A FULLY CONVOLUTIONAL TRI-BRANCH NETWORK (FCTN) FOR DOMAIN ADAPTATION

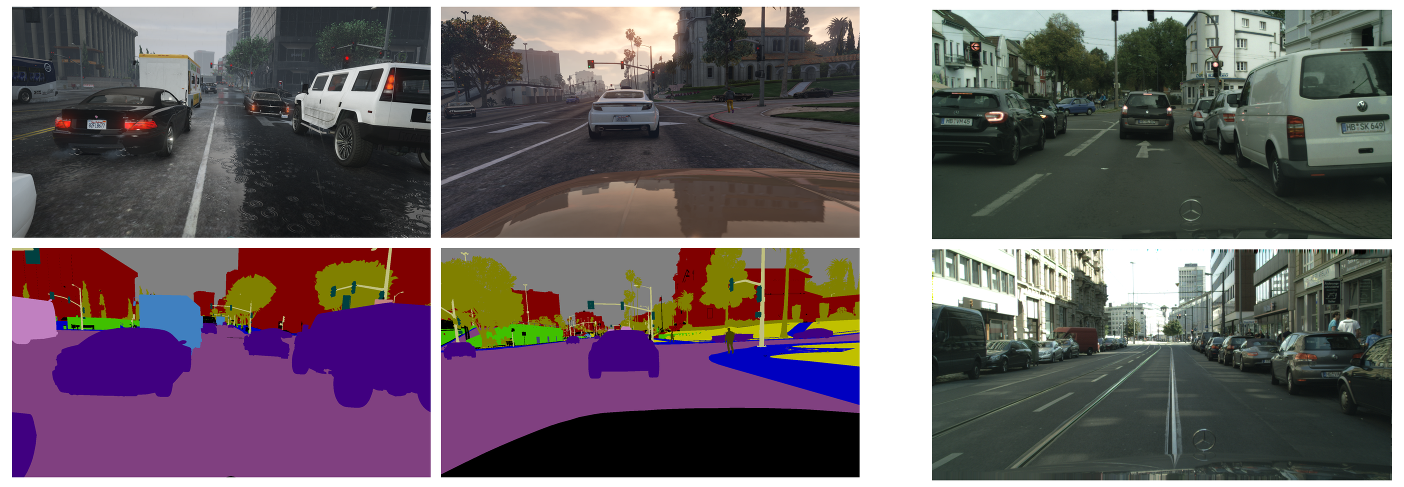
Junting Zhang, Chen Liang and C.-C. Jay Kuo
University of Southern California

Introduction

Motivation

- Major limitation of deep learning: data-hungry
  - Pixel-wise semantic labels are expensive
- Dataset bias is prevalent in many applications

Problem Definition



Source domain: synthetic images with pixel-wise semantic labels

Target domain: unlabeled real world images

- Goal: leverages labeled data in the source domain, to learn a segmenter for unlabeled data in a target domain

Datasets:

- Source Domain: GTA5 (train/val/test: 16k/5k/4k)
- Target Domain: Cityscapes (train/val: 3,149/500)
- # Classes: 34 (19 classes are considered in evaluation)

Related Work

Feature Distribution Alignment

- Distance minimization: maximum mean discrepancy, correlation alignment, etc.
- Adversarial training: domain discriminator
- Major limitation: Assume the existence of a universal classifier that can perform well on samples drawn from whichever domain

