UCLA College | Life Sciences Psychology

Final Project: Building a Computer Vision-Based Object Detection and Classification System in Urban Street Scenario Using U-Net CNN Model

Group members: Clark (Xuanyao) Qian Daphne (Yiyang) Zhang

Psych 186B Instructor: Zili Liu, PhD

3/12/2025

Introduction & Motivation

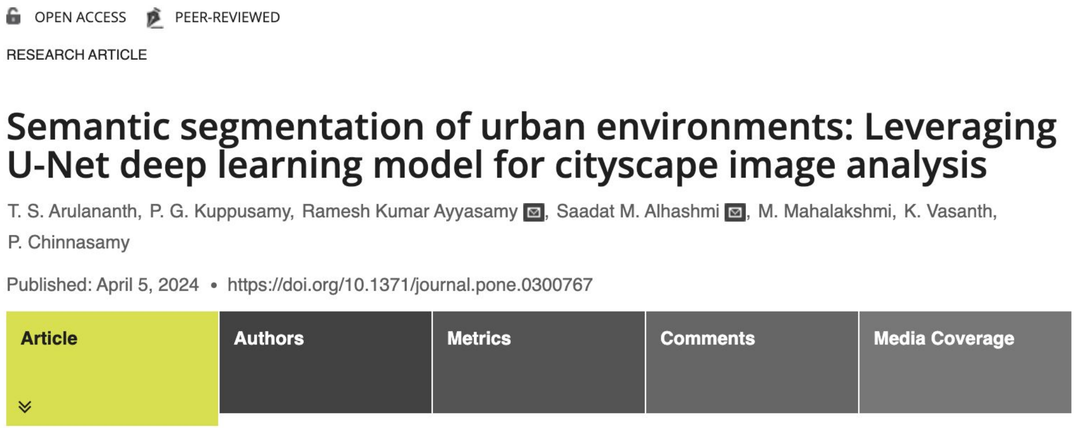
Interest: How are Autonomous Vehicles trained?

Deep Learning: Computer Vision Models (Object Detection & Semantic Segmentation)

Backgound & Literature Review

- Existing models:
- U-Net
- Strengths: Efficient, widely-used baseline.
- Other advanced models: DeepLabV3, Mask-RCNN

- Literature: Semantic segmentation of urban environments: Leveraging U-Net deep learning model for cityscape image analysis



OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Semantic segmentation of urban environments: Leveraging U-Net deep learning model for cityscape image analysis

T. S. Arulananth, P. G. Kuppusamy, Ramesh Kumar Ayyasamy, Saadat M. Alhashmi, M. Mahalakshmi, K. Vasanth, P. Chinnasamy

Published: April 5, 2024 • https://doi.org/10.1371/journal.pone.0300767

Article | Authors | Metrics | Comments | Media Coverage

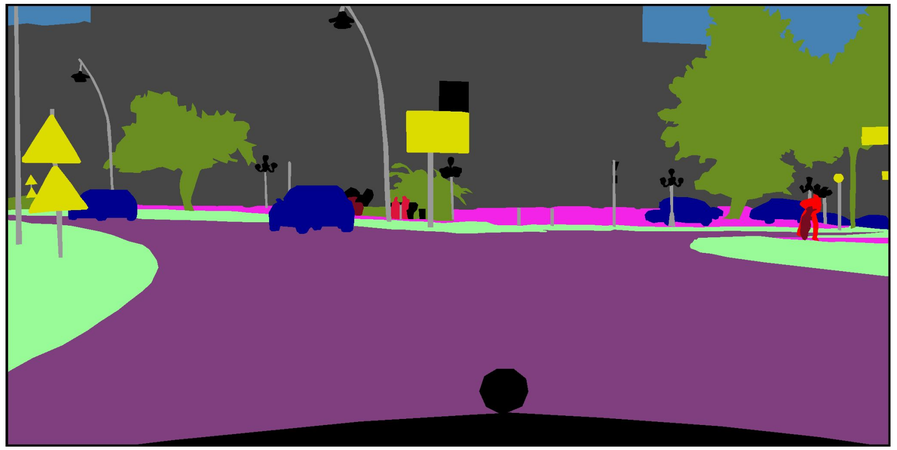
Hypothesis

- Can a basic U-Net architecture provide accurate semantic segmentation?

- Enhancements (e.g., data augmentation, weight adjustment, resolution adjustments) will improve object segmentation accuracy.

Dataset:

