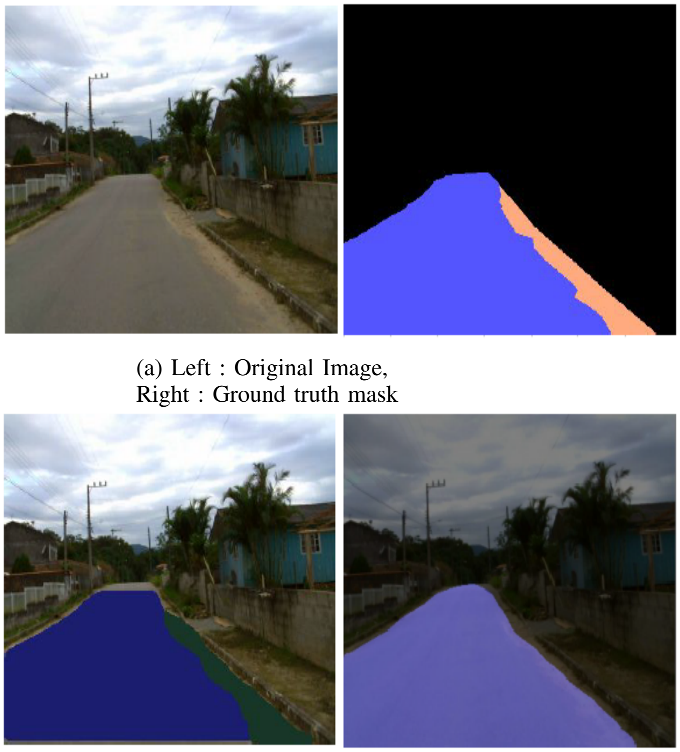
As we can observe from Table 1, the Mean IoU for *roadAsphalt*,*roadPaved* and *roadUnpaved* is very high. The reason behind this is that most of the images contained these three classes and the area of these classes was higher than the rest of the classes in all the images. The models learned these three classes much better than others. For *roadAsphalt* and *roadUnpaved*, the mean IoU of ViT model is much higher than Mask RCNN but for other classes such as *catsEye*, *stormDrain*, etc, the mean IoU is very less compared to Mask RCNN. This is because the ViT requires much more data than Mask RCNN. So the ViT trained in a much better way for classes with a higher amount of data than classes with a lower amount of data. Overall, we can say that Mask RCNN gave us better results than ViT since it was able to predict all the classes more accurately.



