Enhancing Autonomous Vehicle Perception: A Transformer Based Approach for Aerial Road Representation

PURDUE
UNIVERSITY

Authors: William Stevens, Vishal Urs, Karthik Selvaraj, Gabriel Torres, Gaurish Lakhanpal

Mentors: Prof. Edward J. Delp, Prof. Carla Zoltowski

INTRODUCTION

With the prevalence of autonomous vehicles, the computer vision algorithms utilized for autonomous driving must be robust and accurate to assess road features through images captured in real time. We explore a new approach to lane segmentation that produces a novel output to autonomous driving software: an aerial representation of the road structure derived from ground-level images. The model design involves an optimized neural network structure containing multiple transformers, inspired by the LaneSegNet and BEVFormer architectures, which together produce an algorithm that can accurately and efficiently translate ground-level captures into an aerial view of the road features. This will provide autonomous driving software with more useful information and context, such as the position and direction of road lanes relative to the vehicle. This new structure will allow our model to accurately perform the important vision tasks for autonomous driving at a low computation cost, while producing information relevant to the autonomous driving pipelines used in the real world.

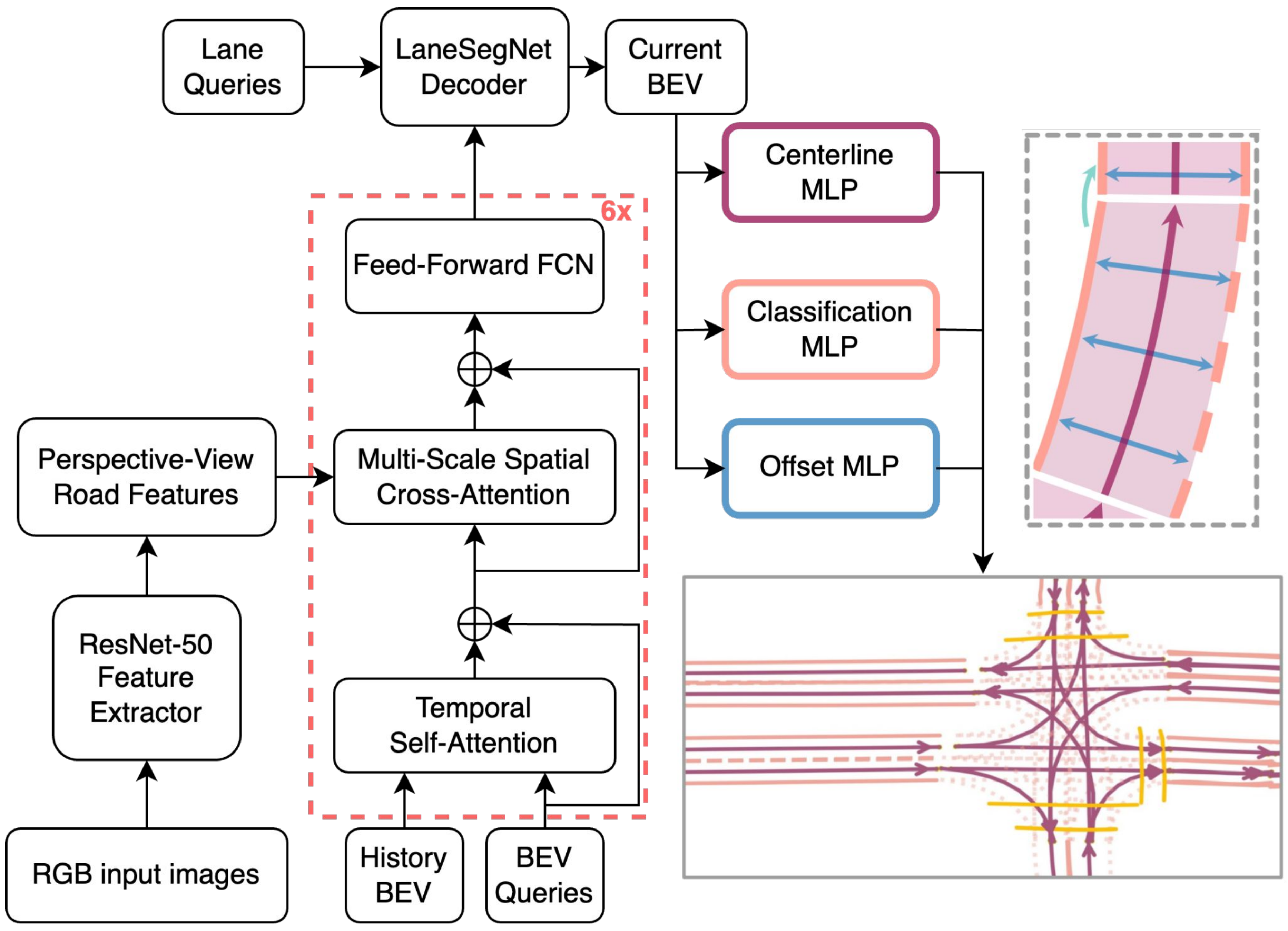
Lane Queries
LaneSegNet Decoder
Current BEV
Centerline MLP
Classification MLP
Offset MLP
6x
Feed-Forward FCN
Perspective-View Road Features
Multi-Scale Spatial Cross-Attention
ResNet-50 Feature Extractor
Temporal Self-Attention
RGB input images
History BEV
BEV Queries