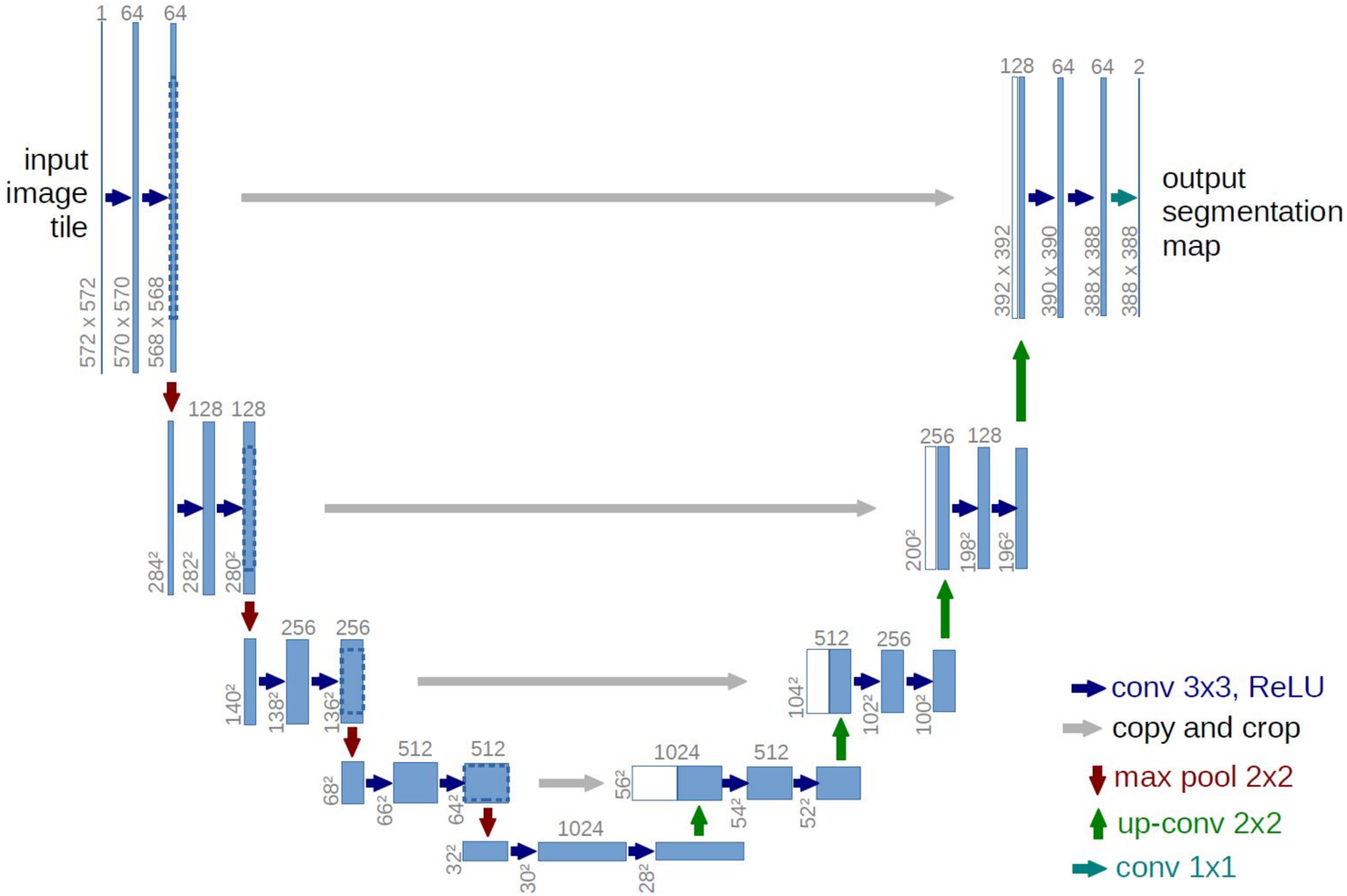
- Dataset: CityScapes images (urban scenarios)
- Diversity:
  - 5000 fine annotated images (1024 * 2048)
  - 50 cities in Germany
  - Summer, Spring, Fall
  - Daytime
  - Good -> Medium (Flare) Weather Condition



- Features:
  - 8 Groups (flat, human, vehicle, construction, object…)
  - 30 Classes (car, person, rider, vegetation, road…)

Model:

- **U-Net Architecture**
- Initially designed for Biomedical Image Segmentation
- Good performance using small-scale dataset
- Encoder-Decoder structure



U-Net Explained

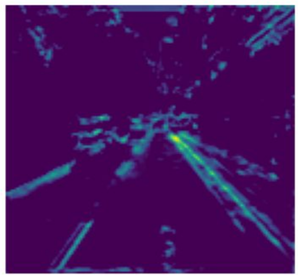
Encoder

Input
1024*1024
x1

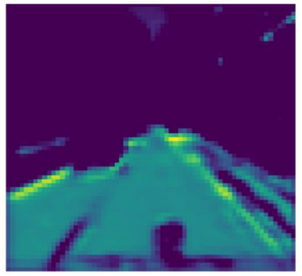
Decoder

stage1
1024 * 1024
x16

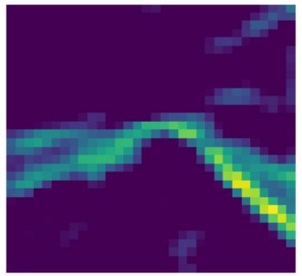
stage2
512 * 512
x32



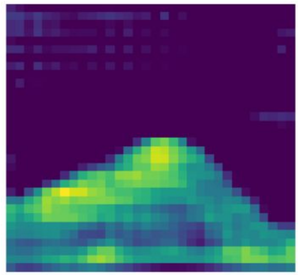
stage3
256 * 256
x64

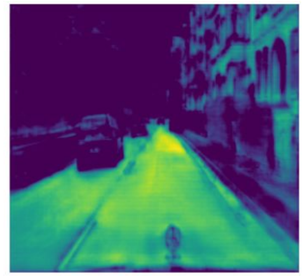


stage4
128 * 128
x128

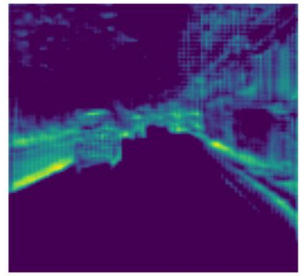


stage1
128 * 128
x128

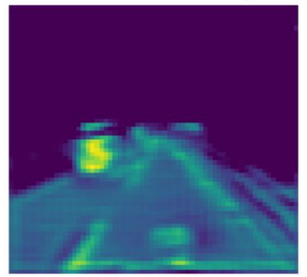




stage4
1024 * 1024
x16



stage3
512 * 512
x32



stage2
256 * 256
x64

Input: An RGB picture

Output: A segmentation map of the same shape that assigns each pixel to one of the classes (road, building, vegetation, sky, etc.).

