Universal Semi-Supervised Semantic Segmentation

Tarun Kalluri[1] Girish Varma[1] Manmohan Chandraker[2] C V Jawahar[1]

[1]Center for Visual Information Technology, IIIT Hyderabad
[2]University of California, San Diego

tarun.05.kalluri@gmail.com

Abstract

In recent years, the need for semantic segmentation has arisen across several different applications and environments. However, the expense and redundancy of annotation often limits the quantity of labels available for training in any domain, while deployment is easier if a single model works well across domains. In this paper, we pose the novel problem of universal semi-supervised semantic segmentation and propose a solution framework, to meet the dual needs of lower annotation and deployment costs. In contrast to counterpoints such as fine tuning, joint training or unsupervised domain adaptation, universal semi-supervised segmentation ensures that across all domains: (i) a single model is deployed, (ii) unlabeled data is used, (iii) performance is improved, (iv) only a few labels are needed and (v) label spaces may differ. To address this, we minimize supervised as well as within and cross-domain unsupervised losses, introducing a novel feature alignment objective based on pixel-aware entropy regularization for the latter. We demonstrate quantitative advantages over other approaches on several combinations of segmentation datasets across different geographies (Germany, England, India) and environments (outdoors, indoors), as well as qualitative insights on the aligned representations.

1. Introduction

Semantic segmentation is the task of pixel level classification of an image into a predefined set of categories. State-of-the-art semantic segmentation architectures [35, 3, 8] pretrain deep networks for a classification task on datasets like ImageNet [13, 53] and then fine-tune on finely annotated labeled examples [12, 64]. The availability of such large-scale labeled datasets has been crucial to achieve high accuracies for semantic segmentation in applications ranging from natural scene understanding [18] to medical imaging [51]. However, performance often suffers even in the presence of a minor domain shift. For example, a segmentation model trained on a driving dataset from a specific geographic location may not generalize to a new city due to differences in weather, lighting or traffic density. Further, a segmentation model trained on traffic scenes for outdoor navigation may not be applicable for an indoor robot.

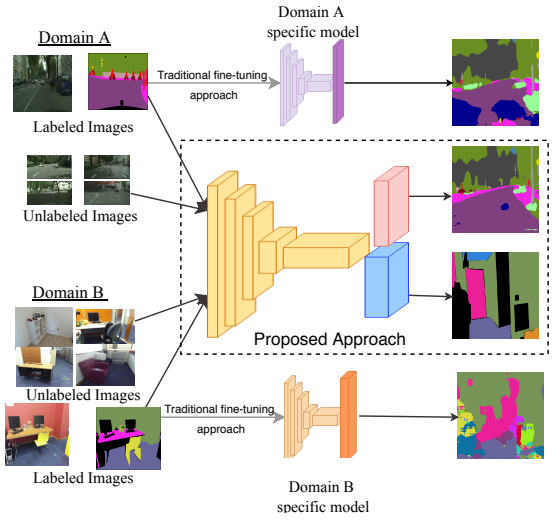


Figure 1: Proposed universal segmentation model can be jointly trained across datasets with different label spaces, making use of the large amounts of unlabeled data available. Traditional transfer learning based approaches typically require training separate models for each domain.

While such domain shift is a challenge for any machine learning problem, it is particularly exacerbated for segmentation where human annotation is highly prohibitive and redundant for different locations and tasks. Thus, there is a growing interest towards learning segmentation representations that may be shared across domains. A prominent line of work addresses this through unsupervised domain adaptation from a labeled source to an unlabeled target domain [25, 61, 10, 42, 6]. But there remain limitations. For instance, unsupervised domain adaptation usually does not leverage target domain data to improve source performance. Further, it is designed for the restrictive setting of large-scale labeled source domain and unlabeled target domain. While

some applications such as self-driving have large-scale annotated datasets for particular source domains (for example synthetic datasets like Synthia [52]), the vast majority of applications only have limited data in any domain. Finally, most of the above works assume that the target label set matches with the source one, which is often not the case in practice. For example, road scene segmentation across different countries, or segmentation across outdoor and indoor scenes, have domain-specific label sets.

