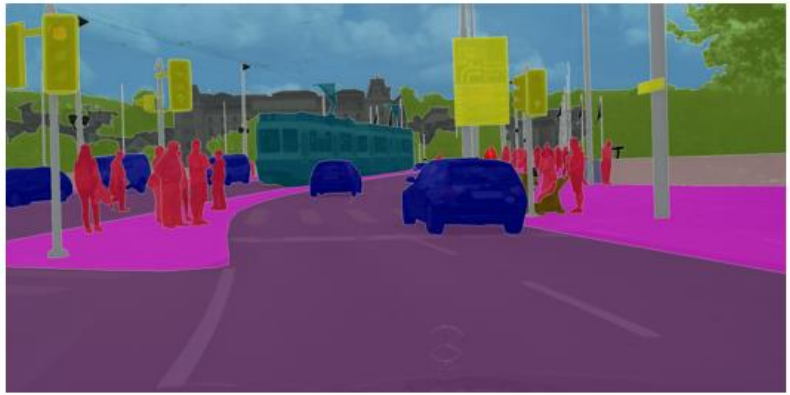
Hierarchical Multi-Scale Attention for Semantic Segmentation

(Tao, Sapra, Catanzaro - May 2020)

Arash Hatefi
Amin Fadaeinedjad
Alireza Forouzandeh Nezhad

What is Semantic Segmentation?

- **The task of semantic segmentation is to label all pixels within an image as belonging to one of N classes.**



