Domain Adaptation for Image Segmentation

Deepak Thukral
Anshul Khantwal
Sharat Agarwal

Introduction and Motivation

- Annotating real data for image segmentation is laborious and time consuming task.
- We aim to adapt the representation learned on synthetic data to real world data.

Learning to Adapt Structured Output Space for Semantic Segmentation

CVPR 2018

Related Work

- Long et al. proposed CNN models can be converted to fully-convolutional network for semantic segmentation.
    - Difficult to obtain annotations
    - May not generalize well to unseen image domains.
- Hoffman et al. introduced Domain adaptation by applying adversarial learning in a fully-convolutional way on feature representations.
- CyCADA transfers source domain images to the target domain with pixel alignment.
    - Generates extra training data along with feature space adversarial learning.

Datasets

Source:

- **GTA-5:** It is curated from the frames of a popular game, ***Grand Theft Auto V*** and consists of 24,966 densely labelled frames.
- **Synthia:** 9400 frames with semantic annotation compatible with Cityscapes.

Target:

- **Cityscapes:** Urban street images of 50 cities with 5000 images.
- ***Indian Roads:*** Videos obtained from youtube.

Proposed Model

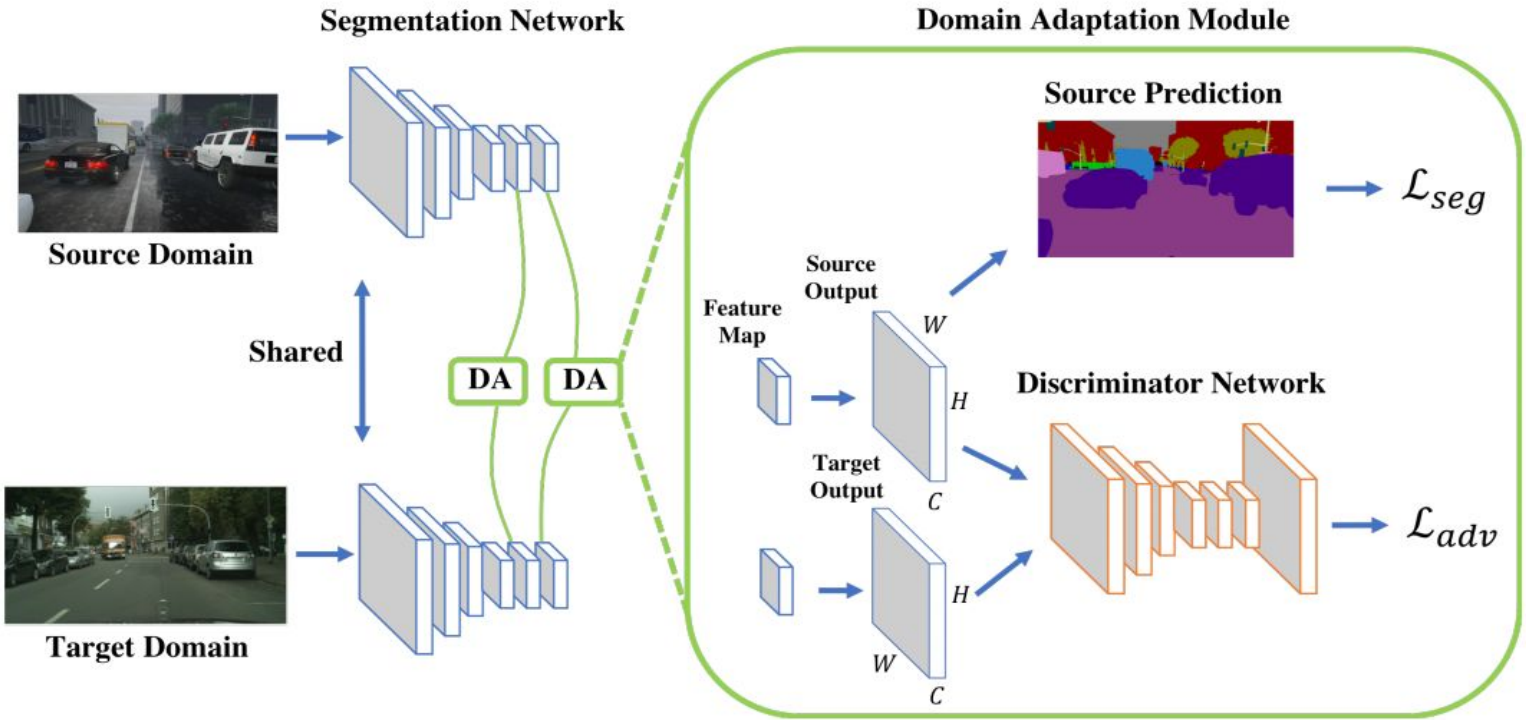


Image Source : https://goo.gl/EcBAq3

Proposed Model

- **Generator G (Segmentation Network)**:
  - Source image is forwarded through the segmentation network to predict the segmentation softmax output.
  - Adversarial loss on the target prediction makes the G generate similar segmentation distribution in the target domain to the source prediction.

- **Discriminator D:**
  - Discriminator is trained to distinguish between the source and target domain.

