Exploring Transfer from Synthetic Images for Semantic Segmentation of Driving Scenes

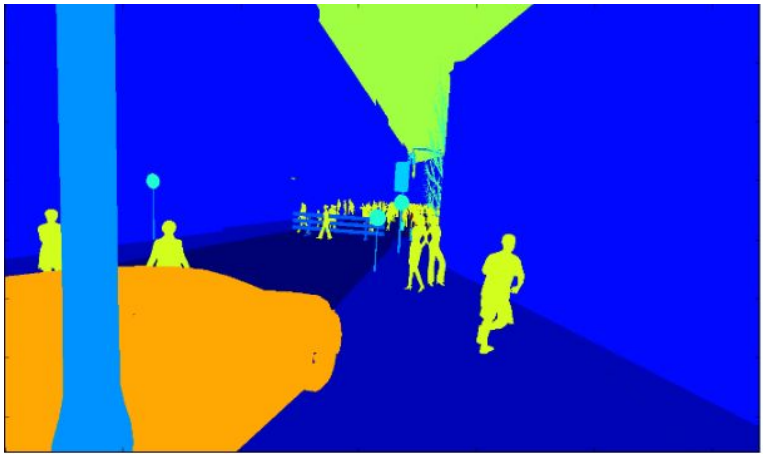
Vashisht Madhavan

Motivation

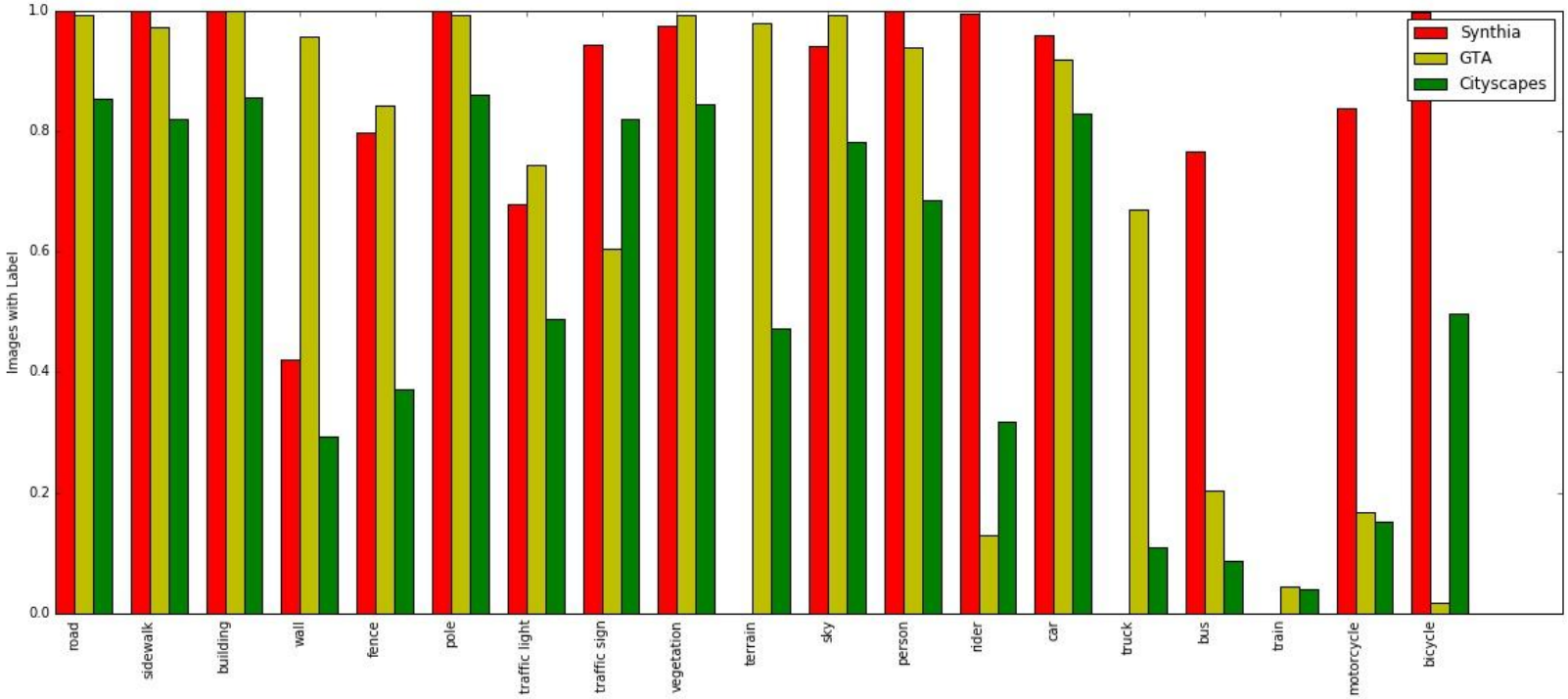
- Games are very realistic these days
- Easy to obtain full pixel labels
- Collecting and labeling real world data is expensive

Problem Statement

- Source model trained on synthetic data
- We have some real world data - level of supervision varies
- What is the best method of transfer to real world data
  - How much supervision is needed?
  - Why do current methods fail to transfer well?
  - Can we leverage raw data from the real domain?