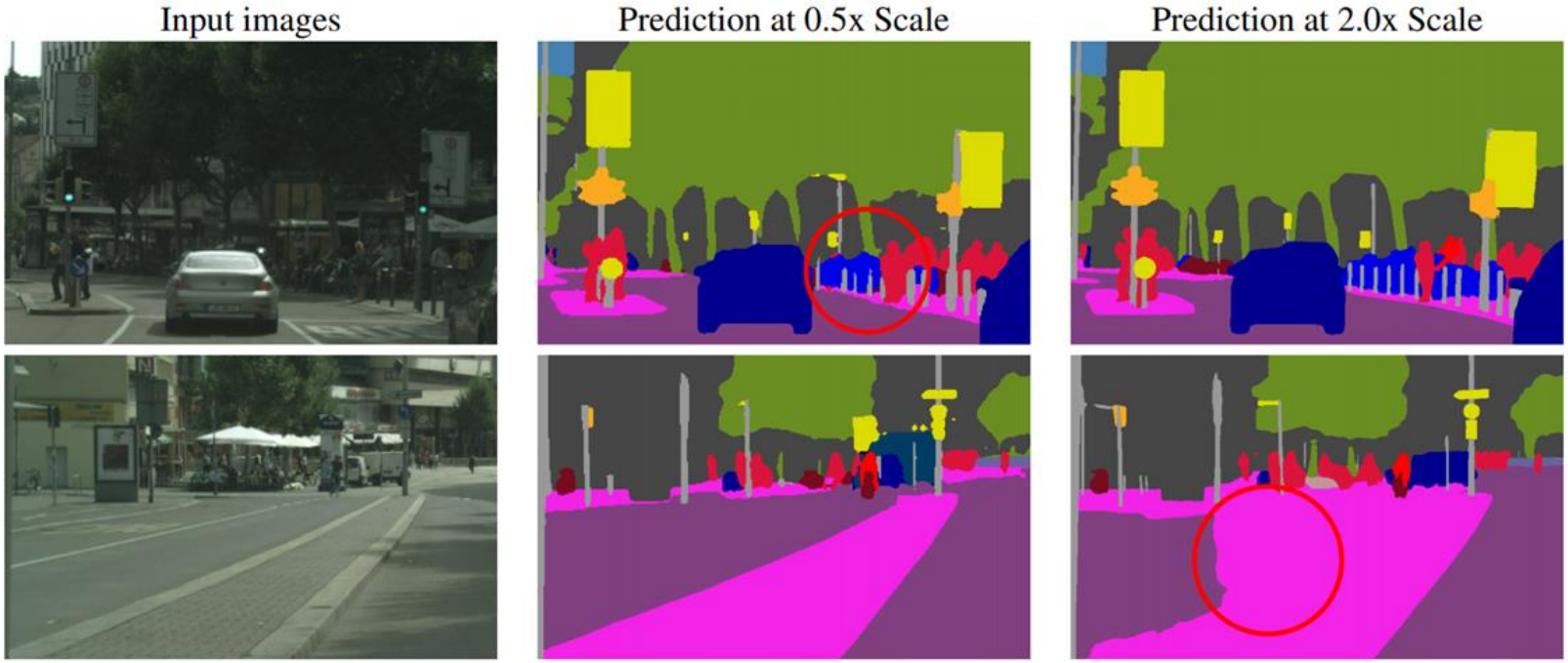
divamgupta.com

Resolution Trade Off in Semantic Segmentation

Certain types of predictions are best in different image resolutions:

1. Fine detail, such as the edges of objects or thin structures, is often better predicted in **scaled up** images.
2. Prediction of large structures, is often done better at **scaled down** image sizes.

Resolution Trade Off in Semantic Segmentation

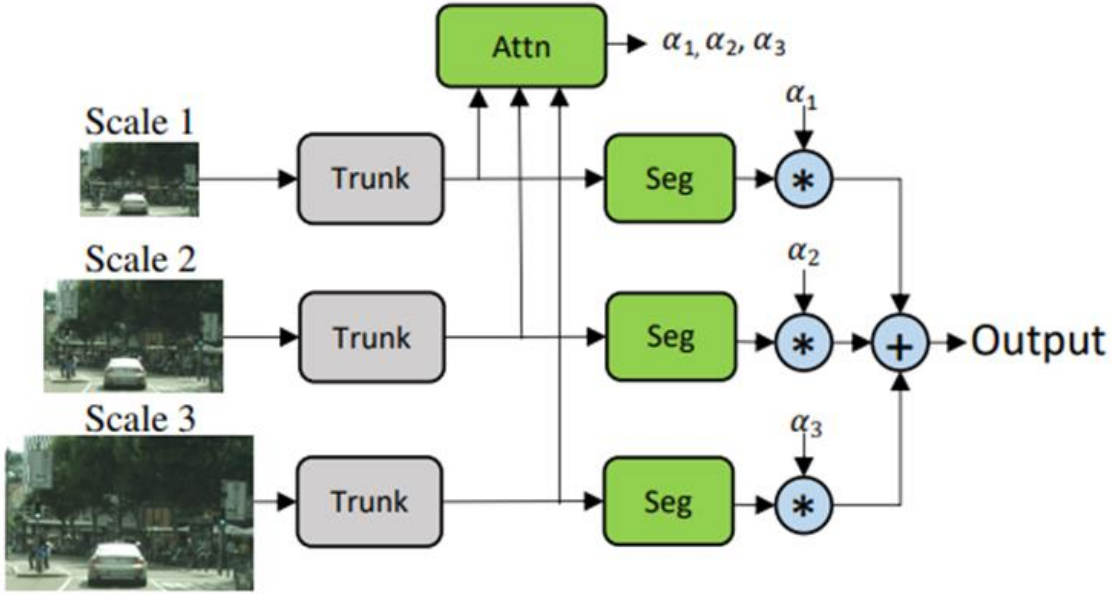


Addressing the Resolution Trade Off

Using multi-scale inference is a common practice to address this trade off. Predictions are done at a range of scales, and the results are finally combined using:

1. **Averaging methods:** suffers the problem of combining the best predictions with poorer ones
2. **Attention-based architectures**

Conventional Attention-Based Architecture



Trunk: Generating the feature map of the image

Segmentation: Classifying each pixels of the input image based on the feature map