Objective Function

- **Segmentation Loss:** Cross-entropy loss using the ground truth annotations in the source domain.
- **Adversarial Loss:** Helps target predictions to adapt to the distribution of the source predictions.
- **Discriminator Loss:** Cross-entropy loss for the two classes (i.e., source and target)

Network Architecture

- **Discriminator:**
  - 5 CONV layers with 4x4 kernel and stride of 2 (no. of output channels: (64, 128, 256, 512, 1).
  - Except for last CONV layer, each layer is followed by a leaky ReLU.
  - Last layer is followed by a upsampling layer to rescale the output to the size of input.
- **Segmentation Network:**
  - A deep convolutional net (VGG-16) with transformed fully connected layers to convolutional layers.
  - Modified stride of last 2 convolutional layers from 2 to 1.
  - An up-sampling layer along with the softmax output to match the size of the input image.

Evaluation

- We pick the common classes between the source and target and evaluate in terms of IOU of these classes.

For instance,

- ***GTA-5 -> Cityscapes*** contain 19 similar classes.
- ***Synthia -> Cityscapes*** contain 16 similar classes.

We compute mean IOU of the similar classes as the metric.

Multi-Level DA Model

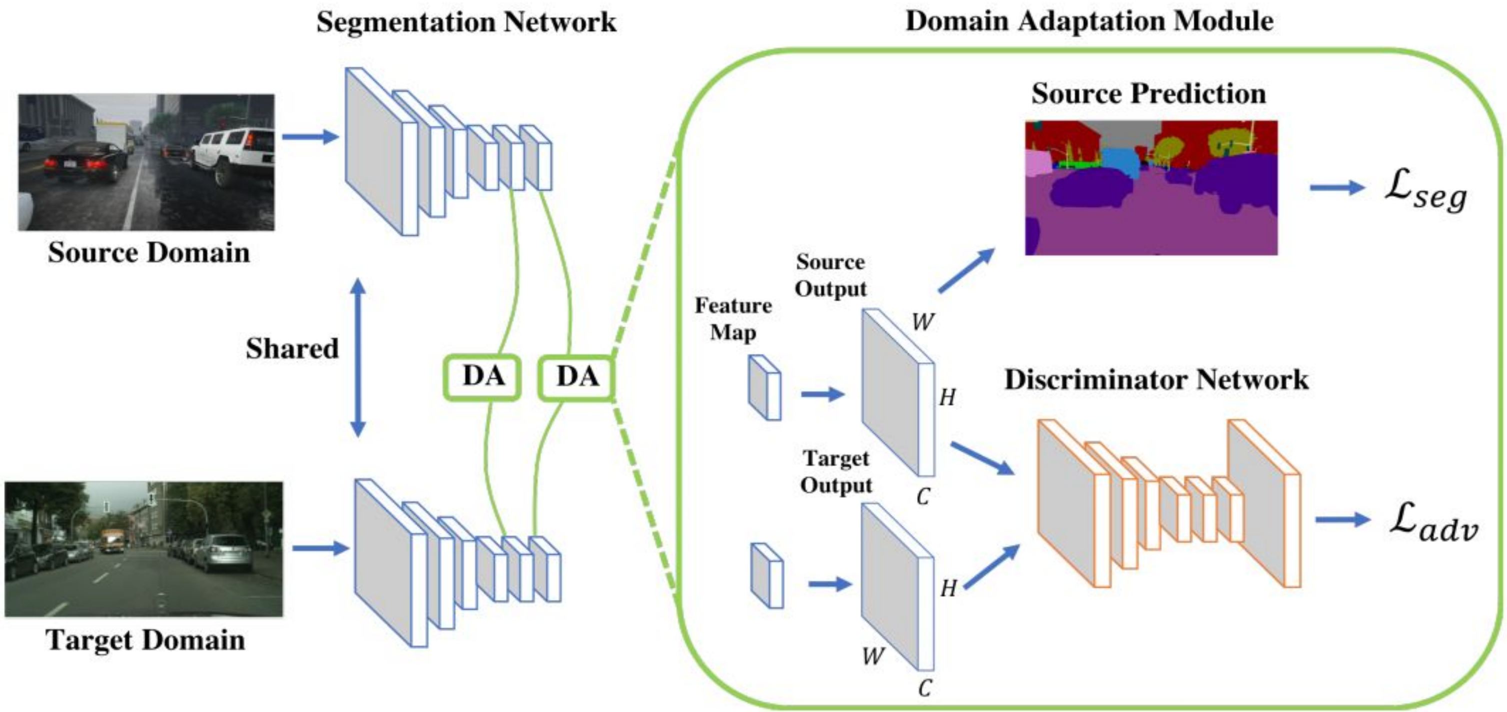


Image Source : https://goo.gl/EcBAq3

Single-Level DA Model (Output Space)