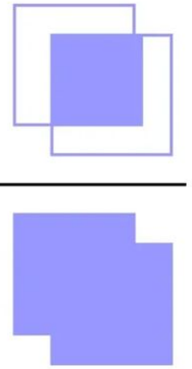
Original Image

Final Prediction

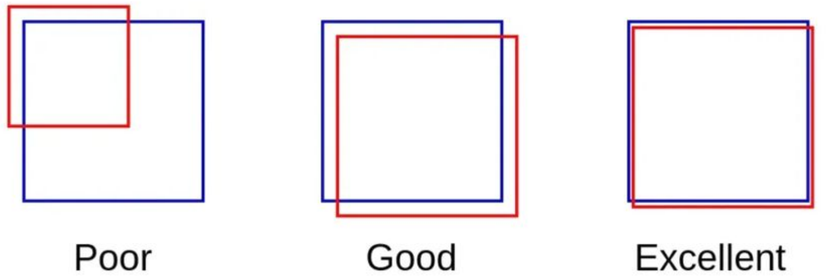
Ground Truth Label

Evaluation criteria

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



Intersection over Union (IoU)



- **Initial Baseline Accuracy**:
  - Mean IoU > 0.5 as the acceptable threshold.
  - Mean IoU < 0.5 indicates poor segmentation (especially people).

Initial Settings:

- **Data Selection:**
  - Randomly selected 50% of the original data (2500)
  - 80:20 Train/Validation Split (2000 Train, 500 Validation)
  - Additional 500 data for testing

- **Selected Classes:**
  - Person, Car, Building, Sky, Vegetation, Road

- **Hyperparameters:**
  - Epoches: 20
  - Activation function: softmax
  - Optimizer: Adam
  - Learning rate: 0.0001
  - Batch size: 8
  - Input Resolution: 128 * 128

Table 1. The proposed model parameter setup.

Hyperparameter	Configurations
Activation function	"softmax"
Optimizer	Adam
Learning rate	0.0001
Batch size	32
Epochs	20–25
Metrics	IoU & mean IoU
Input images size	128 × 128

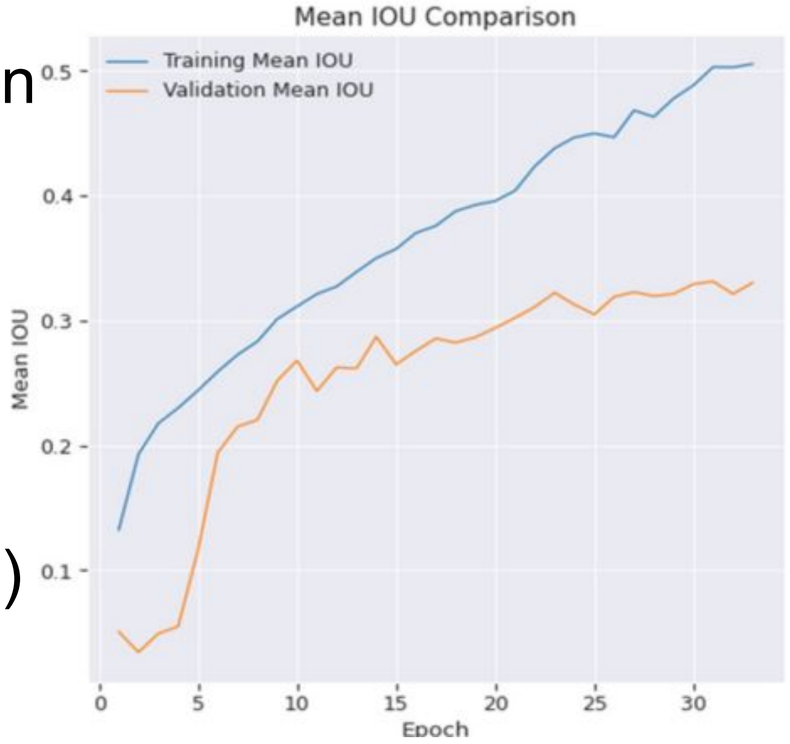
Results:

Higher training and validation IoU compared to the first 20 epochs in the paper.

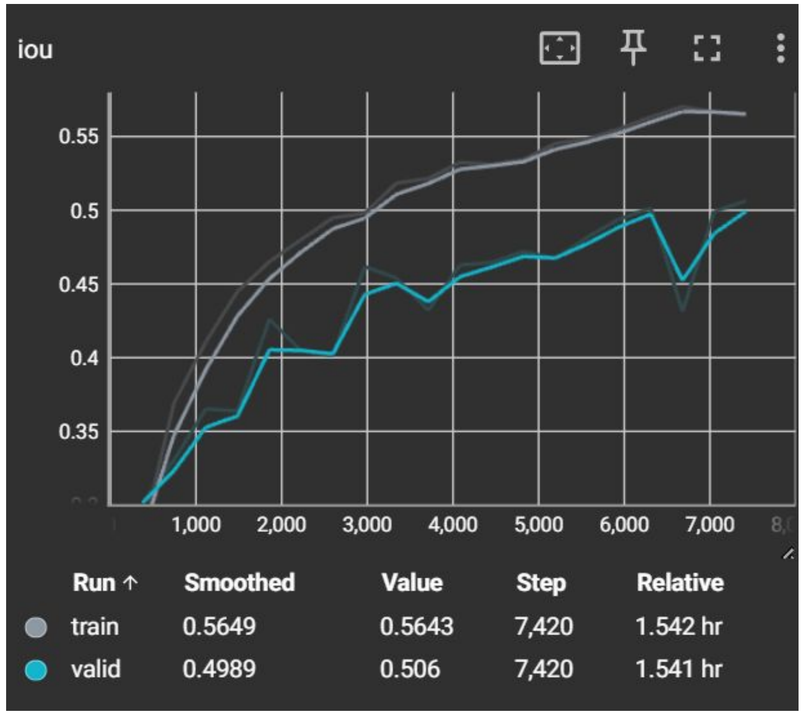
Potential Reason:

- Different Batch Size (8 vs 32)
- Different Training Data
- Different Testing Data

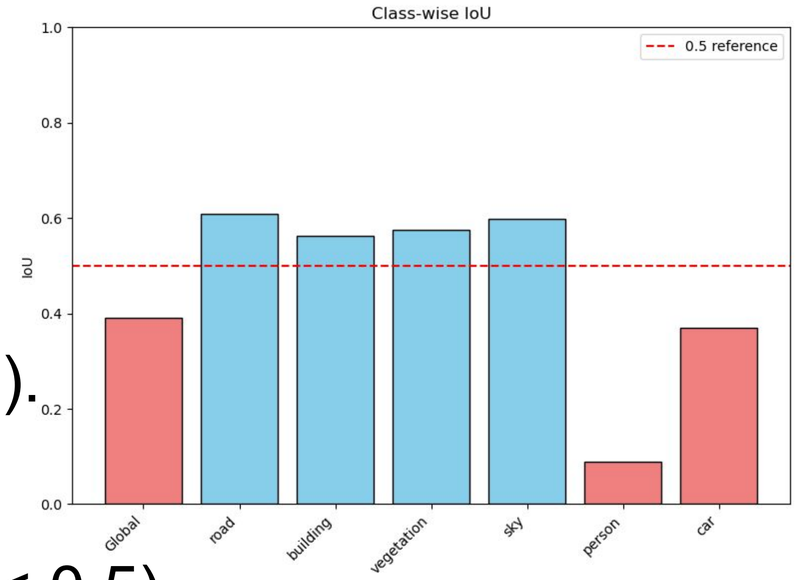
Paper Performance



Project Performance



Further investigation: Class Wise Accuracy



- **Successful Results**:
- Good performance on simple scenarios

  (clear, sky, roads, background, few vehicle/human).
- **Challenges**:
- People and vehicles consistently challenging (IoU < 0.5).
- **Insights**:
- Complexity of scenario significantly affects performance.