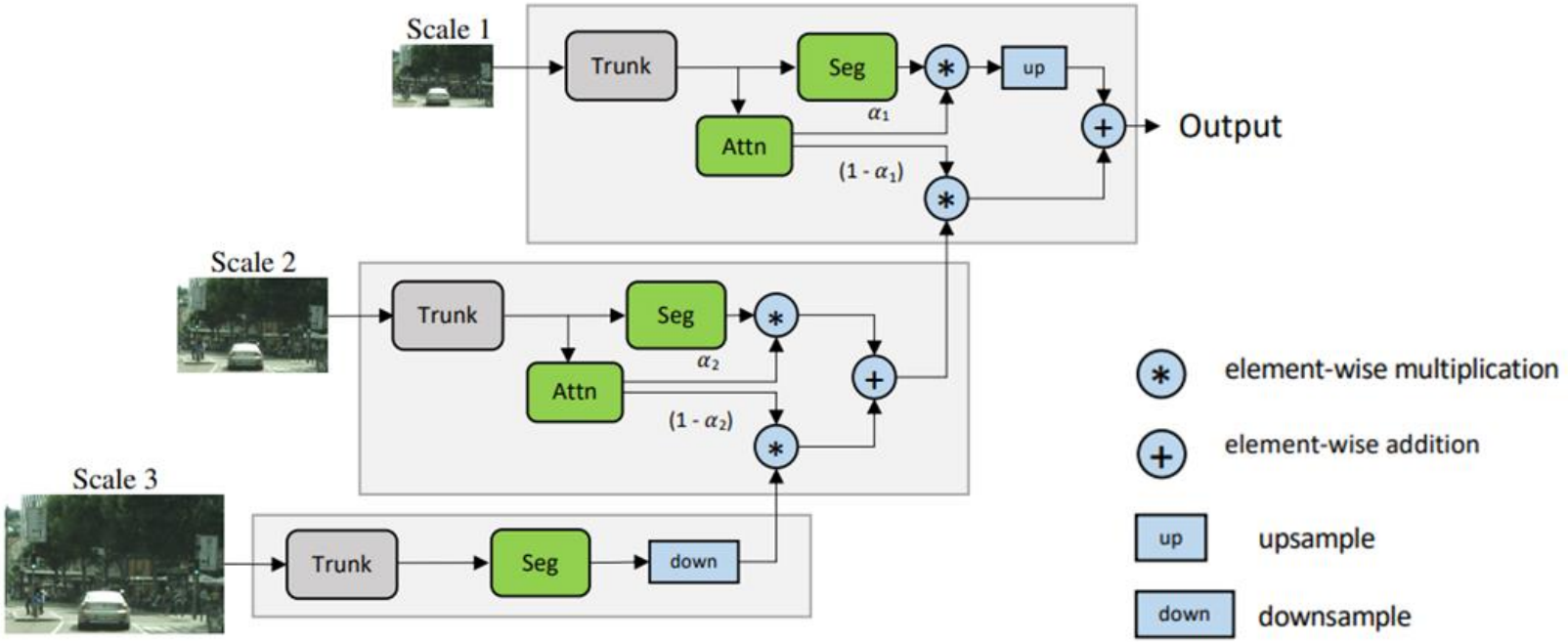
Attention: Generating attention masks

element-wise multiplication

element-wise addition

Encoder-decoder with atrous separable convolution for semantic image segmentation.
Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam.

Conventional Attention-Based Architecture



Hierarchical Multi-scale Attention Architecture for Semantic Segmentation

The Architecture

1. **Trunk:**
    a. ResNet-50
    b. HRNet-OCR * (Used for state-of-art results)
2. **Segmentation:**
    a. DeepLabV3+ **
3. **Attention:**
    a. Based on transpose convolutions

* Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation, 2019

** Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018.

Contributions

1. Achieving state-of-the-art results in Cityscapes (85.1 IOU) and Mapillary Vistas (61.1 IOU)

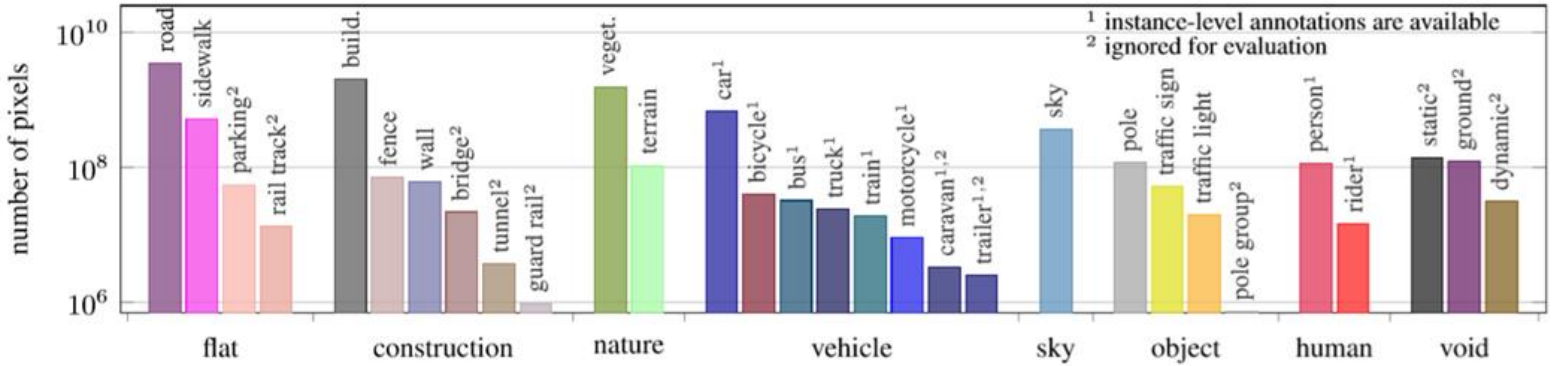
2. Speeding up the training process.

Cityscape Dataset

A large-scale dataset that contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities, with high quality pixel-level annotations of 5 000 frames in addition to a larger set of 20 000 weakly annotated frames.

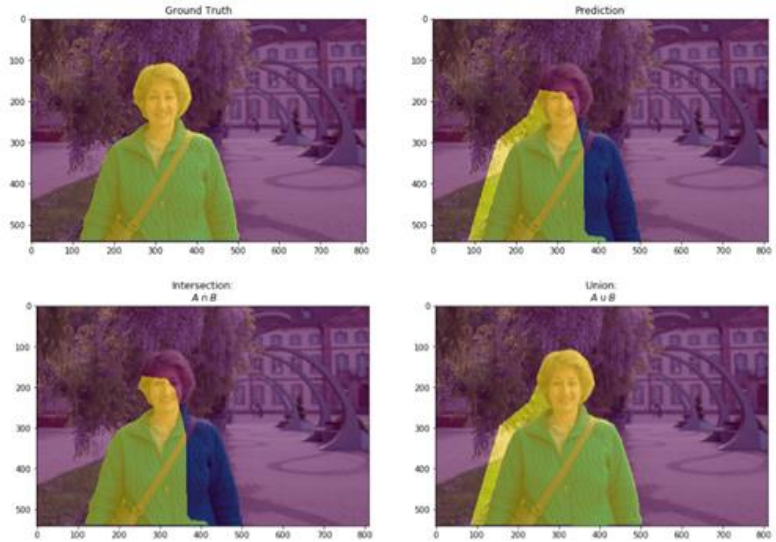
Cityscape Dataset

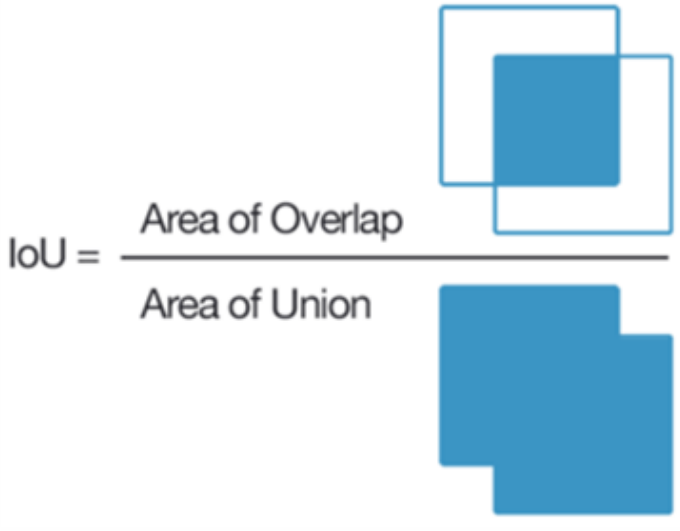
Like Andrew Tao et al. we also only include the 19 most popular labeled objects in our segmentation task while living the other objects as unknown.