In this paper, we propose and address the novel problem of *universal semi-supervised semantic segmentation* as a practical setting for many real-world applications. It seeks to aggregate knowledge from several different domains during training, each of which has few labeled examples but several unlabeled examples. The goal is to simultaneously limit training cost through reduced annotations and deployment cost by obtaining a single model to be used across domains. Label spaces may be partially or fully non-overlapping across domains. While fine-tuning a source model on a small amount of target data is a possible counterpoint, it usually requires plentiful source labels and necessitates deployment of a separate model in every domain due to catastrophic forgetting [40]. Another option is joint training, which does yield a unified model across domains, but does not leverage unlabeled data available in each domain. Our semi-supervised universal segmentation approach leverages both limited labeled and larger-scale unlabeled data in every domain, to obtain a single model that performs well across domains. Table 1 presents the advantage of the proposed semi-supervised universal segmentation over some of the existing approaches.

In particular, we use the labeled examples in each domain to supervise the universal model, akin to multi-tasking [31, 39, 30], albeit with limited labels. We simultaneously make use of the large number of unlabeled examples to align pixel level deep feature representations from multiple domains using entropy regularization based objective functions. Entropy regularization uses unsupervised examples and helps in encouraging low density separation between the feature representations and improve the confidence of predictions. Moreover models trained on one domain typically result in noisy predictions and high entropy output maps when deployed in a different domain, and the proposed cross dataset entropy minimization encourages refined prediction maps across datasets. We calculate the similarity score vector between the encoder outputs at each pixel and the label embeddings (computed from class prototypes [58]), and minimize the entropy of this discrete distribution to positively align similar examples between the labeled and the unlabeled images. We do this unsupervised alignment both within domain, as well as across domains.

We believe such within and cross-domain alignment is fruitful even with non-overlapping label spaces, particularly so for semantic segmentation, since label definitions often encode relationships that may positively reinforce performance in each domain. For instance, two road scene datasets such as Cityscapes [12] and IDD [64] might have different labels, but share similar label hierarchies. Even an outdoor dataset like Cityscapes and an indoor one like SUN [59] may have label relationships, for example, between horizontal (road, floor) and vertical (building, wall) classes. Similar observations have been made for multi-task training [70].

	Source Unlabeled Data	Target Unlabeled Data	Joint Model	Mixed Labels Support
Fine Tuning	✗	✗	✗	✓
Semi-supervised [28, 60]	✓	✗	✗	NA
CyCADA [24]	✗	✓	✓	✗
Joint Training	✗	✗	✓	✓
Our Approach	✓	✓	✓	✓

Table 1: Comparison of Universal Semi-Supervised Segmentation against existing methods.

We posit that our pixel wise entropy-based objective discovers such alignments to improve over joint training, as demonstrated quantitatively and qualitatively in our experiments. Specifically, our experiments lend insights across various notions of domain gaps. With Cityscapes [12] as one of domains (road scenes in Germany), we derive universal models with respect to CamVid (roads in England) [4], IDD (roads in India) [64] and SUN (indoor rooms) [59]. In each case, our semi-supervised universal model improves over fine-tuning and joint training, with visualizations of the learned feature representations demonstrating conceptually meaningful alignments. We use dilated residual networks in our experiments [69], but the framework is equally applicable to any of the existing deep encoder-decoder architectures for semantic segmentation.

Our Contributions

- We propose a universal segmentation framework to train a single joint model on multiple domains with disparate label spaces to improve performance on each domain. This framework adds no extra parameters or significant overhead during inference compared to existing methods for deep semantic segmentation.
- We introduce a pixel-level entropy regularization scheme to train semantic segmentation architectures using datasets with few labeled examples and larger quantities of unlabeled examples (Section 3).
- We demonstrate the effectiveness of our alignment over a wide variety of indoor [59] and outdoor [12, 64, 4] segmentation datasets with various degrees of label overlaps. We also compare our results with other semi-supervised approaches, based on adversarial losses, giving improved results (Section 4).

