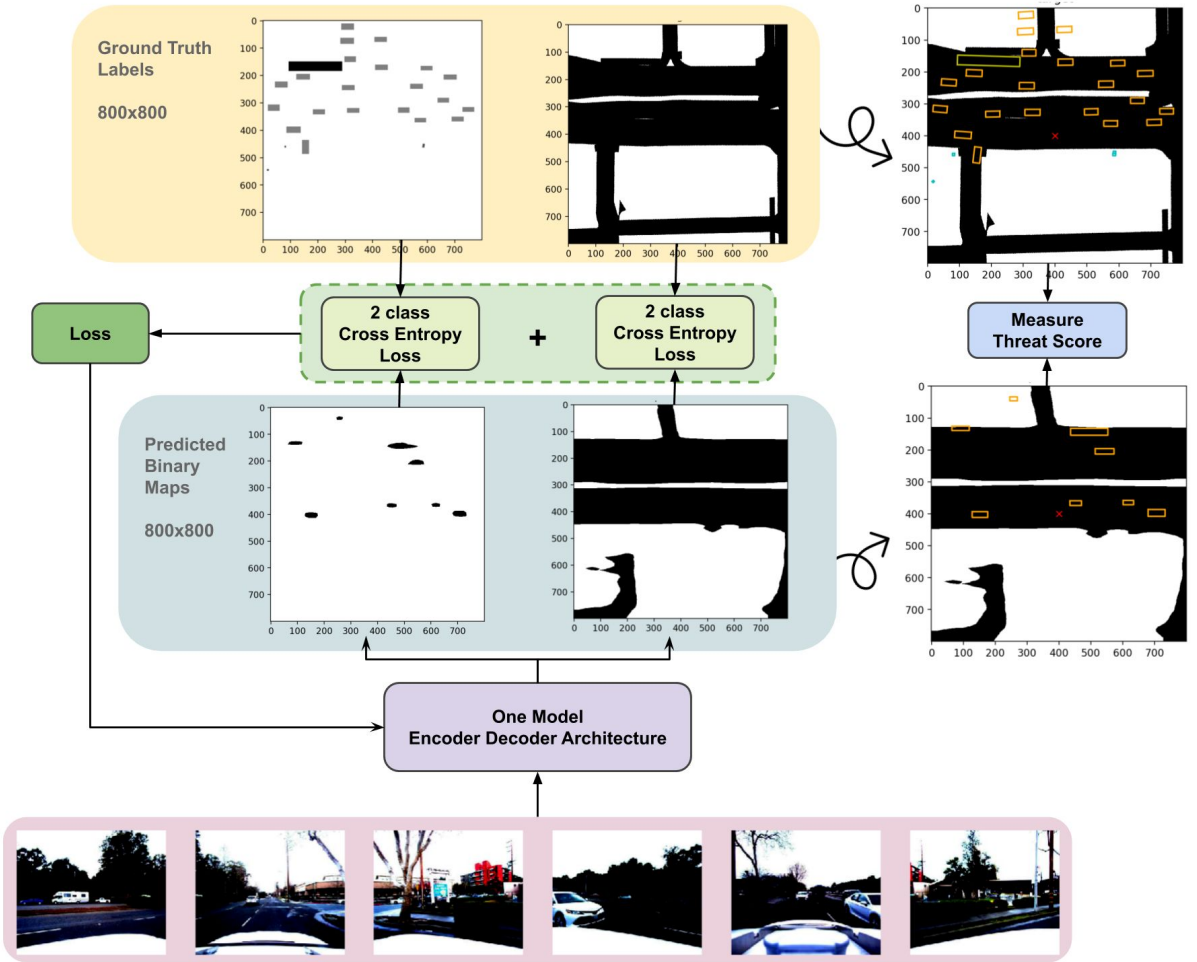
HexaLayout

From six images to roadmap and objects

Team: 33, LetMeDrive
Member: Shiqing Li, Ying Jin, Xiao Li

Task Re-definition

- Train bounding box and road map together
- Turn bounding boxes into a 800 x 800 matrix map
- Two predictions:
  - binary road map
  - binary object map
- Inspired by MonoLayout[1]



1. Mani et al. *MonoLayout: Amodal scene layout from a single image,* arXiv

Self supervised Depth

- **Monodepth2**[1]:

  Use a depth network and a pose network to predict disparity maps for monocular images.