Data



GTA - 25k Images

SYNTHIA - 9k Images

Cityscapes(**Real**) - 5k Images

Label Set

Group	Classes
flat	road · sidewalk · parking+ · rail track+
human	person* · rider*
vehicle	car* · truck* · bus* · on rails* · motorcycle* · bicycle* · caravan*+ · trailer*+
construction	building · wall · fence · guard rail+ · bridge+ · tunnel+
object	pole · pole group+ · traffic sign · traffic light
nature	vegetation · terrain
sky	sky
void	ground+ · dynamic+ · static+

Supervised Approach

- Fine Tuning with various amounts of labeled examples
- Dilation FCN from Yu and Koltun

Very Low = 75 Labels
Low = 450 Labels
Medium = 1000 Labels
High = 2500 Labels

Experiment	Flat	Nature	Object	Sky	Construction	Human	Vehicle	**Mean**
G Baseline	0.737	0.715	0.022	0.676	0.606	0.363	0.652	**0.539**
S Baseline	0.218	0.462	0.076	0.665	0.335	0.516	0.483	**0.394**

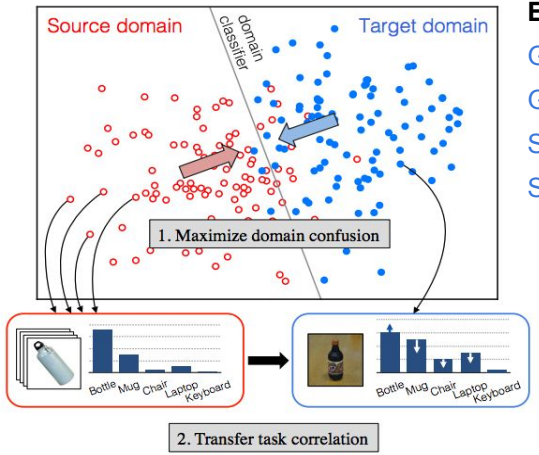
Experiment	Flat	Nature	Object	Sky	Construction	Human	Vehicle	**Mean**
G+C FT Very Low	0.959	0.86	0.318	0.841	0.826	0.554	0.791	**0.736**
G+C FT Low	0.967	0.874	0.385	0.883	0.853	0.617	0.825	**0.772**
G+C FT Medium	0.971	0.883	0.443	0.889	0.865	0.645	0.845	**0.792**
G+C FT High	0.971	0.885	0.436	0.884	0.866	0.648	0.848	**0.791**
S+C FT Very Low	0.961	0.866	0.431	0.84	0.827	0.644	0.797	**0.766**
S+C FT Low	0.973	0.892	0.523	0.889	0.871	0.718	0.861	**0.818**
S+C FT Medium	0.976	0.898	0.538	0.906	0.884	0.725	0.874	**0.829**
S+C FT High	0.976	0.899	0.568	0.909	0.884	0.736	0.878	**0.836**
C Full	0.977	0.902	0.547	0.909	0.89	0.726	0.882	**0.833**

SYNTHIA has lower baseline but higher accuracy with Supervision!

GTA has representations that don't play well with cityscapes

Unsupervised Approach

- Hoffman and Wang (CVPR ‘17 Submission)
- Domain Adversarial Loss to align representations



Experiment	Road	Sidewalk	Building	Traffic Light	Traffic Sign	Vegetation
G Baseline	0.491	0.239	0.582	0.104	0.04	0.775
G DA	0.894	0.374	0.731	0.142	0.057	0.792
S Baseline	0.074	0.155	0.317	0.001	0.008	0.473
S DA	0.115	0.196	0.308	0.01	0.117	0.593

Person	Rider	Car	**Average**
0.363	0.021	0.698	0.245
0.442	0.052	0.724	0.305
0.465	0.029	0.541	0.152
0.512	0.038	0.617	0.181

Tzeng + Hoffman ICCV ‘15

Domain Shift

Global Domain Shift - Image Level Differences between Datasets

Domain	Linear SVM	Poly Kernel SVM	Max Mean Discrepancy
GTA	92%	94.5%	0.459495
SYNTHIA	96.5%	96%	0.72624

Classifier accuracy between of **conv5** features of **VGG-16** pretrained on PASCAL