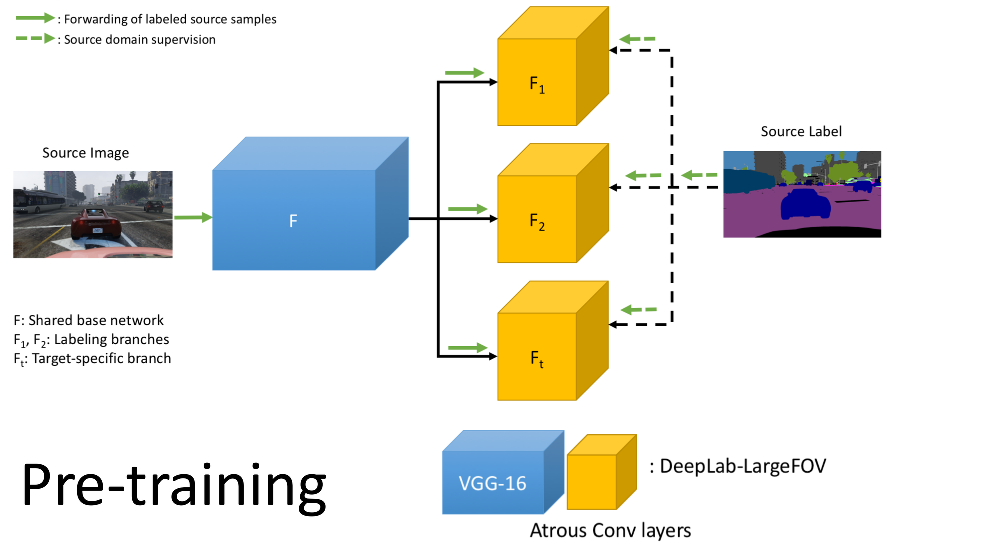
Methodology

Tri-training for Unsupervised Domain Adaptation

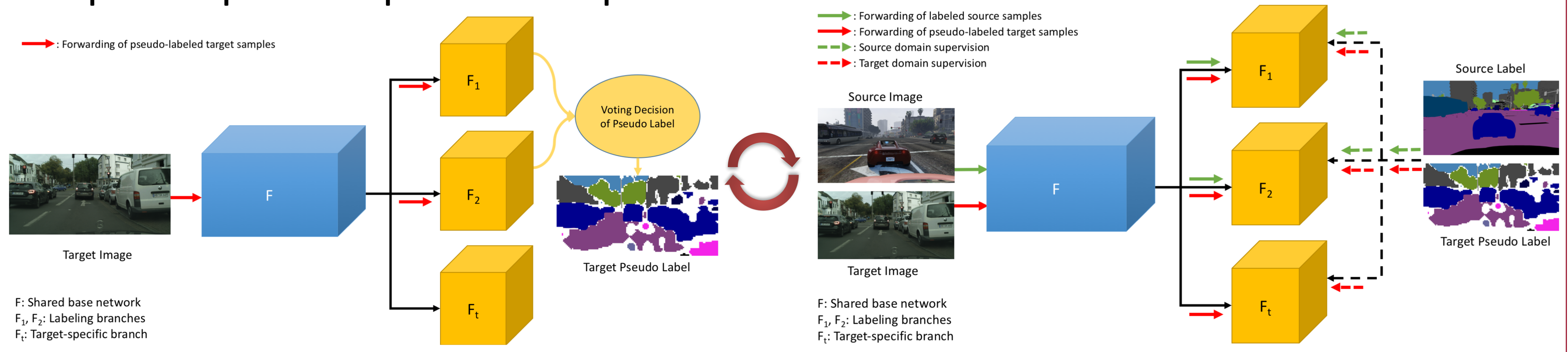
- Classifier 1 (C_1) and Classifier 2 (C_2) are trained with source domain data
- C_1 and C_2 assigns pseudo label to a target sample if:
  1. C_1 and C_2 gives consistent prediction
  2. At least one classifier has high confidence score
- C_t learns from pseudo labels

FCTN Architecture and Training Scheme

Step 1: Pre-train three branches
Step 2: Assign pseudo labels for target domain images
Step 3: Joint train F_1 and F_2 with images from both domains, and train Ft with pseudo-labeled target images
Step 4: Repeat Step 2 and Step 3



Pre-training



Pseudo-labeling

Joint training

Regularized Training

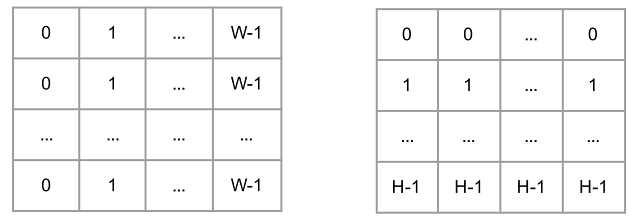
C_1 and C_2 **CAN NOT** be identical:

- Initialize the two branches differently
- Incur a weight-constraint loss among the convolutional kernels of the two branches (F_1 and F_2):

$$L_w = \frac{\vec{w_1}\cdot\vec{w_2}}{\|\vec{w_1}\|\|\vec{w_2}\|}$$

Encoding Prior Knowledge

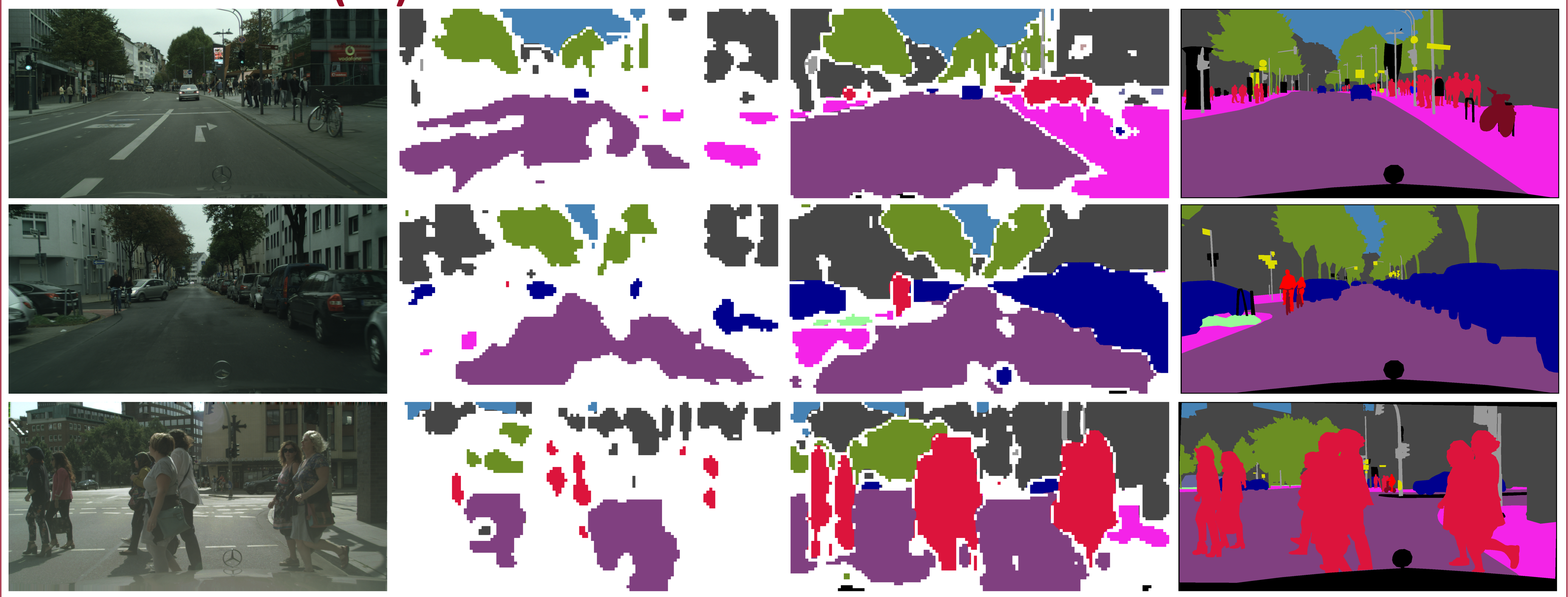
- Layout of the traffic scene images is unique and domain independent
- CNN is translation-invariant
- Two additional feature maps to encode spatial information explicitly