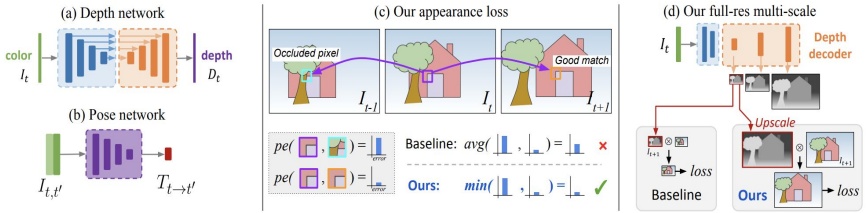
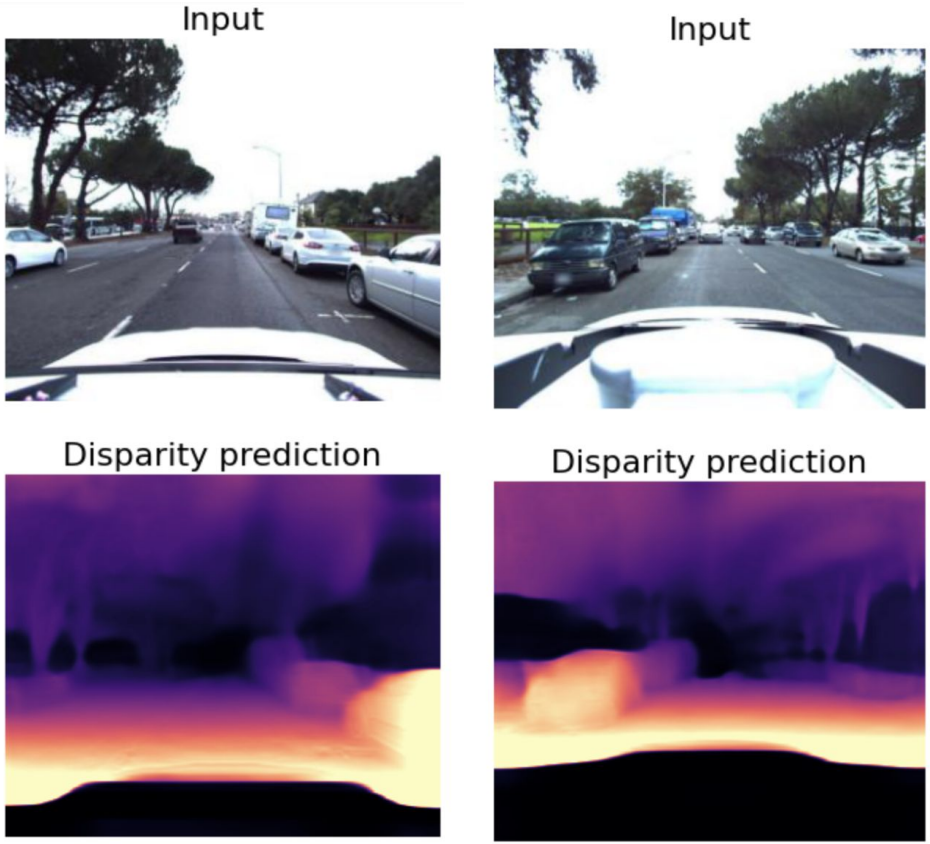


Image from [1]