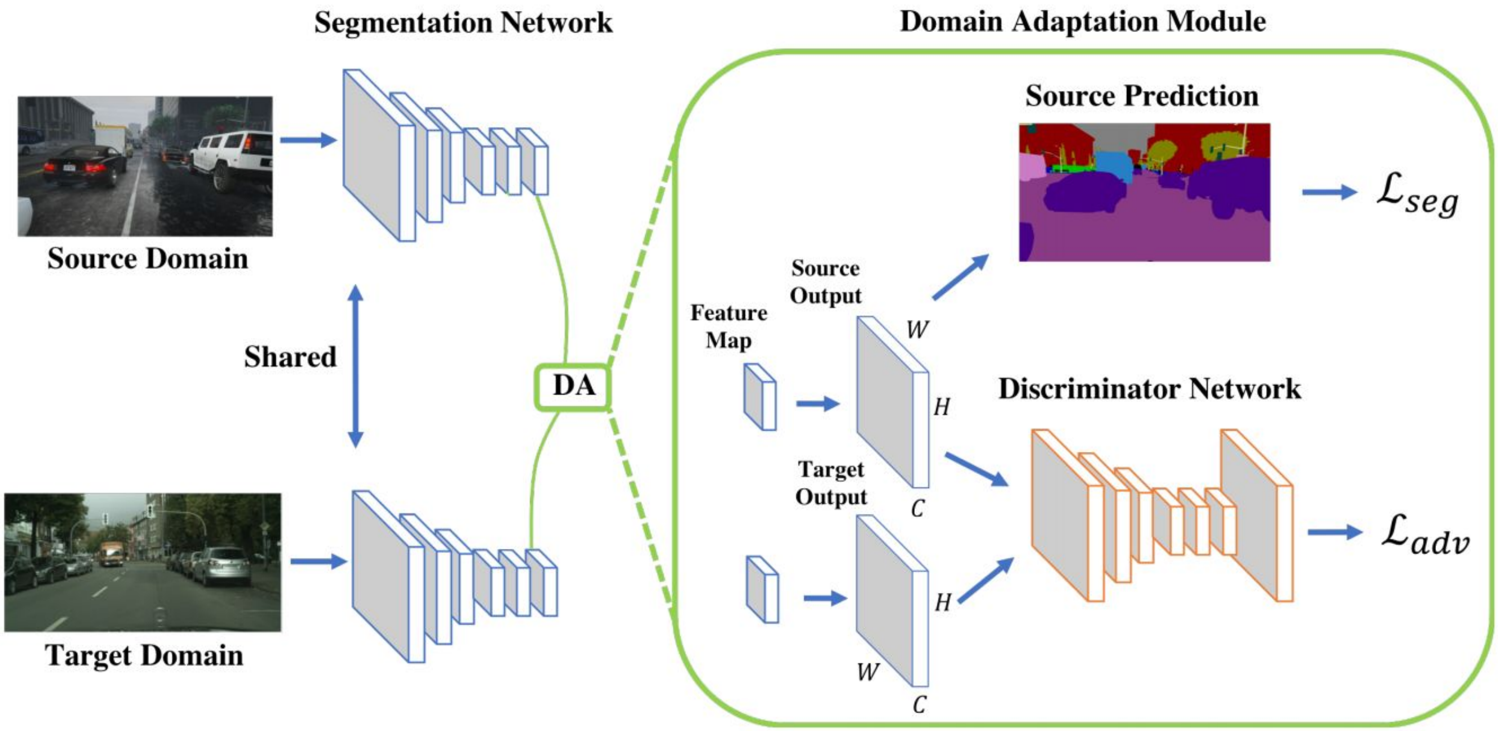


Image Source : https://goo.gl/EcBAq3

Feature-Level DA Model