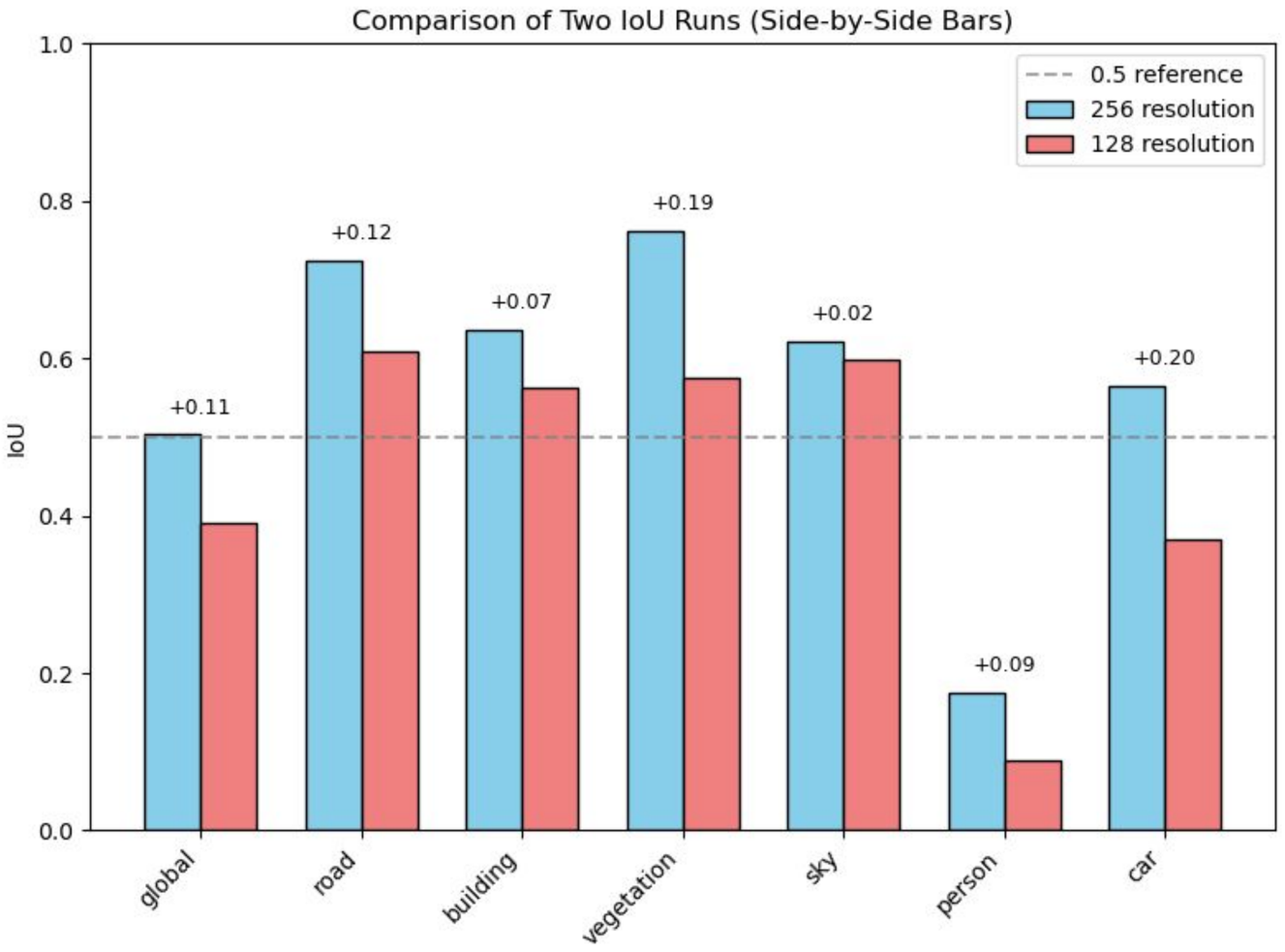
Model improvements

- **Resolution Adjustment**: 128x128 to 256x256 significantly improves overall accuracy (paired t-test $p<0.05$).

- **Limitation**:Computationally Expensive

  100+ minute training time

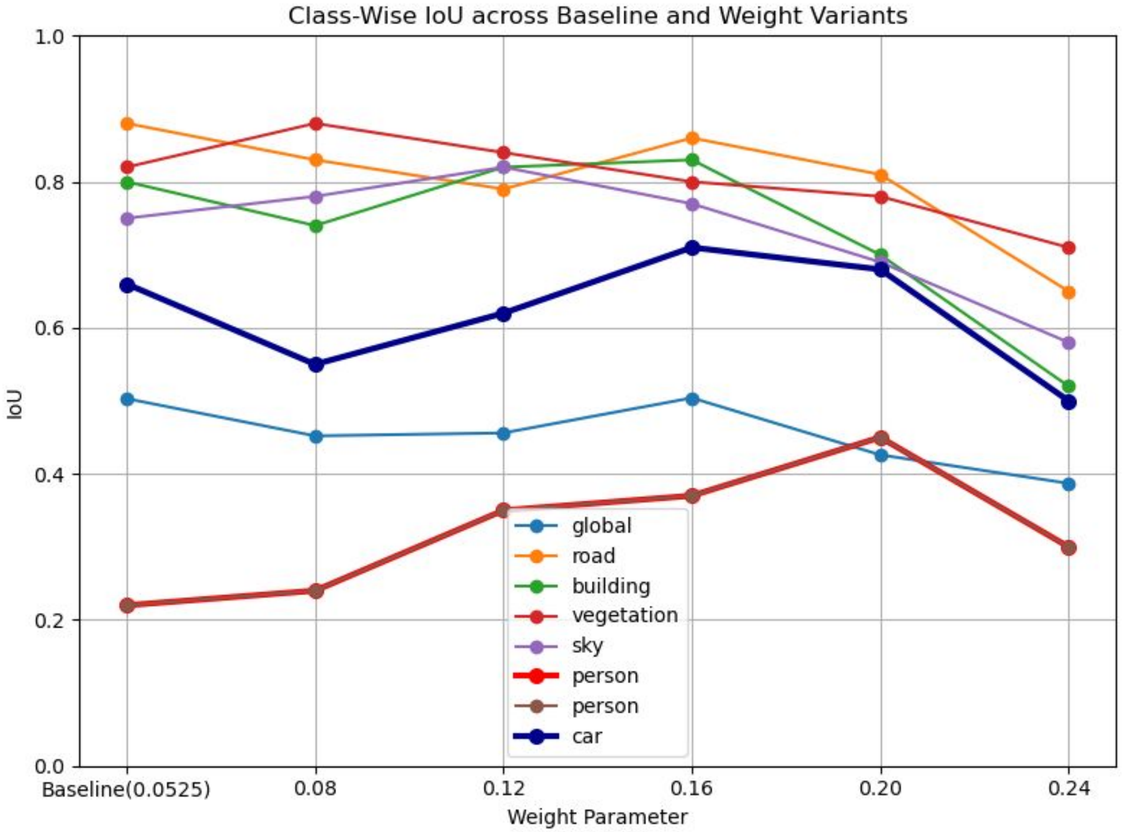
  GPU ran out of memory when training higher resolutions



