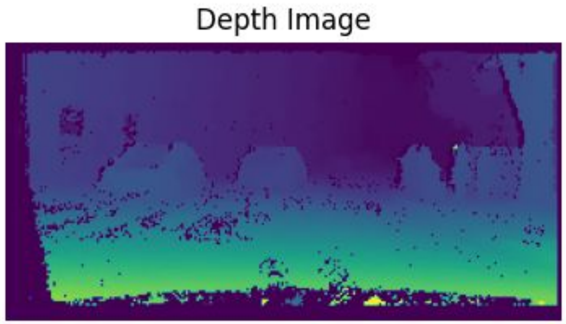
Multi Modal Segmentation

-Khushin, Parth Patel

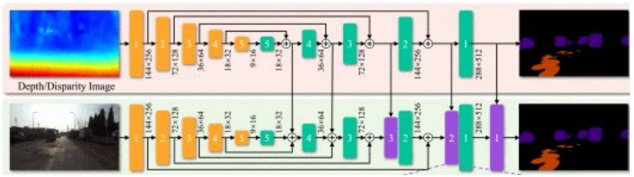
Problem Statement

- Multi-modal data involves information from various sensors or data sources (e.g., RGB images, depth maps, thermal images, LiDAR).
- For autonomous vehicles, these data sources offer different perspectives of the environment.
- Issue: Struggles in detecting obstacles accurately when multi-modal data is inconsistent.
- Multi-modal fusion networks like FuseNet outperform single-modal networks but struggle when one modality (e.g., depth) is missing or degraded.
- Degradation often occurs due to environmental factors like shadows, glares, or limited depth sensing range, leading to poor segmentation performance.



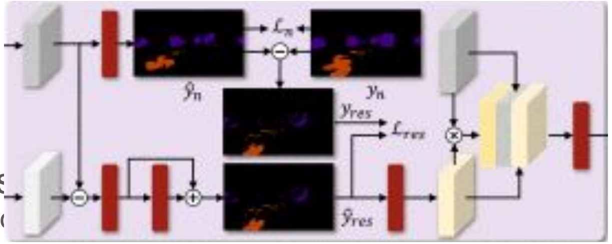
Proposed Solution

- To address the issue, create a network with two connected streams:
  - First stream: encoder-decoder for analyzing depth image.
  - Second stream: encoder-decoder for handling RGB image, with Residual-Guided Fusion (RGF) modules in the decoder.
- Combine the depth stream's decoder output with the RGF modules in the RGB stream.
- The RGB stream will provide the segmentation mask by effectively using both depth and RGB images.
- Purpose of RGF module: Quantify missing features between RGB features and ground truth, addressing performance degradation due to inconsistencies between data types.
- In contrast to Iconseg's 5-stage encoder-decoder, our model uses a more compact 3-stage encoder and 3-stage decoder.
- Unlike Iconseg, which fuses outputs from the last three decoder stages, our model fuses only the output from the final decoder stage into the RGF module.



The RGF Module

- The RGF module takes two inputs: RGB feature maps(dark grey) and depth feature maps(light grey).
- RGB feature maps produce an RGB predicted mask y_hat using a convolutional layer.
- A residual mask y_res is generated through element-wise subtraction between y_hat and the ground truth y, representing the missing features of the RGB feature map.
- Next, complementary features are extracted for the missing features
  - Element-wise subtraction is performed between RGB and depth feature maps t compute their difference.
  - The difference is adjusted to the number of classes using a 1×1 convolution.
  - A residual unit with a 3×3 convolution generates the predicted residual mask y_hat_res, guided by y_res.
  - The adjusted result is fused with the RGB feature maps via element-wise multiplication.
  - The adjusted result, fusion result, and RGB feature maps are concatenated along the channel dimension.