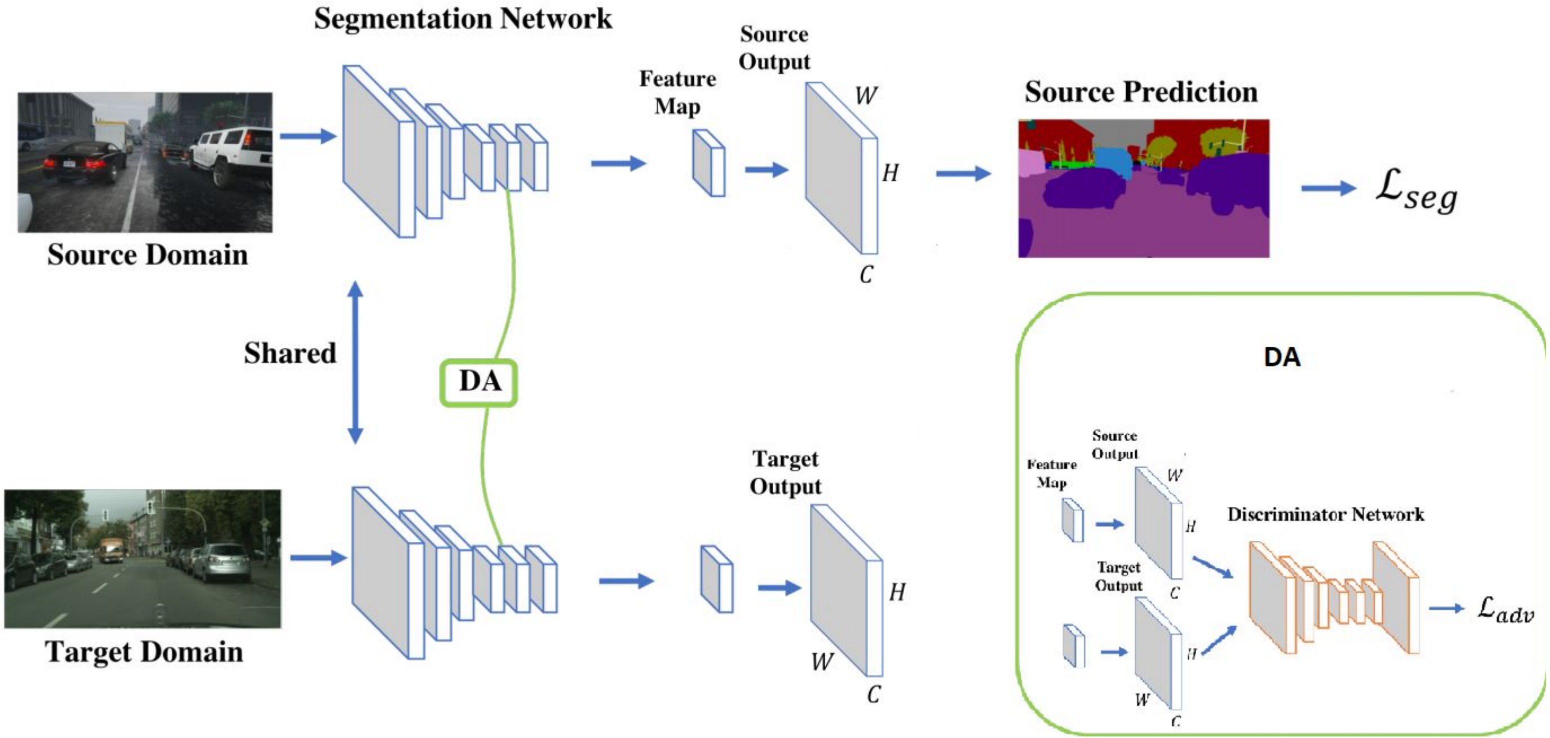
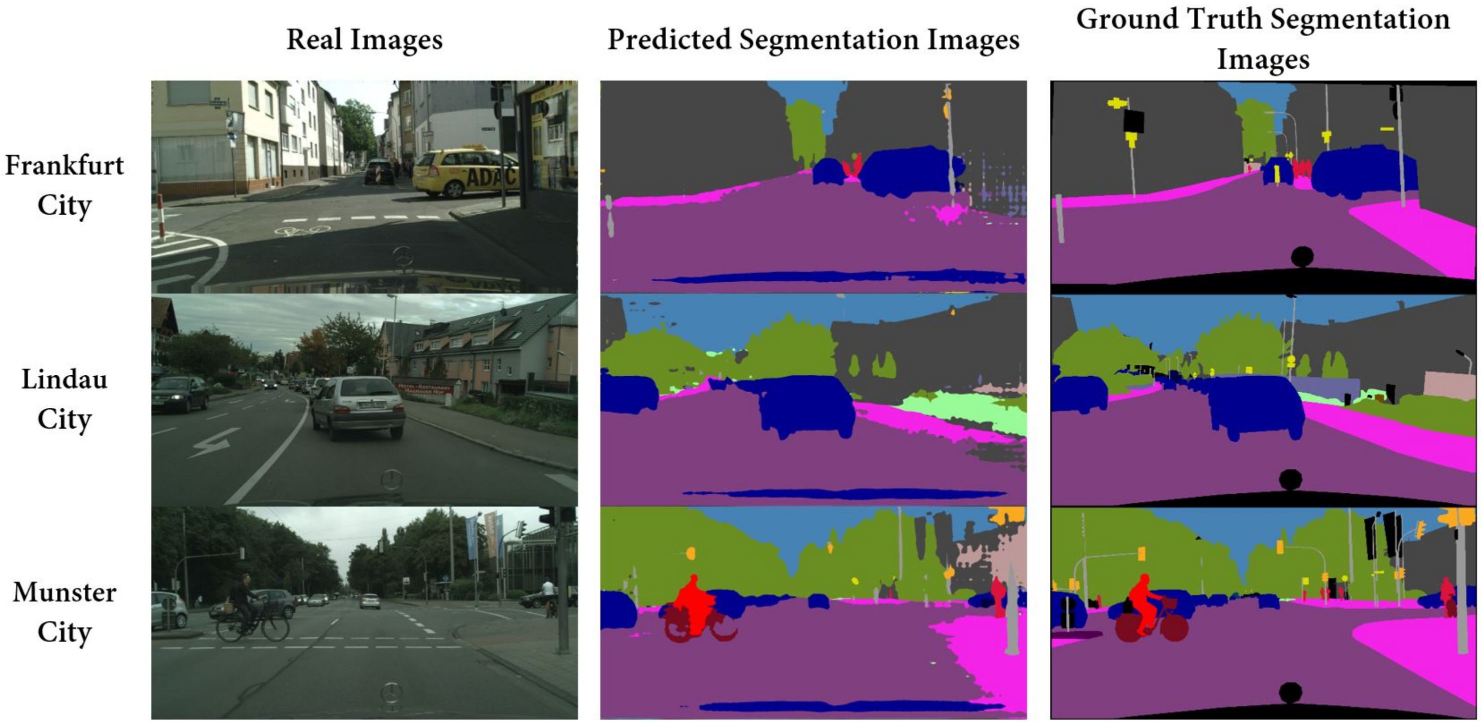