2. Related Work

Semantic Segmentation Semantic segmentation in computer vision is the task of assigning semantic labels to each pixel of an image. Most of the state of the art models for semantic segmentation [69, 35, 43, 3, 8, 50] have been possible largely due to breakthroughs in deep learning that have pushed the boundaries of performance in image classification [32, 22, 23] and related tasks. The pioneering work in [35] proposes an end-to-end trainable network for semantic segmentation by replacing the fully connected layers of pretrained AlexNet [32] and VGG Net [57] with fully convolutional layers that aggregate spatial information across various resolutions. Noh *et al*. [43] use transpose convolutions to build a learnable decoder module, while DeepLab network [8] uses artrous convolutions along with artrous spatial pyramid pooling for better aggregation of spatial features. Segmentation architectures based on dilated convolutions [68] for real time semantic segmentation have also been proposed [69, 50].

Semi Supervised Learning Most of the existing semantic segmentation architectures require large scale annotation of labeled data for achieving good results. To address this limitation, various semi supervised learning methods have been proposed in [60, 46, 28, 26, 66], which make use of easily available large scale unsupervised or weakly supervised data during training. While these approaches deliver competitive results when trained and deployed on a specific dataset, the need for learning efficient segmentation models transferable across domains and environments having limited training data remains.

Transfer Learning and Domain Adaptation Transfer learning [67] involves transferring deep feature representations learned in one domain or task to another domain or task where labeled data availability is low. Previous works demonstrate transfer learning capabilities between related tasks [14, 71, 45, 48] or even completely different tasks [19, 49, 35]. Unsupervised domain adaptation is a related paradigm which leverages labeled data from a source domain to learn a classifier for a new unsupervised target domain in the presence of a domain shift. Various generative and discriminative domain adaptation methods have been proposed for classification tasks in [16, 17, 63, 62, 47, 5] and for semantic scene segmentation in [25, 61, 10, 24, 9, 27, 72].

Most of these works in domain adaptation assume equal source and target dataset label spaces or a subset target label space, which is not the most general case for real world applications. To address this limitation with the domain adaptation approaches, we propose a method similar to [37] which works in the extreme case of non-intersecting label spaces. Moreover, pixel level adaptation based methods are typically focused on using knowledge from a large labeled source domain (eg. Synthia [52]) to improve performance on a specific target domain, while we propose a joint training framework to train a single model that delivers good performance on both the domains.

Universal Segmentation Multitask learning [7] is shown to improve performance for many tasks that share useful relationships between them in computer vision [56, 31, 70], natural language processing [11, 39, 30] and speech recognition [55]. Universal Segmentation builds on this idea by training a single joint model that is useful across multiple semantic segmentation domains with possibly different label spaces to make use of transferable representations at lower levels of the network. Liang *et al*. [34] first propose the idea of universal segmentation by performing dynamic propagation through a label hierarchy graph constructed from an external knowledge source like WordNet. We propose an alternative method to perform universal segmentation without the need for any outside knowledge source or additional model parameters during inference, and instead make efficient use of the large set of unlabeled examples in each of the domains for unsupervised feature alignment. Following the success of metric learning based approaches in tasks such as fine grained classification [2, 1], latent hierarchy learning [54] and zero-shot prediction [44, 15, 33], we use pixel level class prototypes [58] for performing semantic transfer across domains.

3. Problem Description

In this section, we explain the framework used to train a single model across different segmentation datasets with possibly disparate label spaces using a novel pixel aware entropy regularization objective.