Encoder Decoder Structure

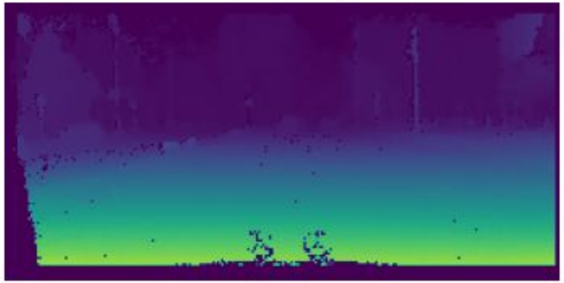
- Encoder
  - 2 Convolution layers 3x3 filters
  - Max Pooling 2x2
- Decoder
  - Upsampling layer
  - Concat with skip connection
  - 2 convolution layers 3x3 filters

Dataset

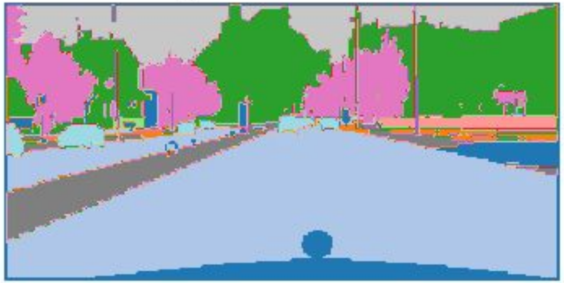
- The dataset used is a preprocessed version of the CityScapes dataset, sourced from the work "End-to-end Multi-task Learning with Attention."
- It includes three types of images:
  - RGB image: standard color image.
  - Depth image: encodes depth information for each pixel.
  - Label image: serves as the segmentation mask.
- The dataset contains 20 classes for semantic segmentation, with each image having a resolution of 128x256 pixels.
- The dataset is divided into three subsets:
  - Training set: 2,380 images.
  - Validation set: 500 images.
  - Test set: 595 images.
- The dataset has a significant class imbalance, with some classes being much rarer than others.