Image Source : https://goo.gl/EcBAq3

GTA5 to CityScapes: Multi-Level DA



GTA5 to CityScapes: Single-Level DA

Real Images

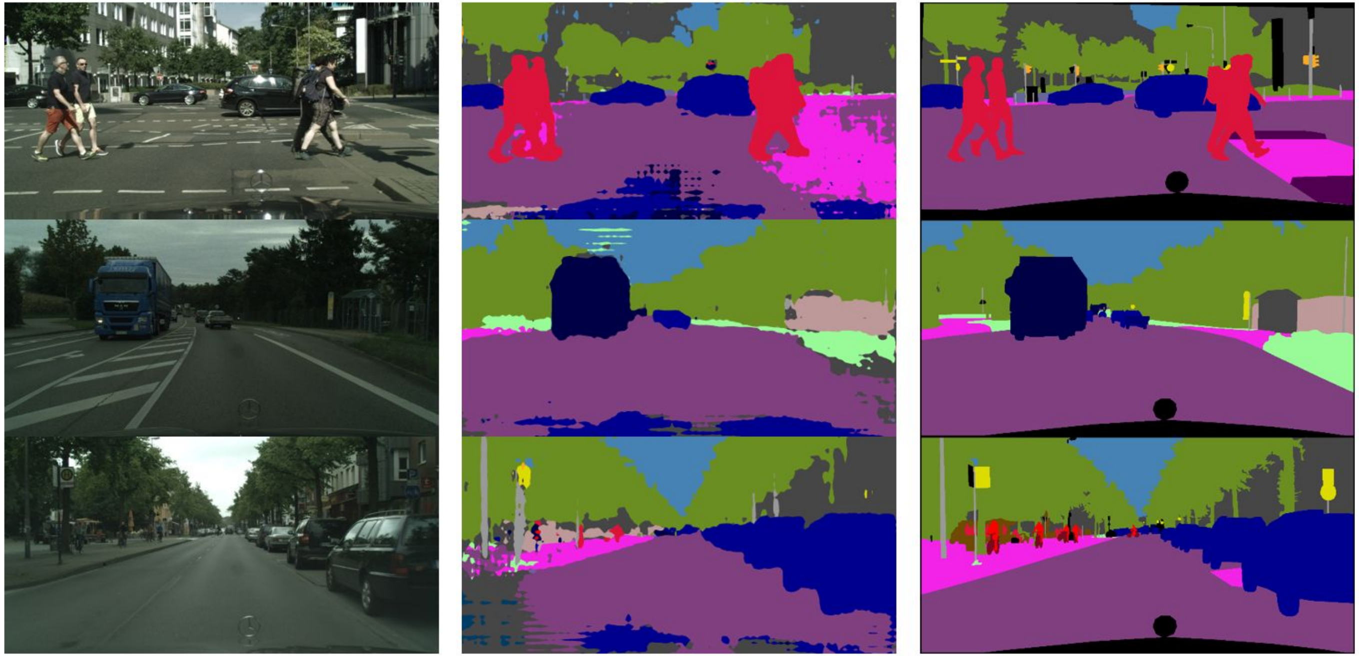
Predicted Segmentation Images

Ground Truth Segmentation Images

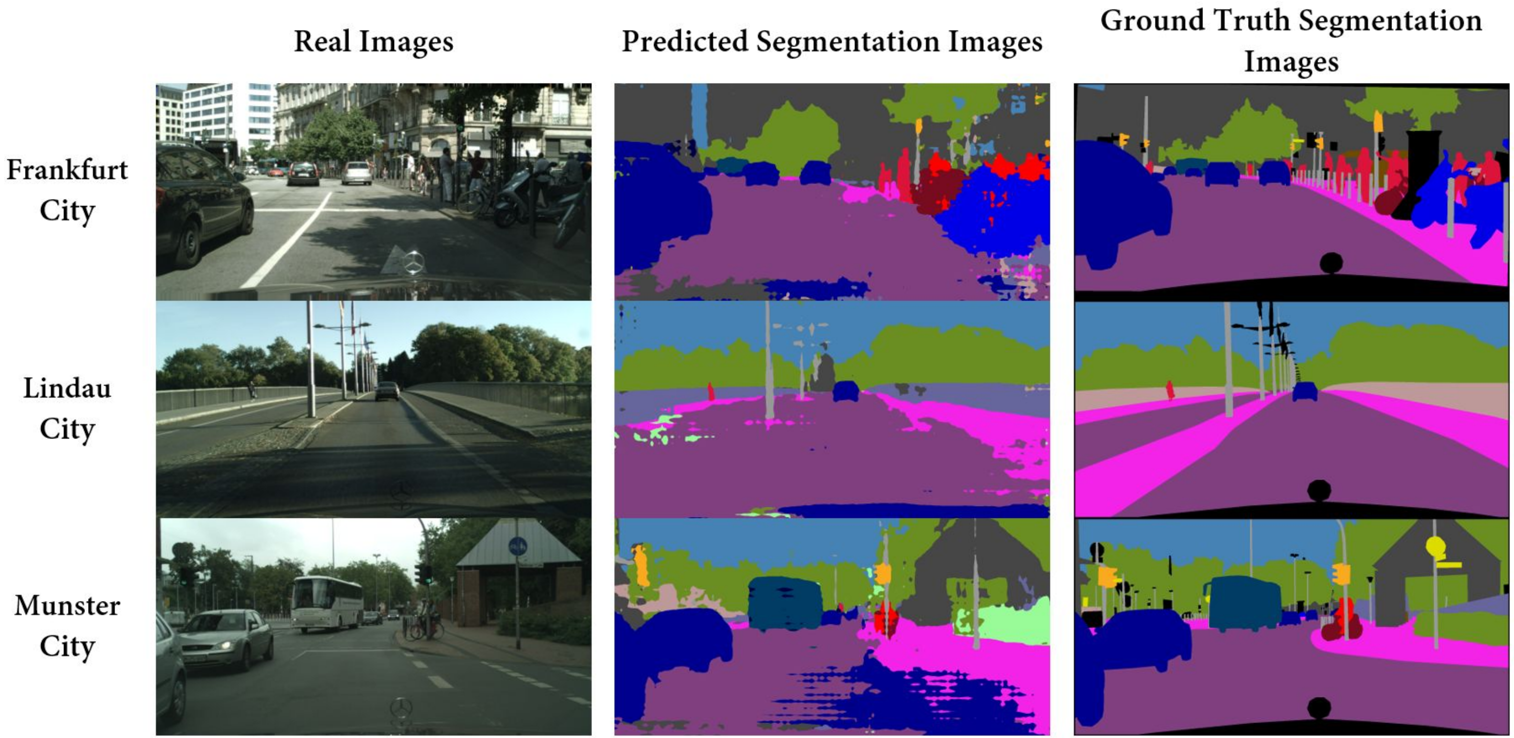
Frankfurt City

Lindau City

Munster City



GTA5 to CityScapes: Feature-Level DA



GTA5 to CityScapes: Method Comparison

Method	mIoU
Feature DA	34.86
Single-Level DA	38.29
Multi-Level DA	**42.35**

GTA5 to CityScapes: Comparison over different Classes

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Mbike	Bike
Feature DA	59.19	29.43	71.51	19.6	19.45	26.57	29.84	17.2	80.28	20.42	73.72	55.94	19.82	43.33	21.46	19.77	0.38	26.19	28.3
Single-Level DA	70.69	26.41	73.65	20.67	21.64	28.39	31.84	17.85	80.49	31.77	72.69	56.97	23.88	66.32	26.92	8.55	2.35	28.08	24.44
Multi-Level DA	**86.46**	**35.96**	**79.92**	**23.41**	**23.27**	**23.87**	**35.24**	**14.77**	**83.35**	**33.25**	**75.62**	**58.49**	**27.55**	**73.65**	**32.48**	**35.42**	**3.85**	**30.05**	**28.11**

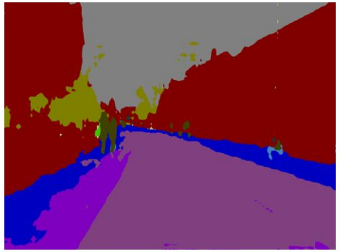
GTA5 to CityScapes (Multi-Level DA)