Model improvements

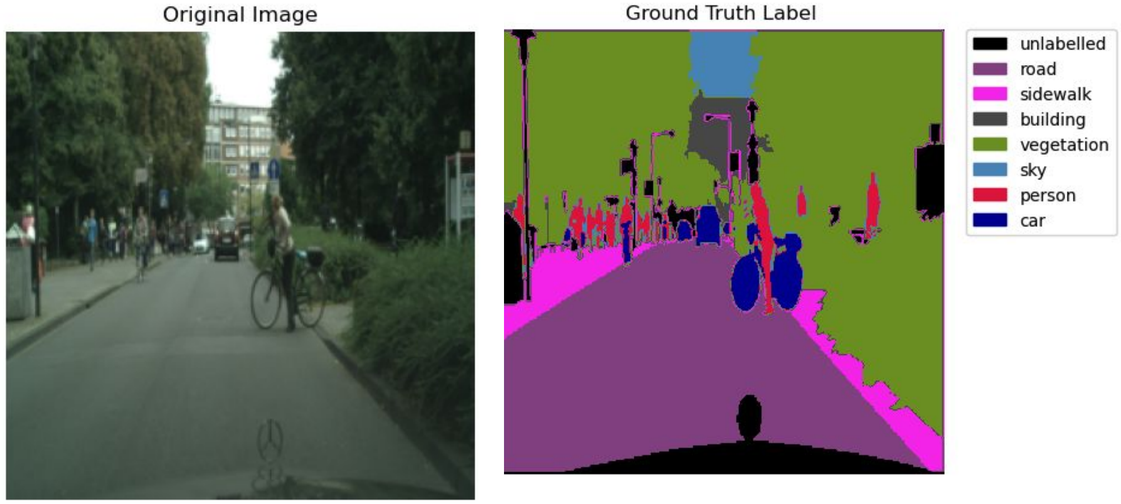
- **Weight Adjustment**: increasing weight for the person/car class gradually increases their accuracy with the cost of slight decrease in accuracy of other class. Model deteriorates when weight > 0.2.

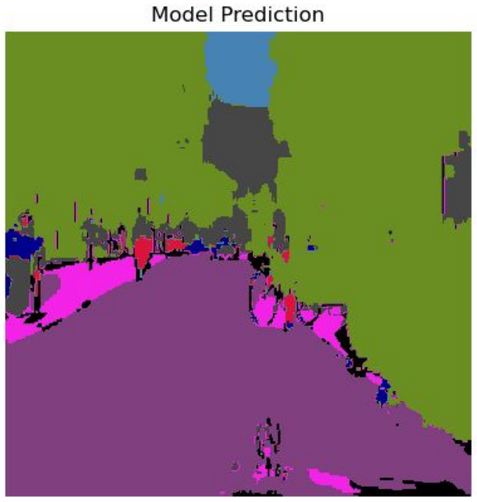
```
### Weights for Focal loss
FOCAL_LOSS_WEIGHTS = [
    0.1825,    # road
    0.0525,    # building
    0.025,     # vegetation
    0.01,      # sky
    0.0525,    # person
    0.0525     # car
]
```