RGB Image

Depth Image



Segmentation Label

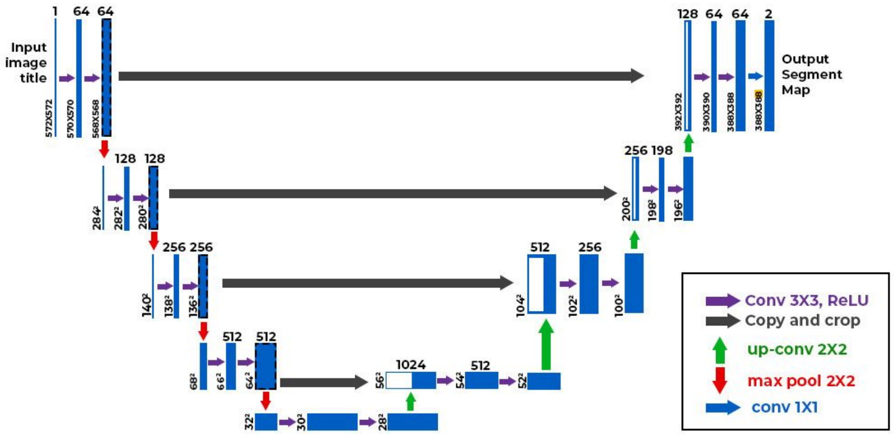


Imbalance

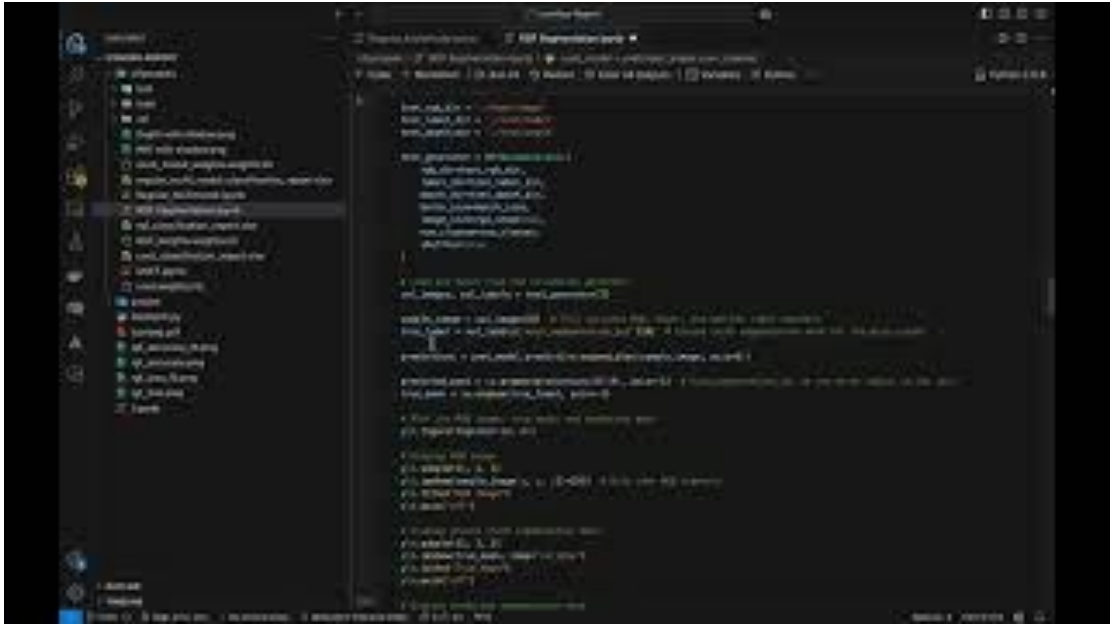
Class Number(original)	Class Number(after offset)	Dataset occurrence(%)
-1	0	12.24624249
0	1	32.44301035
1	2	5.306801676
2	3	20.04192705
5	6	1.079105409
7	8	0.4747958143
8	9	13.95592954
9	10	1.010260061
10	11	3.545020608
11	12	1.033653452
12	13	0.1201418067
13	14	6.146646708
18	19	0.3574544442
3	4	0.5598603577
4	5	0.7466766614
6	7	0.1800101144
17	18	0.08927545628
14	15	0.2416210014
15	16	0.2122253931
16	17	0.2093416102

Models

- We will analyze 3 different models
  - Regular UNET(as designed in the hw)
  - Multi Modal UNET with depth perception
  - Multi Modal UNET with RGF
- Parameters for each
  - Regular UNET: 7,760,724 params
  - Multi Modal UNET: 3,815,284 params
  - Multi Modal UNET with RGF: 3,858,112 params



Our Model Running



Results

Regular UNET(mean IoU= 0.2014)

Class	precision	recall	f1-score
0	0.91	0.59	0.72
1	0.82	0.95	0.88
2	0.44	0.42	0.43
3	0.7	0.77	0.73
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0.68	0.77	0.72
10	0	0	0
11	0.81	0.92	0.86
12	0	0	0
13	0	0	0
14	0.59	0.79	0.68
15	0	0	0
16	0	0	0
17	0	0	0
18	0	0	0
19	0	0	0
accuracy	0.74	0.74	0.74

Multi-modal UNET (mean IoU= 0. 0.228)

Class	precision	recall	f1-score
0	0.93	0.61	0.73
1	0.87	0.95	0.91
2	0.58	0.55	0.57
3	0.68	0.9	0.78
4	0	0	0
5	0.06	0	0
6	0.47	0	0
7	0	0	0
8	0	0	0
9	0.79	0.73	0.76
10	0.24	0.04	0.07
11	0.88	0.89	0.89
12	0.34	0.17	0.22
13	0	0	0
14	0.62	0.84	0.72
15	0	0	0
16	0	0	0
17	0	0	0
18	0	0	0
19	0	0	0
accuracy	0.77	0.77	0.77

Multi-modal UNET w RGF (mean IoU= 0.21)

Class	precision	recall	f1-score
0	0.917884	0.611202	0.733788
1	0.816738	0.962296	0.883562
2	0.500884	0.400119	0.444867
3	0.730892	0.801098	0.764387
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0.673666	0.838151	0.746961
10	0.199059	0.017439	0.032069
11	0.776602	0.958087	0.857851
12	0	0	0
13	0	0	0
14	0.72888	0.729658	0.729269
15	0	0	0
16	0	0	0
17	0	0	0
18	0	0	0
19	0	0	0
accuracy	0.759763	0.759763	0.759763