λ_{adv}	**0.0005**	**0.001**	**0.004**
Output Space	41.75	42.35	41.51

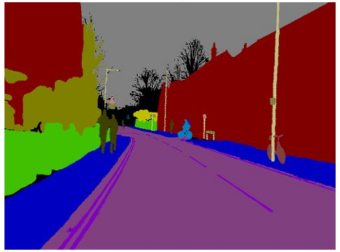
Synthia to Camvid: Multi-Level DA

Real Images

Predicted Segmentation Images

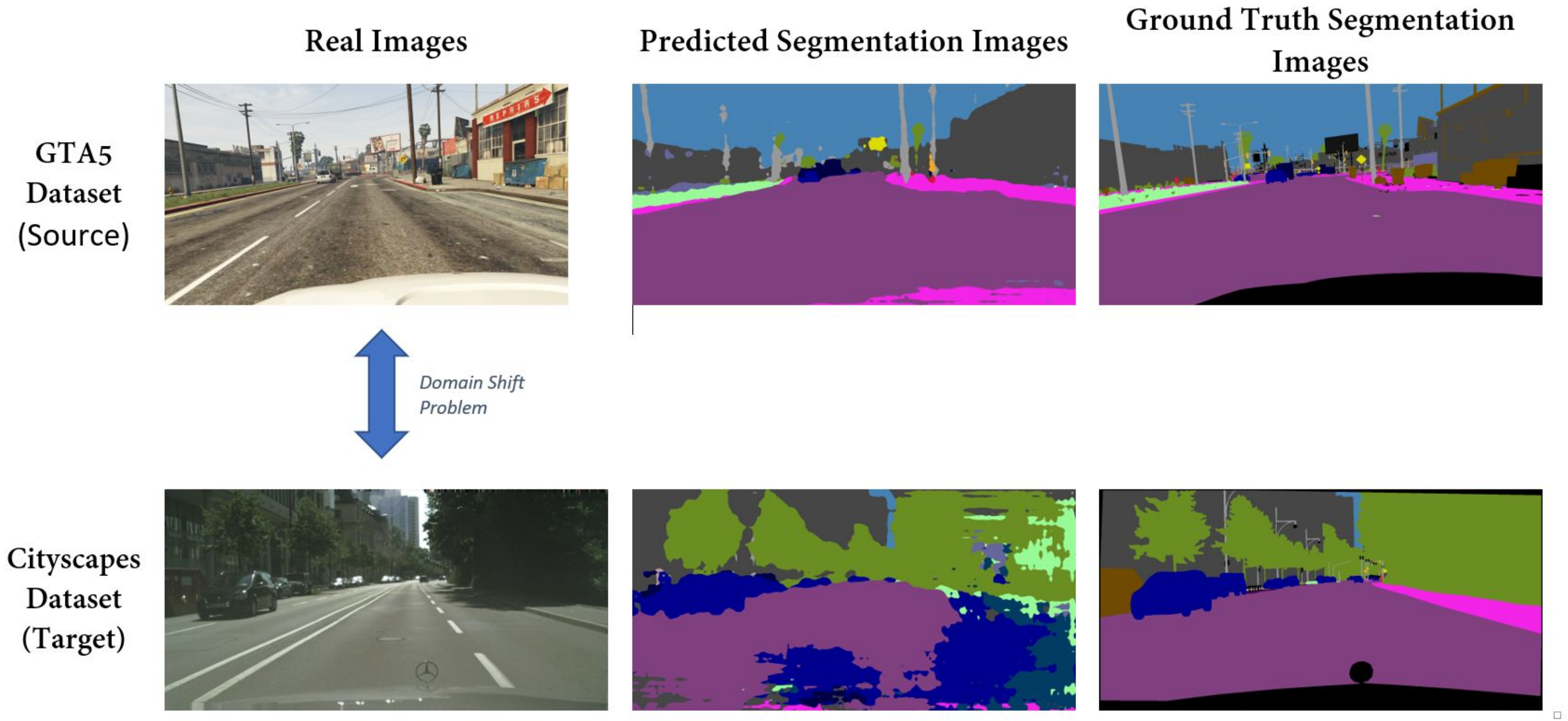


Ground Truth Segmentation Images

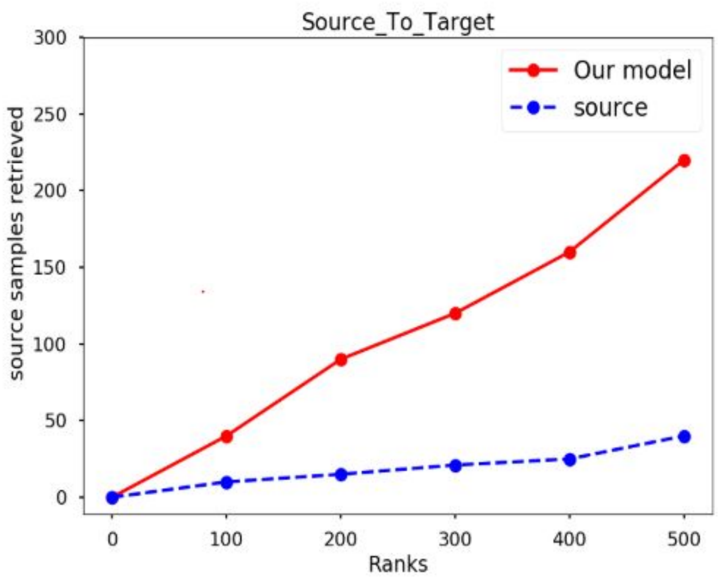


Method	Road	Sidewalk	Building	Pole	Light	Sign	Veg	Terrain	Bus	mIoU
Multi-Level DA	79.43	36.39	72.39	19.07	42.67	49.34	83.69	36.55	17.3	31.2

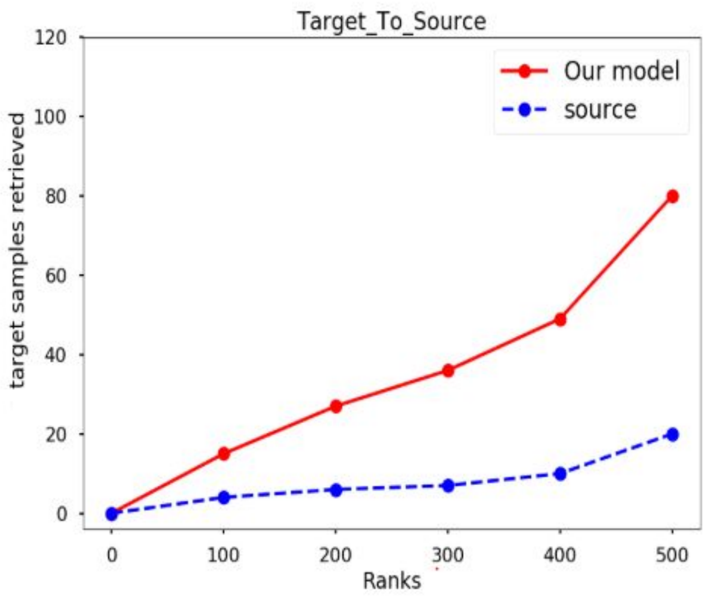
Baseline (Source Only)