Metric: Mean IOU (Intersection Over Union)

The average of the IOUs for of each present classes in an image:





Data Augmentation

We different methods to augment our train data such as:

1. **Random crop**
2. **Changing contrast**
3. **Adding Gaussian noise**
4. **Horizontal flipping**

Loss Function - Cross Entropy

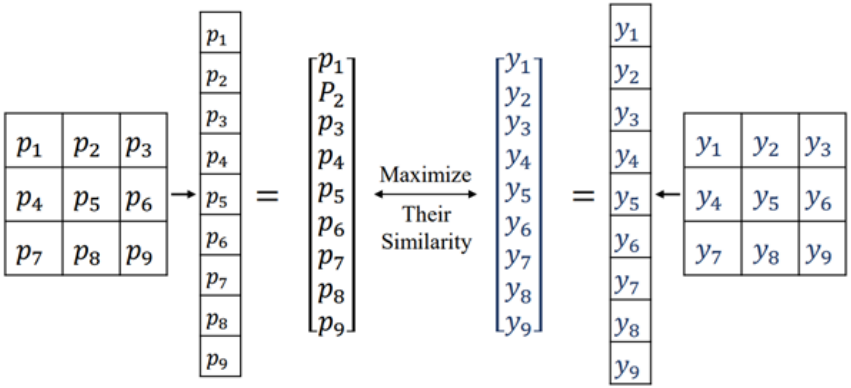
1. **Cross-entropy as the auxiliary loss function**

   Ignores the relationship between pixels and treats them as independent samples.

$$\mathcal{L}_{ce}(y, p) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(p_{n,c})$$

Loss Function - RMI

2. RMI (Region Mutual Information) as the primary loss function



$$\mathcal{L}_{all}(y,p) = \lambda \mathcal{L}_{ce}(y,p) + (1-\lambda)\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C}\left(-I_l^{b,c}(\boldsymbol{Y};\boldsymbol{P})\right)$$

Training - the Challenge

Training hardware:

- Authors implementation: 2 * DGX A100 nodes => **640GB VRAM, 312 TFLOPs**
- Our implementation: 1 * Pascal P100 => **16GB VRAM, 10 TFLOPs**

Training the whole model in an end-to-end manner, was impossible...

Training - the Trick

Use simpler architectures for trunk

1. **MobileNet V2** ~ 3.4 million parameters
2. **ResNet50** ~ 23 million trainable parameters

Training - the Trick

Train a part, freeze others!

1. Train Trunks, Segs in a single stage network for each scale, no attention.
2. Train hierarchical network with trained trunks, freeze trunks, train attentions.
3. Freeze attention layers, fine-tune trunks and segs.

Training – Initial Stage Results

Step-1, Initial stage training – no attention:

index	Trunk	Segmentation Block	Image Scale	#Epochs	Mean IOU On Validation Set	Average Accuracy On Validation Set	Training Time per Epoch
1	MobileNet	DeepLab V3+	0.5	7	51.4%	71.3%	45 min
2	MobileNet	DeepLab V3+	1	7	50.6%	68.7%	60 min
3	MobileNet	DeepLab V3+	2	5	50.3%	68.2%	60 min
4	ResNet	DeepLab V3+	0.5	6	50.1%	68.4%	1.5 h
5	ResNet	DeepLab V3+	1	6	48.0%	76.3%	1.5 h
6	ResNet	DeepLab V3+	2	5	49.0%	68.1%	1.5 h

