Universal Semi-Supervised Semantic Segmentation

Tarun Kalluri[1], Girish Varma[1], Manmohan Chandraker[2], CV Jawahar[1]

CVIT, IIIT Hyderabad[1] University of California San Diego[2]

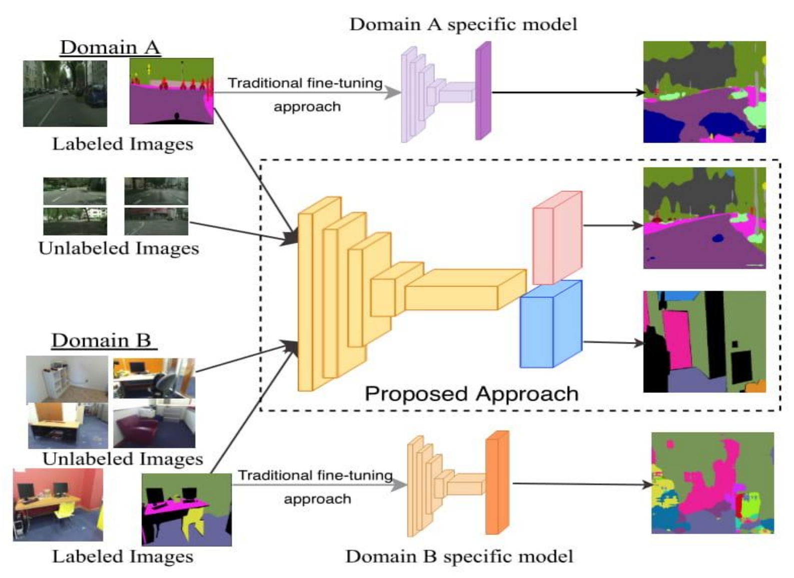
Paper/Code

Overview: Universal Segmentation

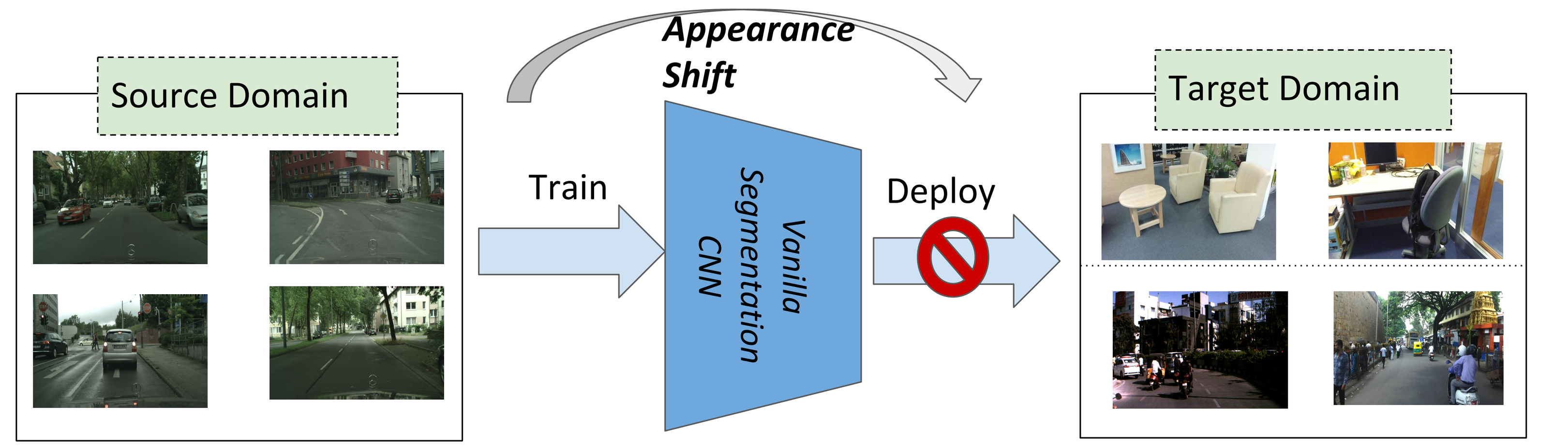
Obtain a common semantic segmentation model across widely disparate domains having limited labeled data.

A good universal model ensures that, across all domains,

- ✓ A single model is deployed
- ✓ Unlabeled data is used
- ✓ Performance is improved
- ✓ And label spaces (semantic content) may differ.