We have d datasets $\{\mathcal{D}^{(i)}\}_{i=1}^{d}$, each of which has few labeled examples and many unlabeled examples. The labeled images and corresponding labels from $\mathcal{D}^{(i)}$ are denoted by $\{\mathrm{X}_l^{(i)}, \mathrm{Y}^{(i)}\}_{i=1}^{N_l^{(i)}}$, where $\mathrm{Y}^{(i)} \in \mathcal{Y}_i$, and $N_l^{(i)}$ is the number of labeled examples. The unlabeled images are represented by $\{\mathrm{X}_u^{(i)}\}_{i=1}^{N_u^{(i)}}$, and $N_u^{(i)}$ is the number of unlabeled examples. We work with domains with very few labeled examples ($N_u^{(i)} \gg N_l^{(i)}$), and consider the general case of non-intersecting label spaces, so that $\mathcal{Y}_p \neq \mathcal{Y}_q$ for any p, q. The label spaces might still have a partial overlap between them, which is common in the case of segmentation datasets. For ease of notation, we consider the special case of two datasets $\{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}\}$, but similar idea can be applied for the case of multiple datasets as well.

The proposed universal segmentation model is summarized in Figure 2. Deep semantic segmentation architectures generally consist of an encoder module which aggregates the

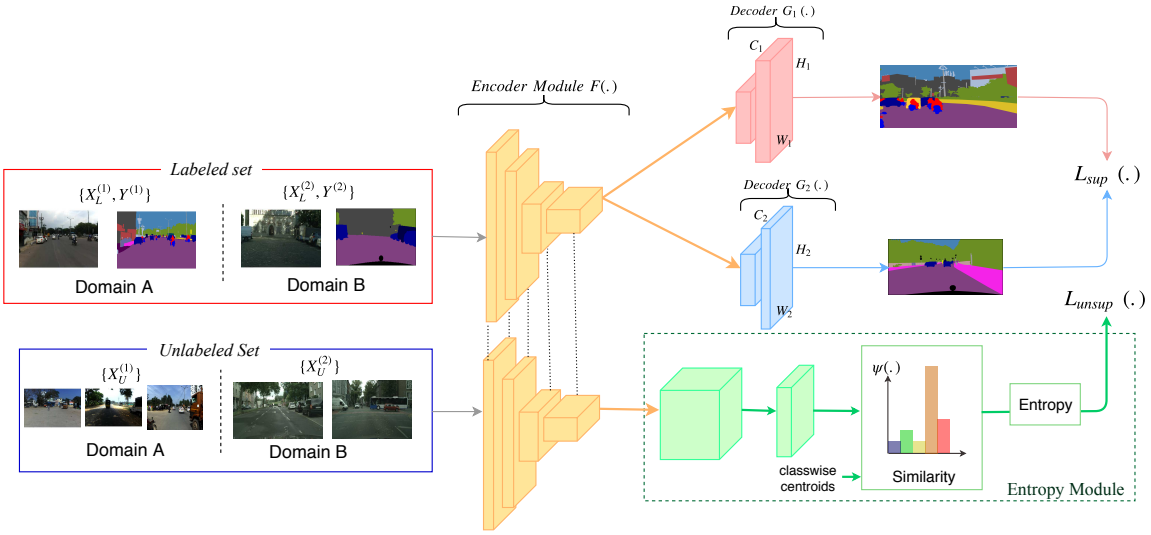


Figure 2: Different modules in the proposed universal semantic segmentation framework. $\{\mathrm{X}_l^{(1)}, \mathrm{Y}^{(1)}\}$, $\{\mathrm{X}_l^{(2)}, \mathrm{Y}^{(2)}\}$ are the set of labeled examples and $\mathrm{X}_u^{(1)}$, $\mathrm{X}_u^{(2)}$ are the set of unlabeled examples. The entropy module uses the unlabeled examples to perform alignment of pixel wise features from multiple domains by calculating pixel wise similarity with the labels, and minimizing the entropy of this discrete distribution.