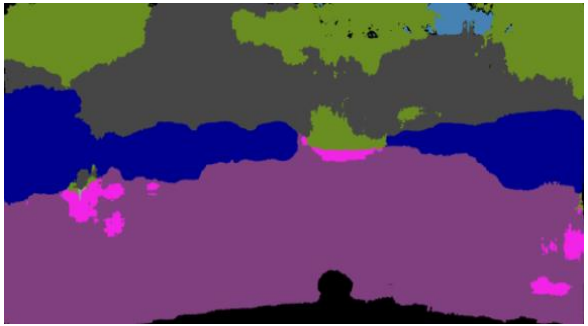
Training - Initial Stage Outputs 1 (ResNet Trunk)

Input

Ground Truth

Scale 2x

Scale 1x

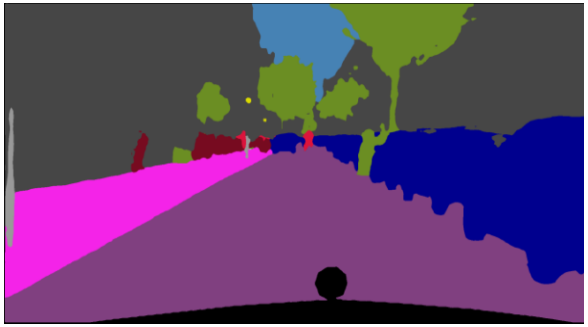


Scale 0.5x

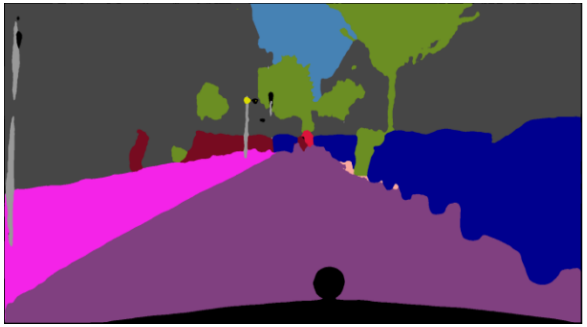
Training - Initial Stage Outputs 2 (MobileNet Trunk)

Input

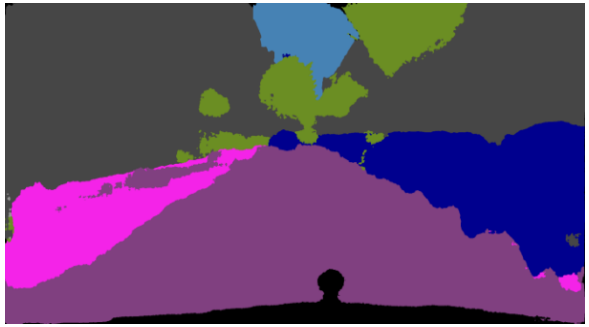
Ground Truth



Scale 2x



Scale 1x



Scale 0.5x

Training - Hierarchical Model Results

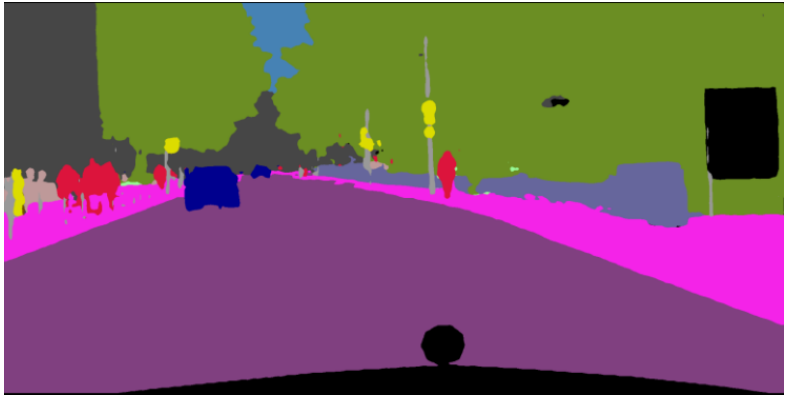
#Stages	Image Scales	#Epochs	Mean IOU On Validation Set	Average Accuracy On Validation Set
2	1/0.5	10	53.5%	75.2%
3	2/1/0.5	10	51.8%	74.6%