Challenge: Domain Shift + Different Labels

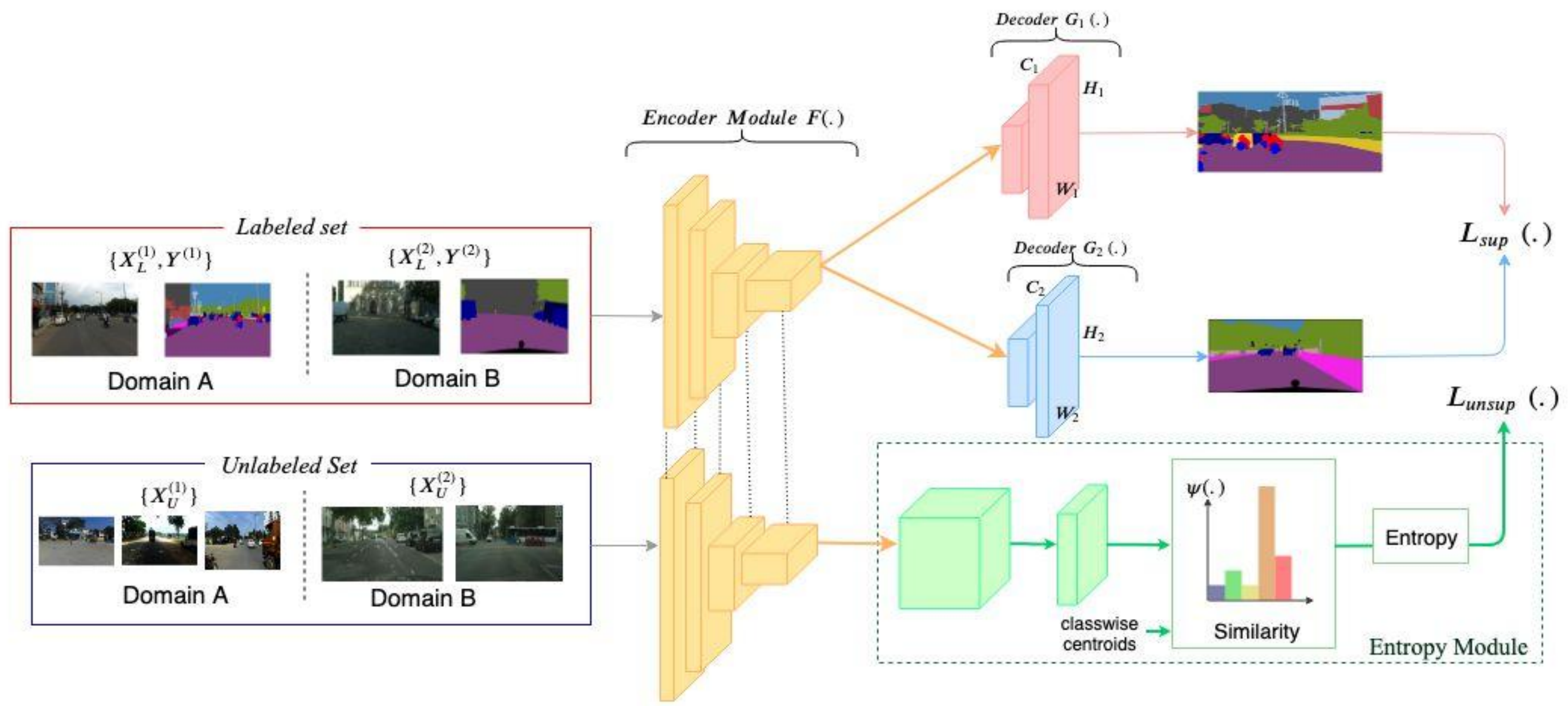


- ➢ Models trained on a single domain are not usable in other domains due to *Domain Shift* and *Semantic Shift*.
- ➢ Training individual models for different domains results in deployment overhead, doesn't exploit shared structure among these domains.

	Source Unlabeled Data	Target Unlabeled Data	Joint Model	Mixed Labels Support
Fine Tuning	✗	✗	✗	✓
Semi-supervised [Hung 2018]	✓	✗	✗	NA
CyCADA [Hoffman 2018]	✗	✓	✓	✗
Joint Training	✗	✗	✓	✓
Our Approach	✓	✓	✓	✓

Prior works fall short in addressing the semantic change, which we do by using large scale unsupervised images.

Approach: Feature Alignment Using Entropy Regularization



Training Objective: Supervised + Unsupervised Losses

Unsupervised Losses

➢ $L_{u,c} = \mathcal{H}(\sigma([v_{12}])) + \mathcal{H}(\sigma([v_{21}]))$

➢ $L_{u,w} = \mathcal{H}(\sigma([v_{11}])) + \mathcal{H}(\sigma([v_{22}]))$

$$[v_{ij}] = \phi\left(\mathcal{E}\left(\mathcal{F}\left(x_u^{(i)}\right)\right), \mathrm{c}^{(j)}\right)$$

Supervised Loss

➢ $L_{sup} = \sum_k \frac{1}{N_l^{(k)}} \sum_{x_i \in D^{(k)}} \psi_k\left(y_i, \mathcal{G}_k\left(\mathcal{F}\left(x_i\right)\right)\right)$

Total Loss

$$L_t = L_{sup} + \lambda_1 \cdot L_{u,c} + \lambda_2 \cdot L_{u,w}$$

Datasets

Cityscapes (N=2975)

✚

CamVid (N=366)

IDD (Indian Roads) (N = 6993)

SUN (Indoor) (N = 5050)