1. Godard et al. *Digging into Self-Supervised Monocular Depth Prediction,* ICCV 2019

Model Architecture

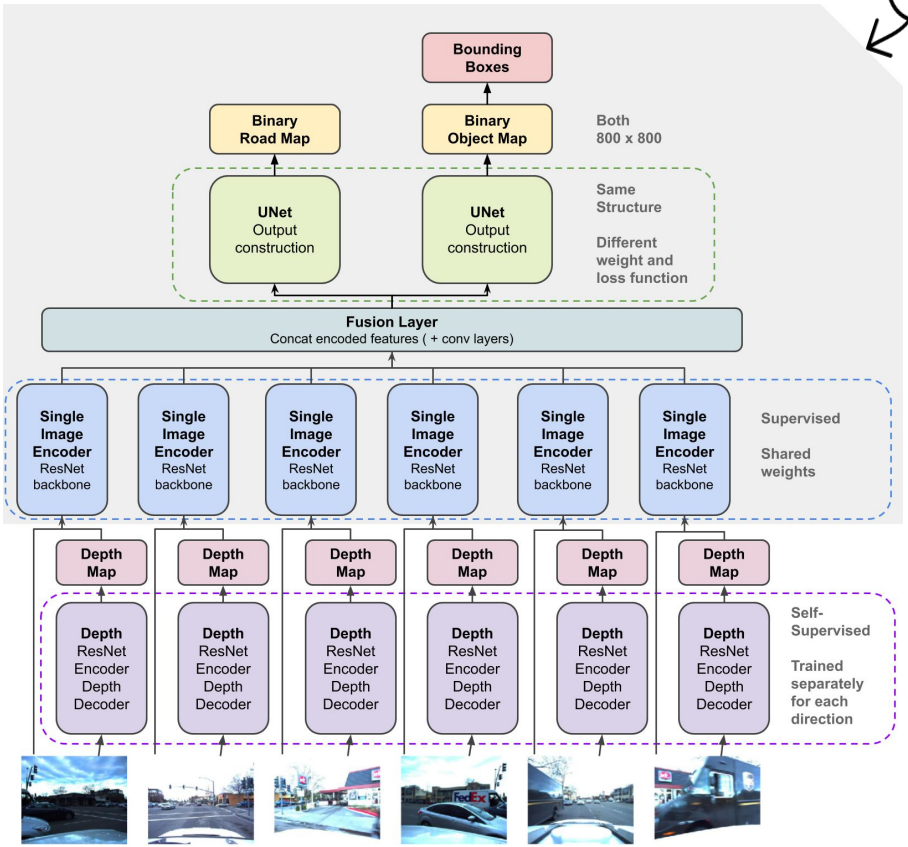


Figure: Model architecture for Variation 1 - Depth as Input

HexaLayout Architecture

- 6 shared weights Encoders (ResNet)
- Concat encoded features
- 2 Decoders to perform binary image construction (UNet)

Variation 1: Depth as input

- use Depth Map, generated through self supervised learning, as the 4th channel input

Variation 2: Encoded Depth Feature

- Additional 6 Single Image Encoder to encode depth map
- Concatenate along channels at fusion layer

Variation 3: with Discriminator

- Patch-level discriminators (Pix2Pix) consist of 4 convolutional layers

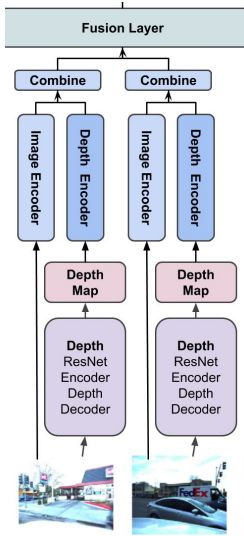


Figure: Illustration for Variation 2 - Encoded depth feature

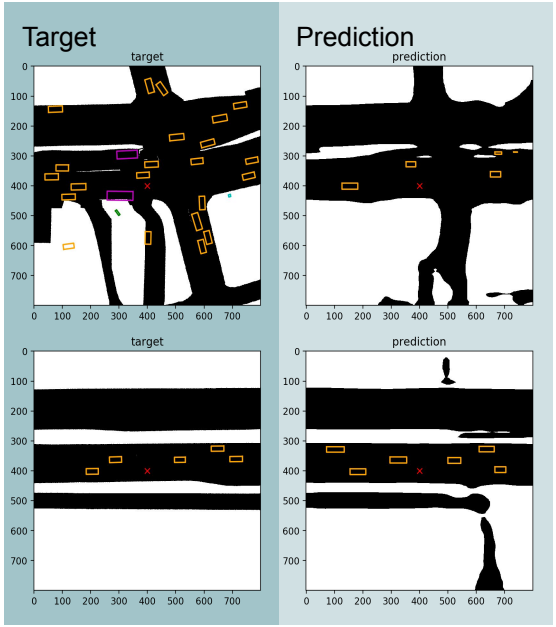
Results on validation set

Road Map	AUC	Accuracy	Threat Score
HexaLayout Architecture	0.9267	0.8828	**0.7523**
Variation 1: Depth as input	0.9283	0.8818	0.7445
Variation 2: Encoded Depth Feature	0.9302	0.8869	0.7490
Variation 3: with Discriminator	0.9067	0.8631	0.7120

Object Map (/bounding box)	AUC	Accuracy	Threat Score (box)
HexaLayout Architecture	0.9069	0.9676	**0.0343**
Variation 1: Depth as input	0.9543	0.9696	0.0087
Variation 2: Encoded Depth Feature	0.7715	0.9718	0.0112
Variation 3: with Discriminator	0.9051	0.9695	0.0216

*Note: AUC and Accuracy for Cap Map is measured on the 800 x 800 binary matrix.