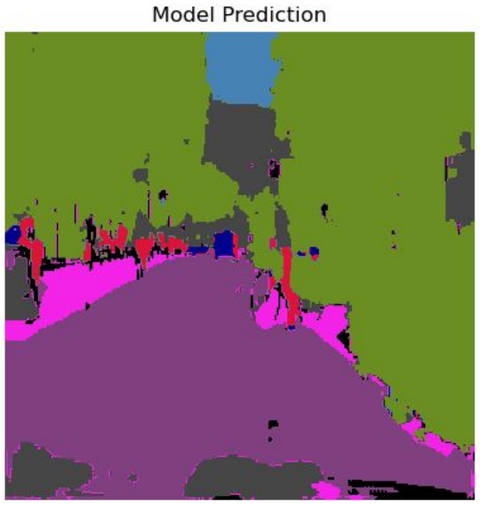
Model improvements

Weight Adjustment - case study

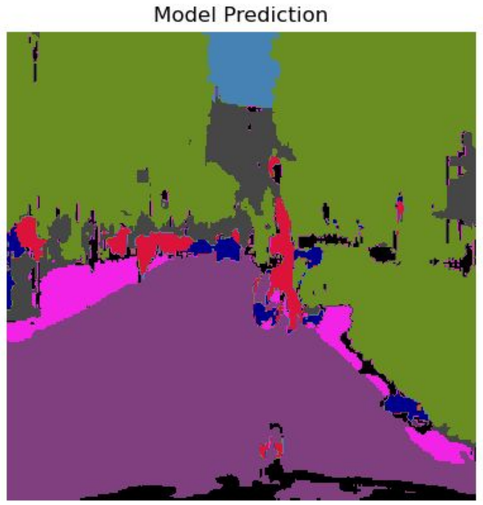




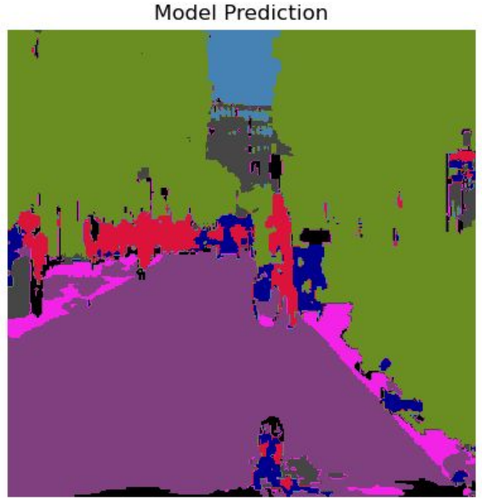
Weight=0.08



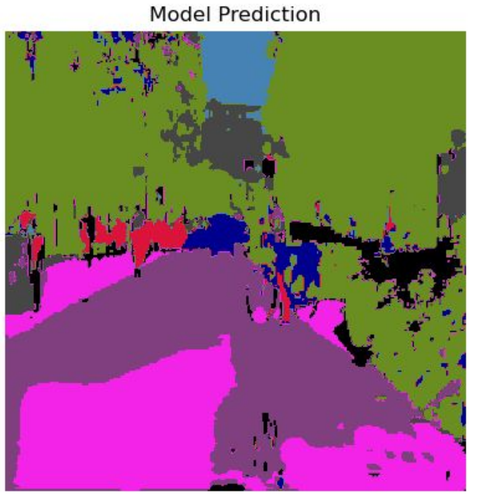
Weight=0.12



Weight=0.16



Weight=0.20



Weight=0.24

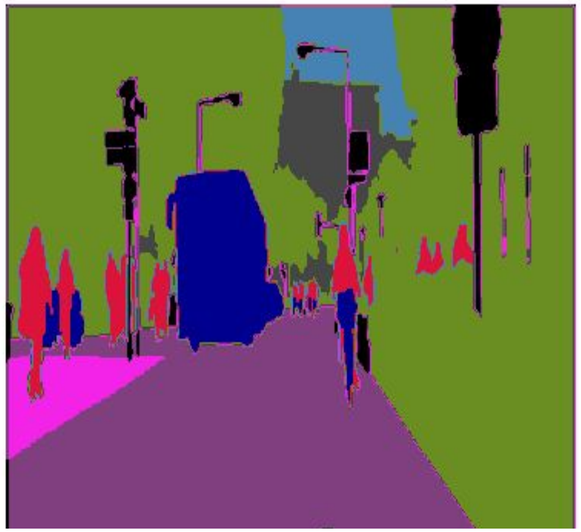
Beyond 0.2 threshold, model starts to over-generalize classifications and make significant prediction errors

Model improvements

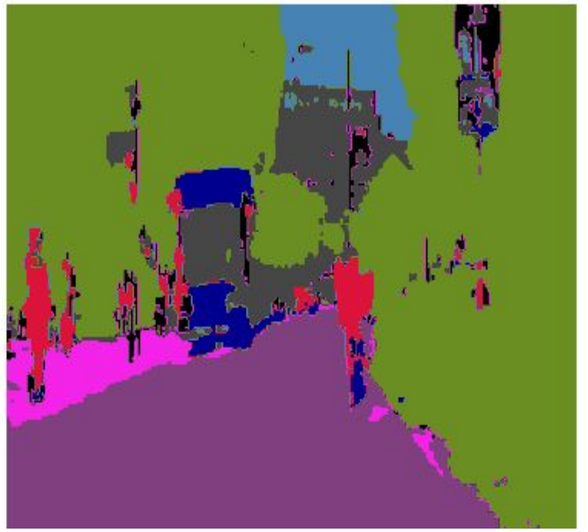
- Incomplete prediction on vehicles (especially false prediction on window reflections)

Original Image

Ground Truth Label



Model Prediction



Original (256×256)