Experimental Results

Method	N=375		
	CS	CamVid	Avg.
Train on CS	55.07	48.52	51.80
Train on CVD	26.45	60.61	43.53
Hung *et al.* 2018	58.80	-	-
Souly *et al.* 2017	-	58.20	-
Univ-basic ($\mathcal{L}_s$)	53.14	65.33	59.24
Univ-cross (+ $\mathcal{L}_c$)	56.36	63.34	59.85
Univ-full (+ $\mathcal{L}_c, \mathcal{L}_w$)	55.92	64.72	**60.32**

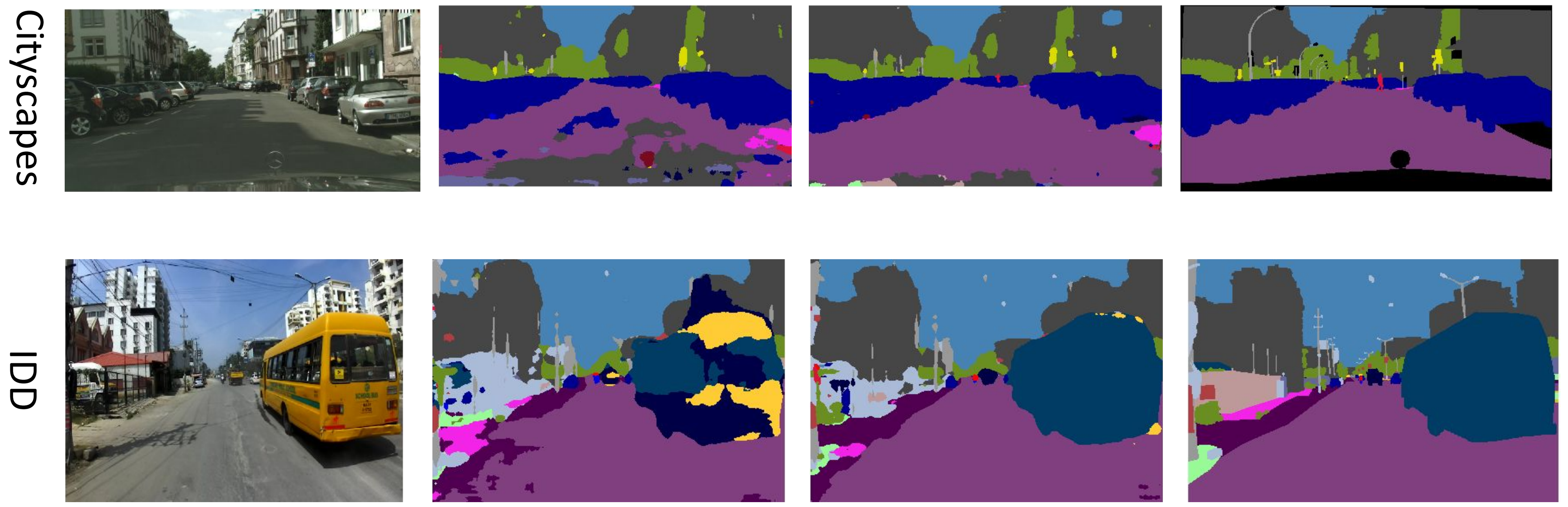
New SOTA with semi supervised data!

Method	N=100 (Resnet-18)		
	CS	IDD	Avg.
Train on CS	40.97	14.64	27.81
Train on IDD	25.05	26.53	25.79
Univ-basic	37.94	25.21	31.58
Univ-full	36.48	27.45	**31.97**

Method	Labeled Examples	CS	SUN	Avg.
Train on CS	1.5k	64.23	15.47	39.85
Train on SUN	1.5k	15.61	42.52	29.07
SceneNet [McCormac 2017]	Full(5.3k)	-	49.8	-
Univ-basic	1.5k	58.01	31.55	44.78
Ours[Univ-full]	1.5k	57.91	43.12	**50.52**

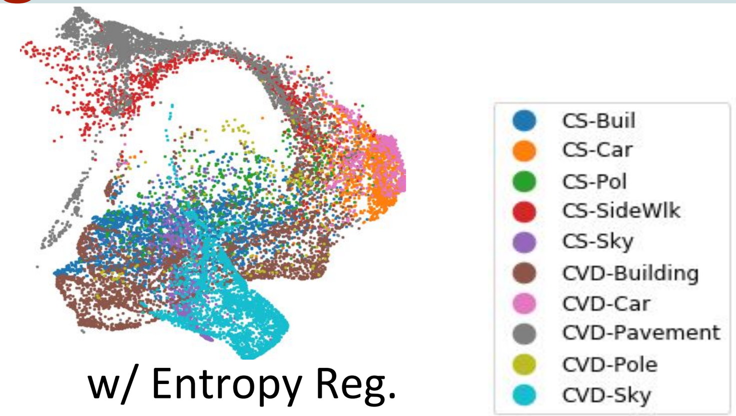
28% labeled data from SUN RGB dataset with no synthetic examples, recovers ~88% of performance obtained with full dataset.

Qualitative Improvements In Segmentation



tSNE Embedding Visualization

w/o Entropy Reg.



w/ Entropy Reg.

Visually similar features, like *Building* and *SideWalk* from Cityscapes and CamVid are positively aligned, helping in learning agnostic discriminative features.