Cross Domain Retrieval

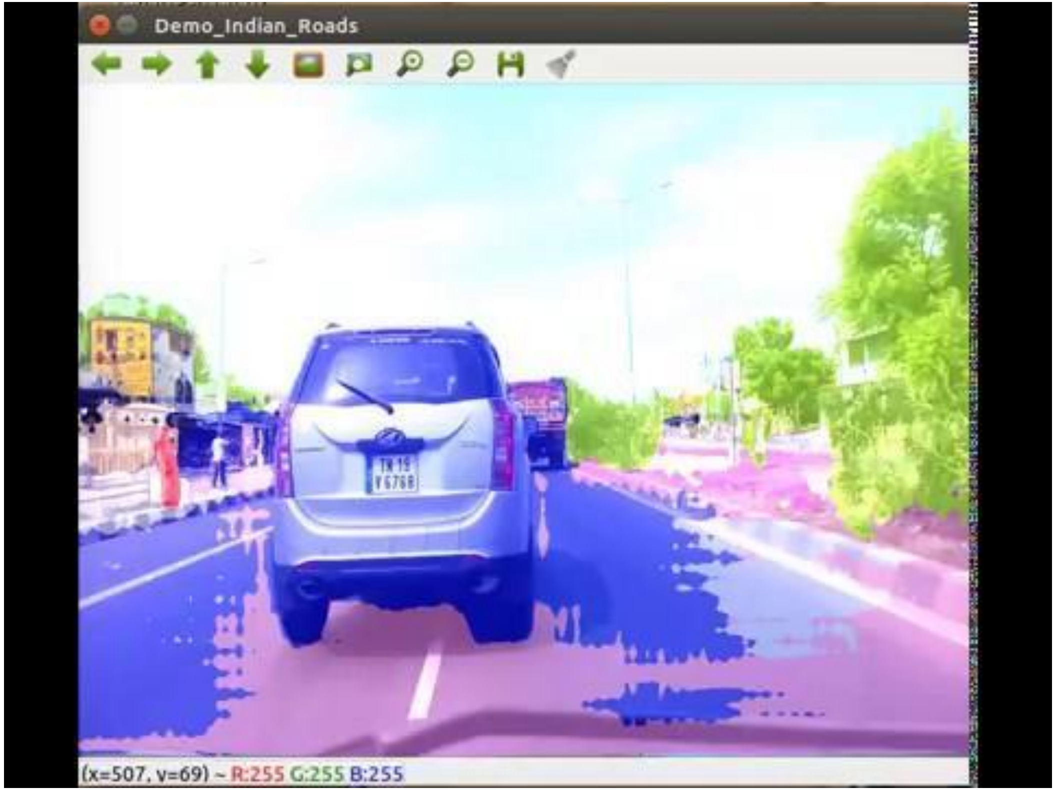


Queried from source and seen if they belong to the target

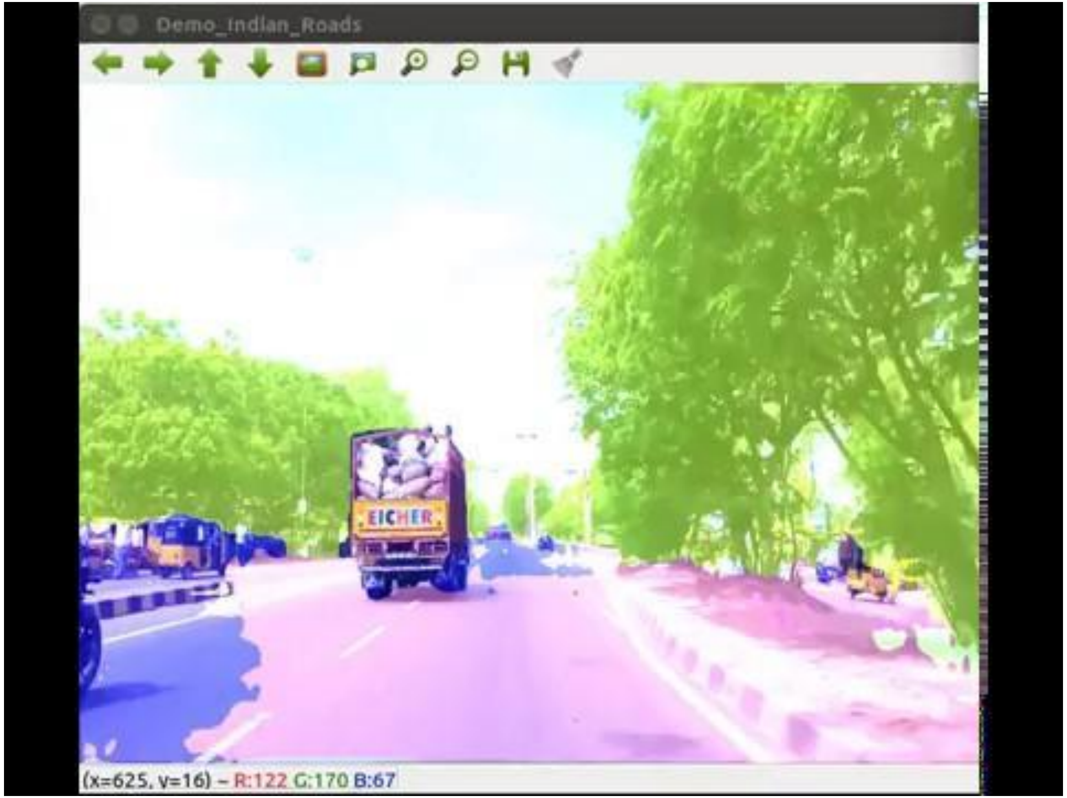


Queried from target and seen if they belong to the source

Demo on Indian Roads - I



Demo on Indian Roads - II



References

- https://arxiv.org/pdf/1802.10349.pdf
- https://arxiv.org/pdf/1711.06969.pdf
- http://synthia-dataset.net/
- https://download.visinf.tu-darmstadt.de/data/from_games/
- https://www.cityscapes-dataset.com/

Thank You