(a) Left : Original Image,
Right : Ground truth mask

(b) Left : Mask RCNN prediction,
Right : ViT preditcion

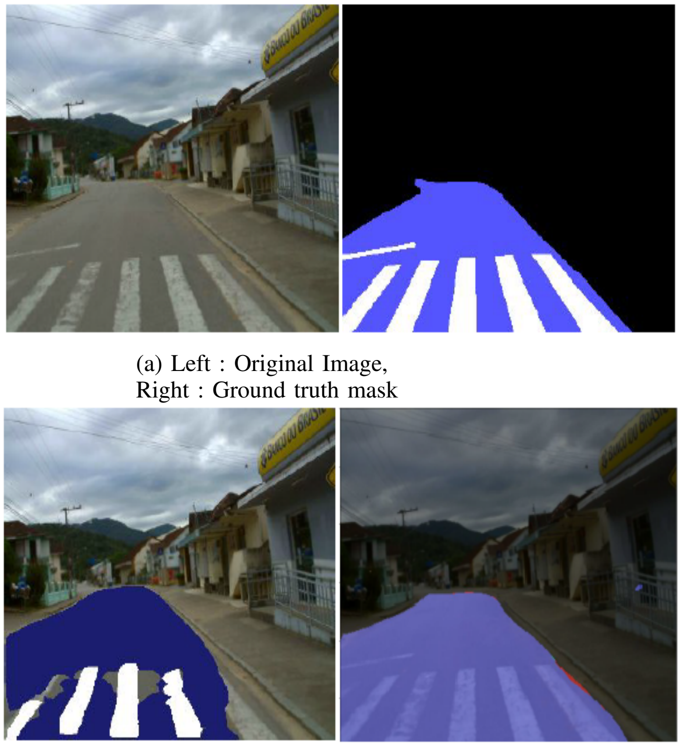
Fig. 6



(a) Left : Original Image,
Right : Ground truth mask

(b) Left : Mask RCNN prediction,
Right : ViT preditcion

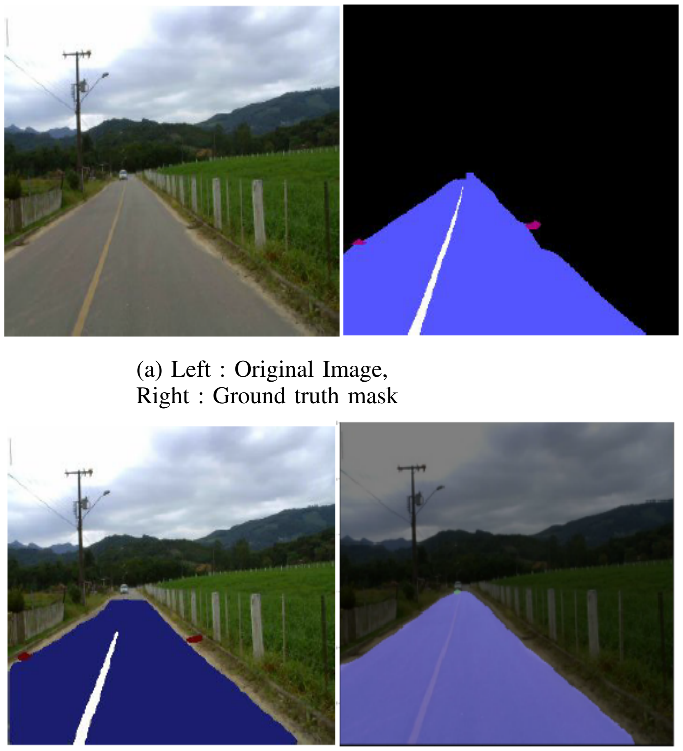
Fig. 7



(a) Left : Original Image,
Right : Ground truth mask

(b) Left : Mask RCNN prediction,
Right : ViT preditcion

Fig. 8



(a) Left : Original Image,
Right : Ground truth mask

(b) Left : Mask RCNN prediction,
Right : ViT preditcion

Fig. 9

References

[1] J. Fritsch, T. Kuehnl, and A. Geiger, ‘A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms’, in International Conference on Intelligent Transportation Systems (ITSC), 2013.

[2] T. Rateke and A. von Wangenheim, “Road surface detection and differentiation considering surface damages,” Autonomous Robots, vol. 45, no. 2. Springer Science and Business Media LLC, pp. 299–312, Jan. 11, 2021. doi: 10.1007/s10514-020-09964-3.

[3] T. T. P. Quoc, T. T. Linh and T. N. T. Minh, ”Comparing U-Net Convolutional Network with Mask R-CNN in Agricultural Area Segmentation on Satellite Images,” 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), 2020, pp. 124-129, doi: 10.1109/NICS51282.2020.9335856.

[4] Madeleine S. Durkee, Rebecca Abraham, Junting Ai, Jordan D. Fuhrman, Marcus R. Clark, and Maryellen L. Giger ”Comparing Mask R-CNN and U-Net architectures for robust automatic segmentation of immune cells in immunofluorescence images of Lupus Nephritis biopsies”, Proc. SPIE 11647, Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XIX, 116470X (5 March 2021); https://doi.org/10.1117/12.2577785

[5] J. Hui, “Image segmentation with mask R-CNN,” Medium, 06-Sep-2022. [Online]. Available: https://jonathan-hui.medium.com/image-segmentation-with-mask-r-cnn-ebe6d793272. [Accessed: 03-Dec-2022].

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ‘Mask R-CNN’. arXiv, 2017.

[7] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, 2020. doi: 10.48550/ARXIV.2010.11929.

[8] Papers with code - vision transformer explained. Explained — Papers With Code. (n.d.). Retrieved December 3, 2022, from https://paperswithcode.com/method/vision-transformer