spatial information across various resolutions and a decoder module that consists of a classifier and an up sampler to enable pixel wise predictions at a resolution that matches the input. In order to enable joint training with multiple datasets, we modify this encoder decoder architecture by having a shared encoder module $\mathcal{F}$ and different decoder layers $\mathcal{G}_1(.)$, $\mathcal{G}_2(.)$ for prediction in different label spaces. For a labeled input image x_l, the pixel wise predictions are denoted by $\hat{y}^{(k)} = \mathcal{G}_k(\mathcal{F}(\mathrm{x}_l))$ for $k = 1, 2$ which, along with the labeled annotations, gives us the supervised loss. To make use of the semantic information from the unlabeled examples X_u, we propose an entropy regularization module $\mathcal{E}$. This entropy module takes as input the output of the encoder $\mathcal{F}(.)$ to give pixel wise representation outputs in an embedding space. The entropy of the similarity score vector of these embedding representations with the label embeddings results in the unsupervised loss term. Each of these loss terms is explained in detail in the following sections.

Supervised Loss The supervised loss is the softmax cross entropy loss between the predicted segmentation mask $\hat{y}$ and the corresponding pixel wise ground truth segmentation masks for all labeled examples. Specifically, for the samples from dataset k,

$$\mathcal{L}_S^{(k)} = \frac{1}{N_l^{(k)}} \sum_{\mathrm{x}_i \in \mathcal{D}^{(k)}} \psi_k \left(\mathrm{y}_i, \mathcal{G}_k \left(\mathcal{F} \left(\mathrm{x}_i \right) \right) \right), \qquad (1)$$

where ψ_k is the softmax cross entropy loss function over the label space $\mathcal{Y}_k$, which is averaged over all the pixels of the segmentation map. $\mathcal{L}_S^{(1)}$ and $\mathcal{L}_S^{(2)}$ together comprise the supervised loss term $\mathcal{L}_S$.

Entropy Module The large number of unsupervised images available provides us with rich information regarding the visual similarity between the domains and the label structure, which the existing methods on adversarial based semi supervised segmentation [60, 28] or universal segmentation [34] do not exploit. To address this limitation, we propose using entropy regularization to transfer the information from labeled images to the unlabeled images, as well as among the unlabeled images between the datasets. Entropy regularization is proved to encourage low density separation between the clusters in the feature space [20], hence resulting in high confidence predictions and smooth output maps for semi supervised learning. A crucial difference between some previous works which use entropy regularization for semi supervised learning [20, 36, 65] and ours is that we perform entropy regularization in a separate embedding space using an *entropy module* $\mathcal{E}$, unlike the other works which apply this objective directly in the softmax output space. This embedding approach helps in achieving semantic transfer between datasets with disparate label sets, hence aiding in closely aligning the visually similar pixel level features calculated from the segmentation network from both the datasets.

The entropy module is explained in Figure 3, and works similar to the decoder module in a segmentation architecture. Firstly, we project the encoder outputs from the segmentation network from both datasets into a common d dimensional embedding space $\mathbb{R}^d$, and upsample this output map to match the size of the input. Then, a similarity metric ϕ, which operates on each pixel, is used to calculate the similarity score of the embedding representations with each of the d

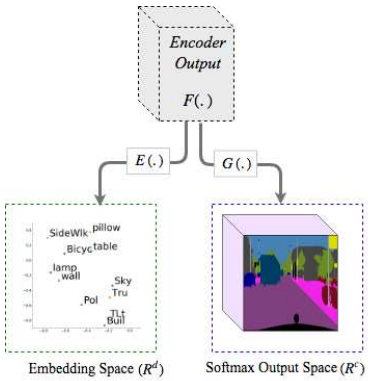


Figure 3: In addition to a traditional decoder layer that outputs predictions in the respective label spaces $\mathbb{R}^c$, we also have an entropy module $\mathcal{E}(.)$ that first maps the features of both the domains into a common embedding space $\mathbb{R}^d$, and then calculates similarity scores with the label embeddings of respective datasets.