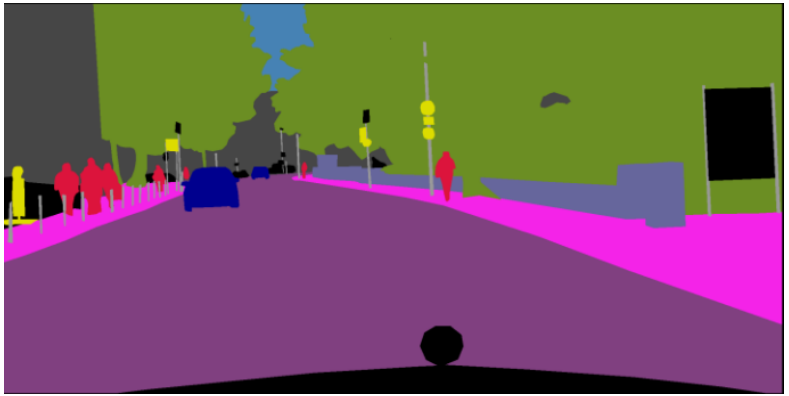
Hyper - Parameters

- **Learning Rates:**
  a. Step-1: 1e-3, 0.8 decay
  b. Step-2: 5e-4
  c. Step-3: 1e-4
- **Loss:**
  Cross-Entropy (faster, same result)
- **Trunk Network:**
  MobileNet v2 (best IOU, fastest)
- **Seg Network:**
  DeepLab v3
- **Train method:**
  Semi-supervised

Full 2-Stages Network Outputs:

Predicted



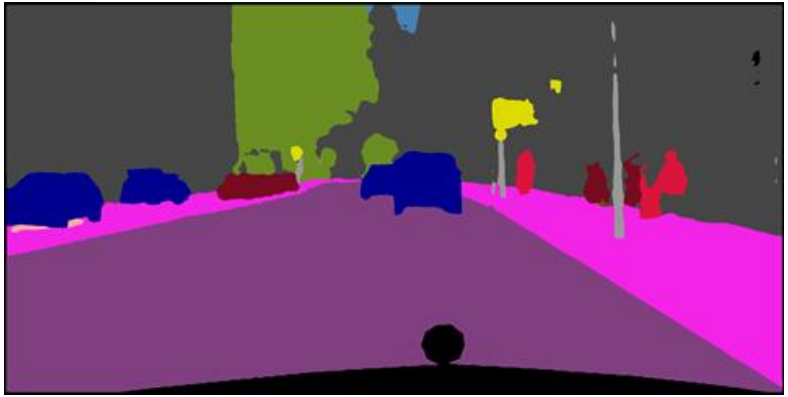


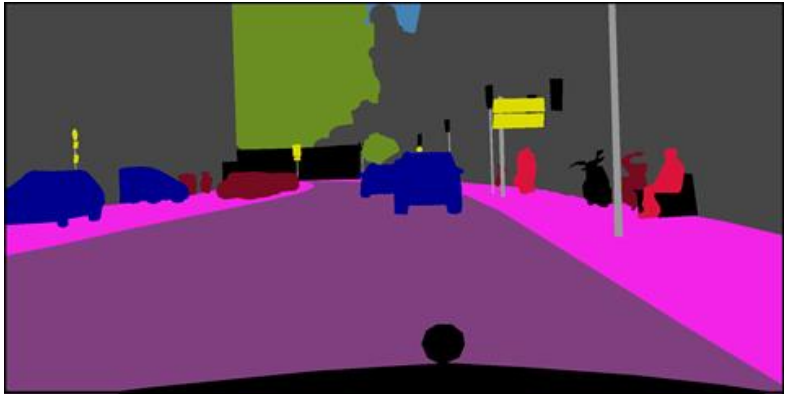
Ground Truth

Input

Full 2-Stages Network Outputs:

Predicted





Ground Truth

Input

Full 3-Stages Network Outputs:

Predicted

Ground Truth

Input

Full 3-Stages Network Outputs:

Predicted

Ground Truth

Input

Thanks !

- All references in this presentation has been cited in the attached report.