Examples of results

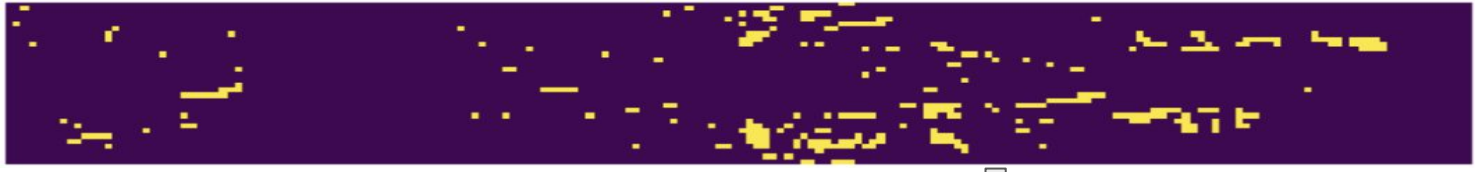


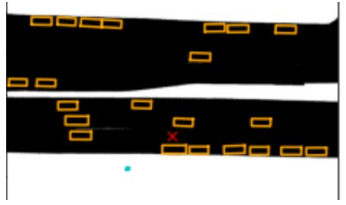
Results without shared encoder

Train by itself	Best TS
Road Map	0.7244
Bounding Box	0.0006

Analysis - Encoder Features

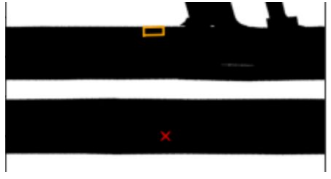
(1) Simple two directional horizontal road maps examples:



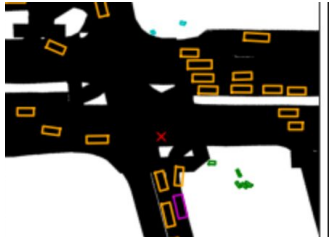


Many objects on road

Few objects on road



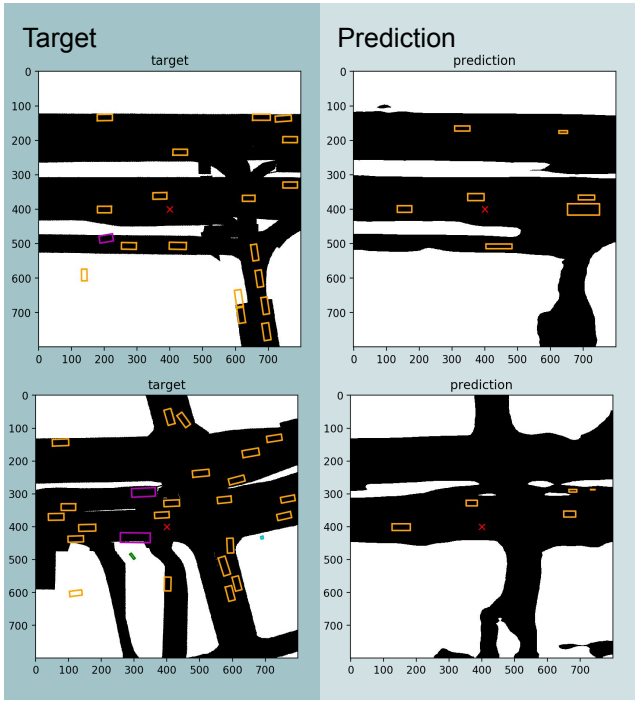
(2) Complex intersection road maps examples:



Many objects on road

Few objects on road

Analysis - Finding from Results



The bounding box prediction performs better for cars that are closer to the cameras (temporal info could help).

The road map prediction almost never predict cases with one horizontal lane (not enough seen in training).