Figure I: LaneSegNet Model Architecture

DATA

We are using the OpenLane-V2 dataset for our project this semester because it contains 2,000 road scenarios in the form of 15-second videos, with seven camera angles in a 360º span around the car (See Figure II). For each of these videos, OpenLane-V2 contains a groundtruth annotation representative of what a satellite view of the road scenario looks like. This Bird's-Eye-View perspective will be used as a measurement for how accurate the model's output is, as it will be what the model is learning to replicate.

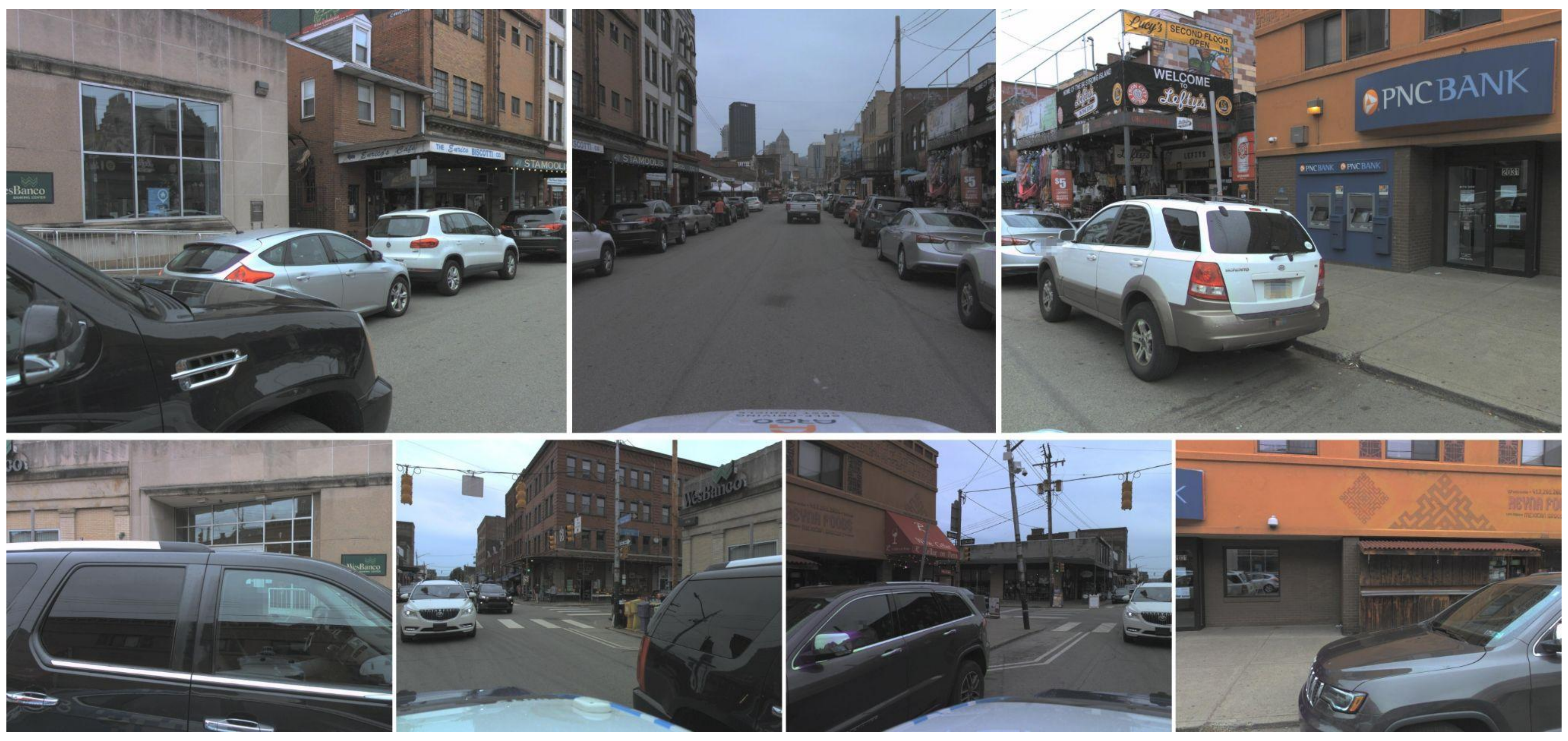

Figure II: Perspective View Road Scenario

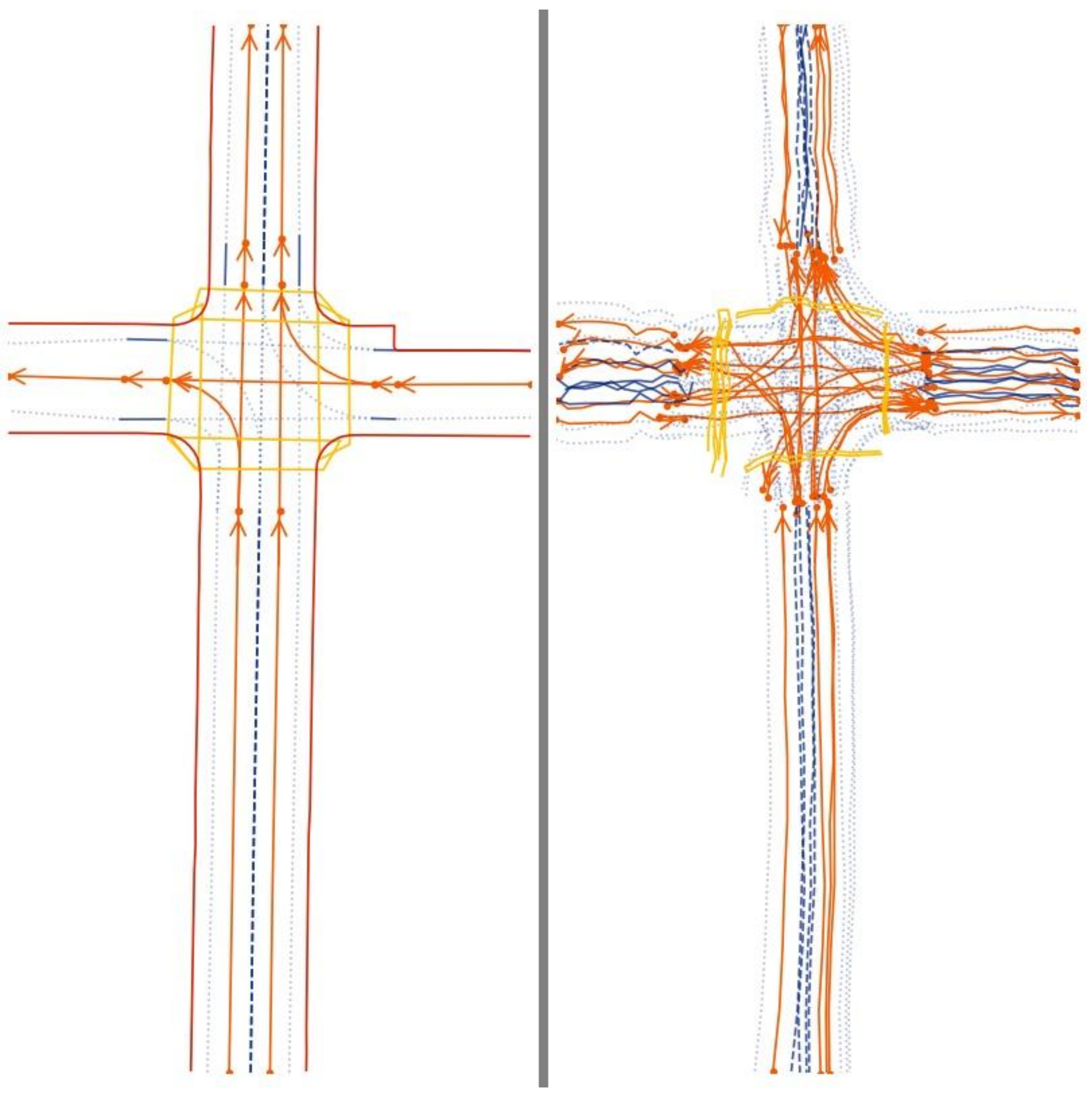

Figure III: Bird's-Eye-View Groundtruth (Left) and Model Output (Right)

METHODS

The LaneSegNet architecture takes in as input the seven perspective view images at a given frame in the video. These are passed through a ResNet-50 model, which has 50 convolutional layers with residual connections, to extract perspective view features from the road scenario. This model will learn to extract features representative of the patterns, lines, and edges in the images. These perspective view features are sent to the encoder component of the LaneSegNet transformer. The encoder is designed to transform these perspective view features into bird's-eye-view features, by incorporating temporal information from previous frames with the spatial information from the perspective features. These bird's-eye-view features are sent to the decoder, which performs lane attention through a heads-to-regions mechanism. Then, three perception branches use Multi-Layered Perceptrons to detect the lanes' centerlines, offsets, and lane types to produce the final output as shown in Figure I.

RESULTS

We have managed to implement and train the above architecture to a notable extent. The results we have obtained are based on a shorter training time of 4-5 hours for 2 epochs. The perspective features are pre-processed in an identical manner to that of the original LaneSegNet architecture and then processed through the BEVformer segment. The lane mapping is developed based on the outputs of the LaneSegNet Decoder combined with the various MLPs in accordance to the particular area of focus. The results we have obtained have 7% MaP scores. While this may seem low, it is very promising considering that we have only run two epochs. The centerlines are the easiest to develop, while the intersection boundaries require stronger segmentation information to be depicted accurately. We are confident that with more extensive training, the architecture will be able to learn significantly more information and develop a mapping much more similar to the ground-truth depicted on the left side of Figure III.

Purdue Undergraduate Research Conference Spring 2024