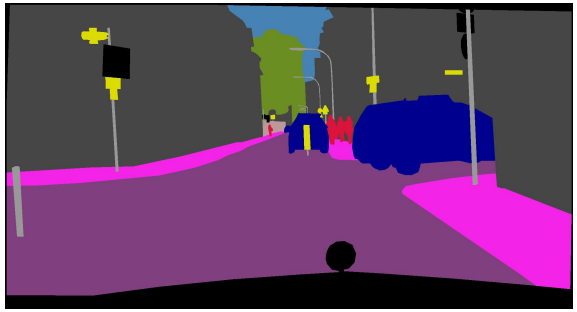
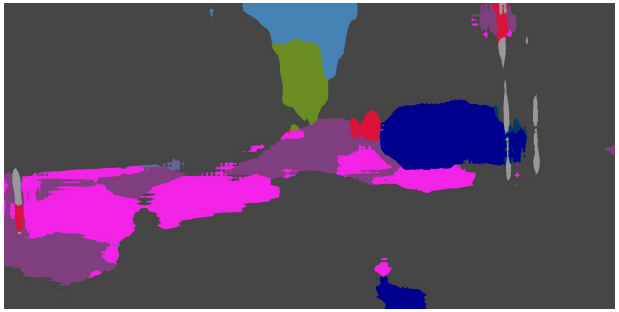
Category-Specific Shift - Distribution and Appearance Differences for Objects

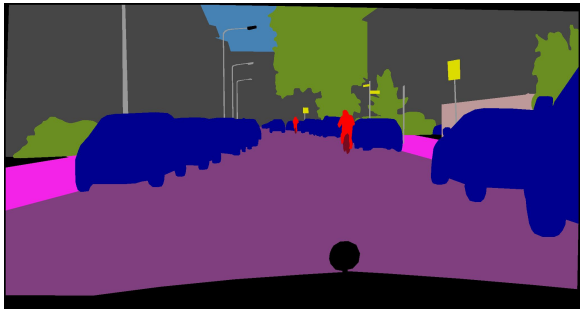
Domain	Road	Building	Pole	Traffic Light	Traffic Sign	Person	Car
GTA	10.172	6.08	0.27	0.68	5.037	4.14	12.08
SYNTHIA	14.78	10.72	0.204	0.51	4.02	0.68	12.6

KL Divergence between histograms of proportion of object in image

Results

SYNTHIA Fine Tuning



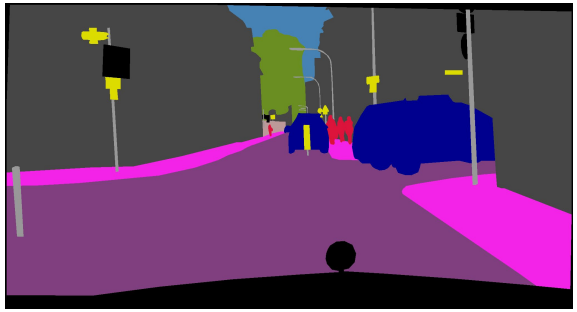
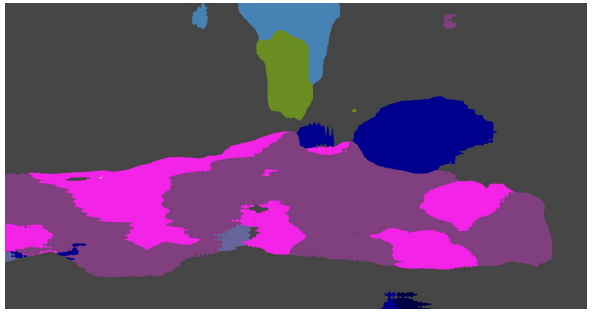
Ground Truth

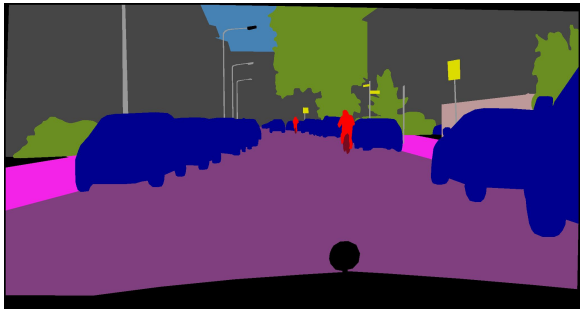
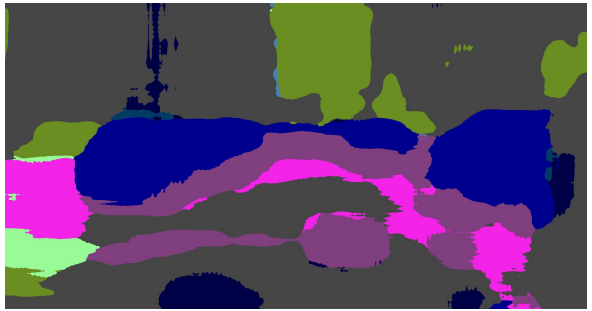
Baseline

Very Low Supervision

Results

GTA Fine Tuning


Ground Truth

Baseline

Very Low Supervision