X Coordinate Map Y Coordinate Map

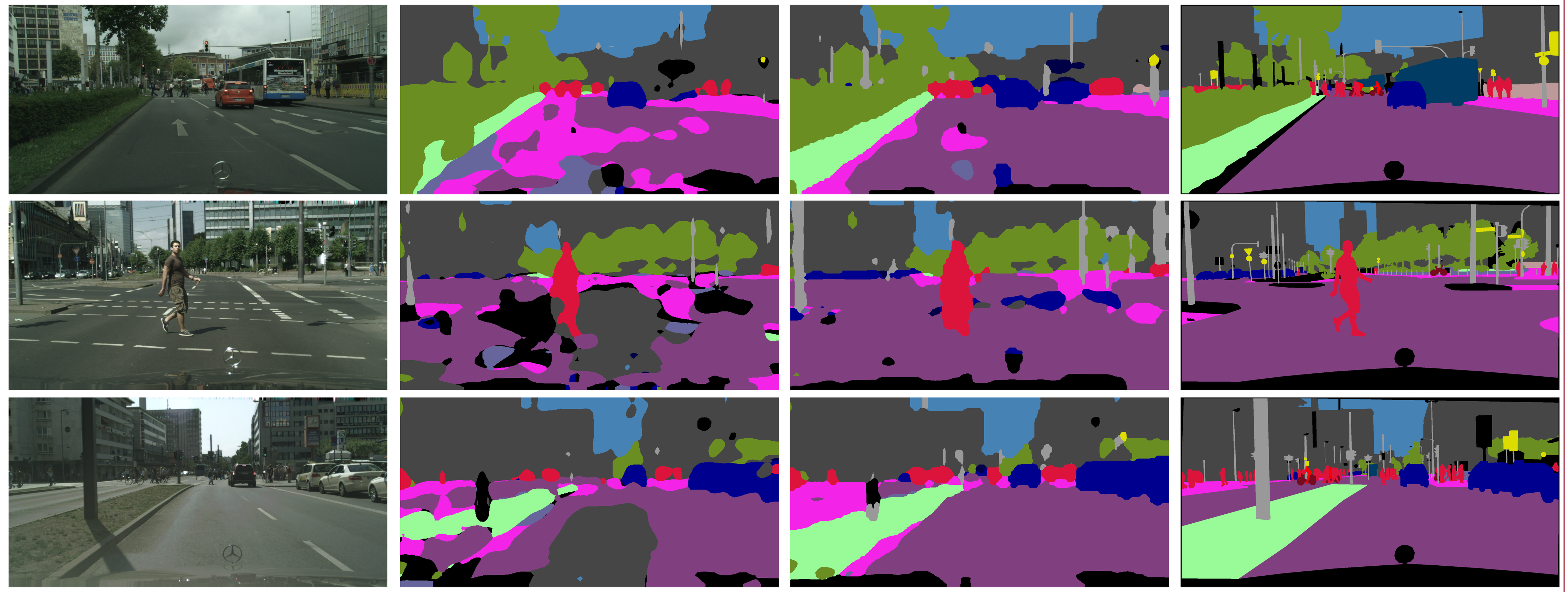
Experiments

Pseudo Labels (PL)



Original Image | 1st Round PL | 2nd Round PL | Ground Truth

Qualitative Results



Original Image | No Adaptation | After Adaptation | Ground Truth

Quantitative Results

Model	road	sidewlk	bldg.	wall	fence	pole	t. light	t. sign	veg.	terr.	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
	per-class IoU																			
No Adapt	31.9	18.9	47.7	7.4	3.1	16.0	10.4	1.0	76.5	13.0	58.9	36.0	1.0	67.1	9.5	3.7	0.0	0.0	0.0	21.1
FCN [13]	70.4	**32.4**	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
No Adapt	18.1	6.8	64.1	7.3	8.7	21.0	14.9	16.8	45.9	2.4	64.4	41.6	**17.5**	55.3	8.4	5.0	**6.9**	4.3	13.8	22.3
CDA [5]	26.4	22.0	74.7	6.0	**11.9**	8.4	16.3	11.1	75.7	13.3	**66.5**	38.0	9.3	55.2	**18.8**	**18.9**	0.0	**16.8**	**14.6**	27.8
No Adapt	59.7	24.8	66.8	12.8	7.9	11.9	14.2	4.2	78.7	22.3	65.2	44.1	2.0	67.8	9.6	2.4	0.6	2.2	0.0	26.2
Round 1	66.9	25.6	74.7	17.5	10.3	17.1	18.4	8.0	79.7	34.8	59.7	**46.7**	0.0	77.1	10.0	1.8	0.0	0.0	0.0	28.9
Round 2	**72.2**	28.4	**74.9**	**18.3**	10.8	**24.0**	**25.3**	**17.9**	**80.1**	**36.7**	61.1	44.7	0.0	**74.5**	8.9	1.5	0.0	0.0	0.0	**30.5**

Table 1: Adaptation from GTA to Cityscapes. All numbers are measured in %. The last three rows show our results before adaptation, after one and two rounds of curriculum learning using the proposed FCTN, respectively.