Results(cont.) IMG w/o shadow

RGB Image

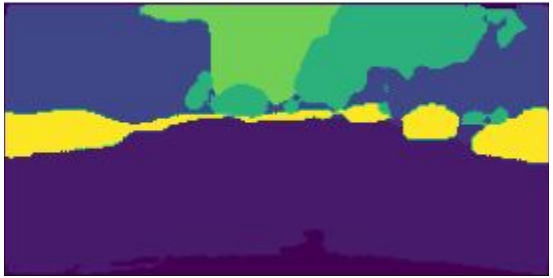
True Mask

Regular UNET

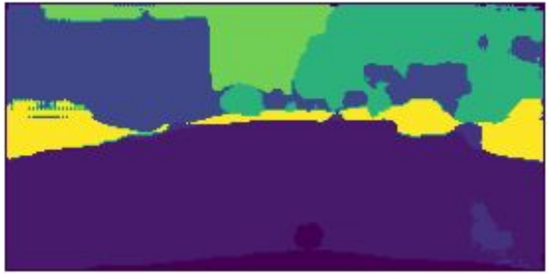
Multi-modal UNET

Multi-modal UNET w RGF

Predicted Mask



Predicted Mask

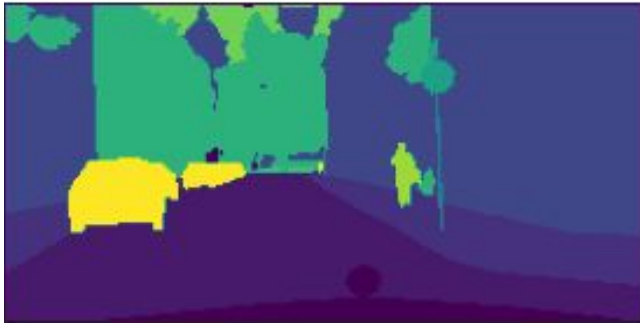


Predicted Mask

Results(cont.) IMG w shadow

RGB Image

True Mask

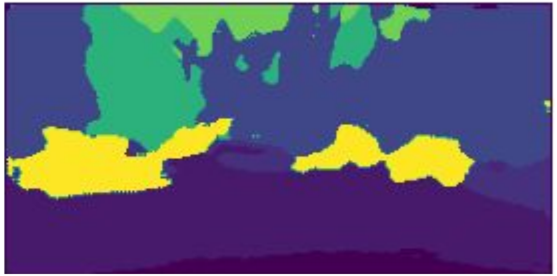


Regular UNET

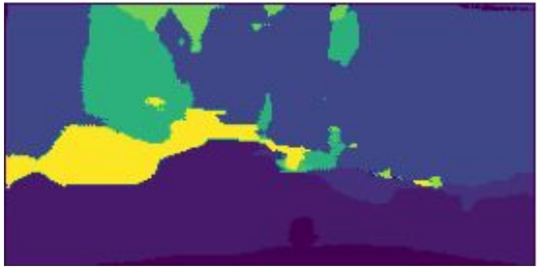
Multi-modal UNET

Multi-modal UNET w RGF

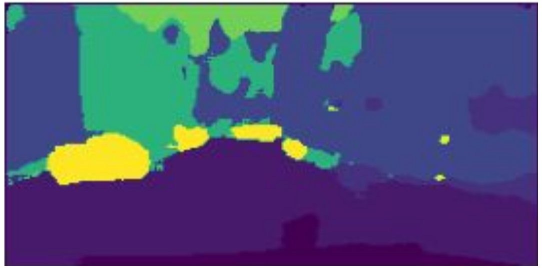
Predicted Mask



Predicted Mask



Predicted Mask



Conclusion

- RGF-Enhanced Network: Improves segmentation in shadowed regions compared to baseline UNet and standard multimodal models.
- Performance and Efficiency: Achieves higher accuracy in shadow detection without a substantial increase in model size.
- Further Improvements: Increasing network depth, adding more RGF modules, and training for more than 10 epochs could unlock additional potential.