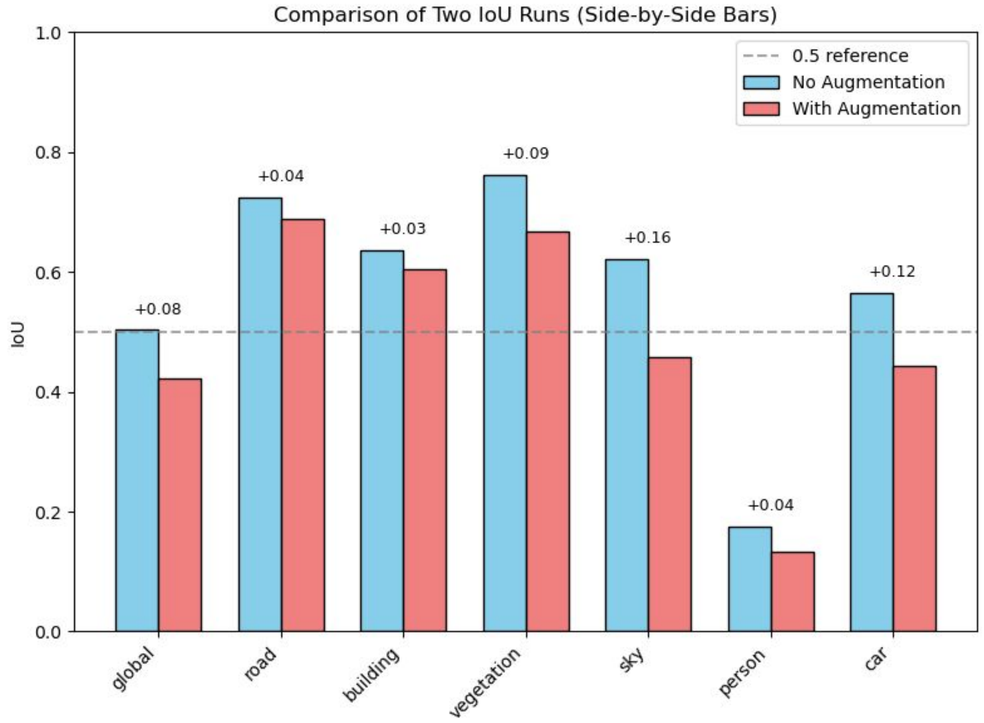
Color Jitter

Color Jitter: Randomly change saturation, contrast, brightness of the image

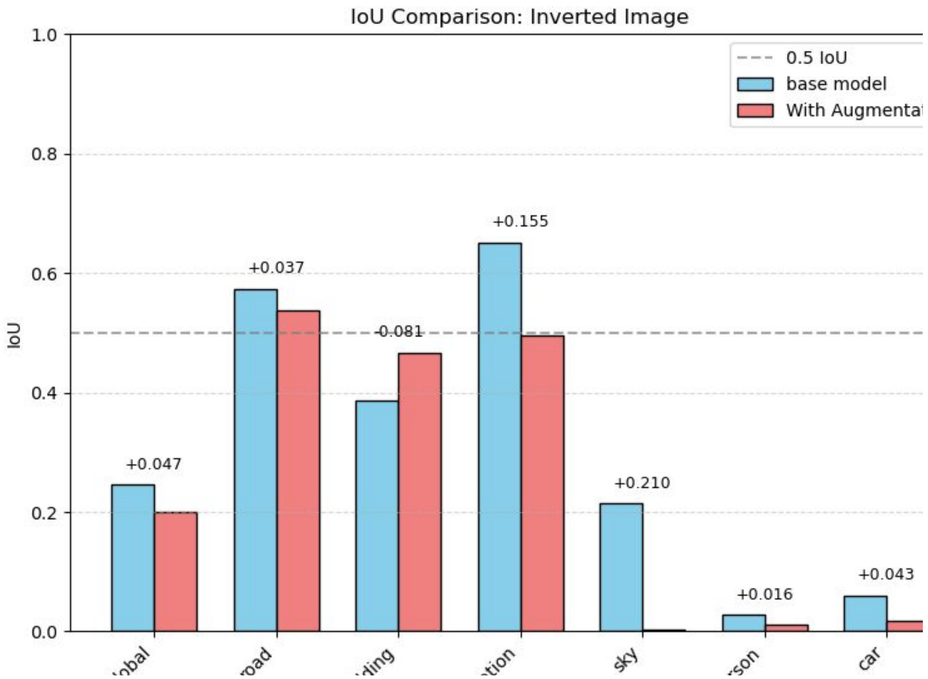
Purpose: Robustness when dealing with different color/texture on cars

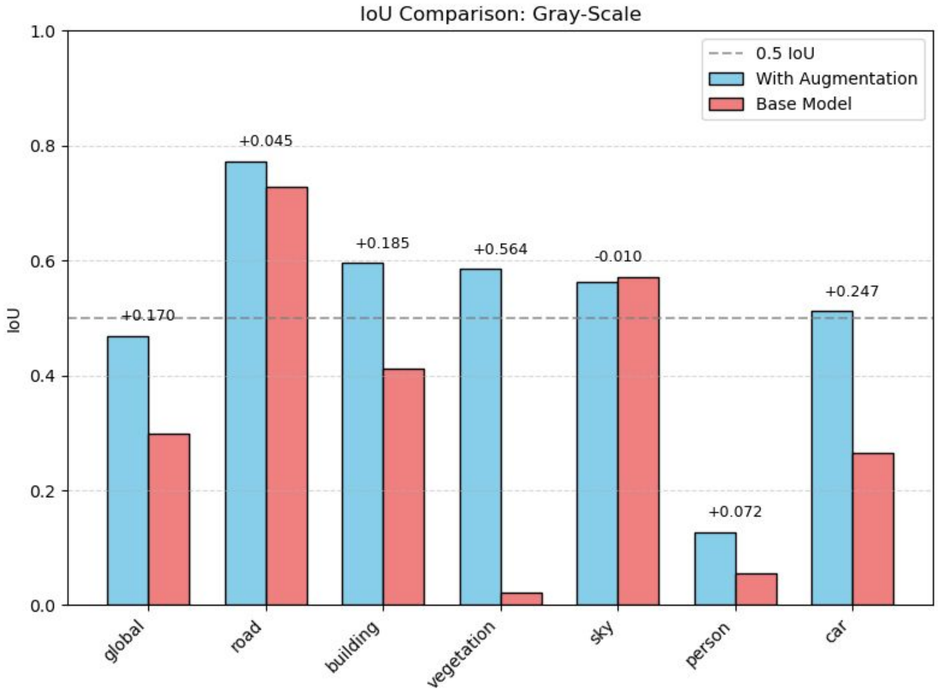
Model improvements

- **Data Augmentation**: to improve general robustness
- Horizontal Flip, Random Crop
- Result: No significant improvement on car or overall prediction



Generalization to Peculiar Scenarios





Limitations

Model deficiency:

Small Object Detection

3D Urban Morphology Interpretation

Model Interpretability

Project deficiency:

Computation power

Limited data-Reliability Across Atmospheric Conditions

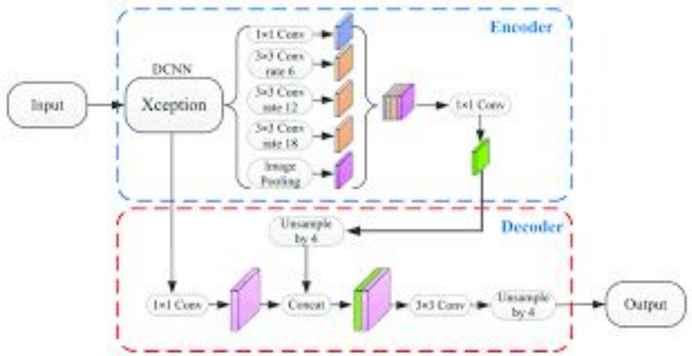
Pixel-wise compared to regional-wise evaluation

Conclusion

Summary of Results:

- U-Net effective baseline; limited accuracy in complex scenarios.

Improvements:



Choose diverse datasets

Improve computation power

Incorporate regional wise metric

Explore more advanced architectures (DeepLab V3+, PointNet++).

Implications:

- Segmentation accuracy directly impacts real-world decision-making.

Thank you