dimensional label embeddings using the equation

$$[v_{ij}]_k = \phi\left(\mathcal{E}\left(\mathcal{F}\left(\mathrm{x}_u^{(i)}\right)\right), \mathrm{c}_k^{(j)}\right) \quad \forall k \in \{|\mathcal{Y}_j|\}, \tag{2}$$

where $\mathrm{x}_u^{(i)}$ is an image from the i^{th} unlabeled set, $\mathrm{c}_k^{(j)} \in \mathbb{R}^d$ is the label embedding corresponding to the k^{th} label from the j^{th} dataset and $[v_{ij}] \in \mathbb{R}^{|\mathcal{Y}_j|}$. When $i = j$, the scores correspond to the similarity scores within a dataset, and when $i \neq j$, they provide the cross dataset similarity scores. The label embeddings are just the prototype features calculated using the labeled data. They are pre computed and kept fixed over the course of training the network, since we found that the limited supervised data was not sufficient to jointly train a universal segmentation model as well as fine tune the label embeddings. More details on calculating label embeddings is presented in the supplementary section.

Unsupervised Loss We have two parts for the unsupervised entropy loss. The first part, the cross dataset entropy loss, is obtained by minimizing the entropy of the cross dataset similarity vectors.

$$\mathcal{L}_{US,c} = \mathcal{H}(\sigma([v_{12}])) + \mathcal{H}(\sigma([v_{21}])), \tag{3}$$

where $\mathcal{H}(.)$ is the entropy measure of a discrete distribution, $\sigma(.)$ is the softmax operator and the similarity vector $[v]$ is from Eq (2). Minimizing $\mathcal{L}_{US,c}$ makes the probability distribution *peaky* over a single label from a dataset, which helps in label side semantic transfer across datasets and hence improving the overall prediction certainty of the network. In addition, we also have a within dataset entropy loss given by

$$\mathcal{L}_{US,w} = \mathcal{H}(\sigma([v_{11}])) + \mathcal{H}(\sigma([v_{22}])) \tag{4}$$

which aligns the unlabeled examples within the same domain.

The total loss $\mathcal{L}_T$ is the sum of the supervised loss from Eq (1), and the unsupervised losses from Eq (3) and Eq (4), written as

$$\begin{aligned}\mathcal{L}_T = \quad & \mathcal{L}_S(\mathrm{X}_l^{(1)}, \mathrm{Y}^{(1)}, \mathrm{X}_l^{(2)}, \mathrm{Y}^{(2)}) + \alpha \cdot \mathcal{L}_{US,c}(\mathrm{X}_u^{(1)}, \mathrm{X}_u^{(2)}) \\ & + \beta \cdot \mathcal{L}_{US,w}(\mathrm{X}_u^{(1)}, \mathrm{X}_u^{(2)})\end{aligned} \tag{5}$$

where α and β are a hyper parameters that control the influence of the unsupervised loss in the total loss.

Inference For a query image $\mathrm{q}^{(k)}$ from dataset k during test time, the output $\hat{\mathrm{y}}^{(k)} = \mathcal{G}_k(\mathcal{F}(\mathrm{q}^{(k)}))$ gives us the segmentation map over the label set $\mathcal{Y}_k$ and the pixel wise label predictions. This adds no computation overhead or extra parameters to our approach during inference compared to existing deep semantic segmentation approaches. We note that although we calculate feature and label embeddings in our method and metric based inference schemes like nearest neighbor search might enable prediction in a label set agnostic manner, calculating pixel wise nearest neighbors using existing methods can prove very slow and costly for images with high resolution.

4. Experiments and Results

We provide the performance results of the proposed approach on a wide variety of real world datasets used in autonomous driving as well as indoor segmentation settings. We show the superiority of the our method over the existing baselines (Section 4.2), demonstrate improvement upon the state of the art semi-supervised approaches (Section 4.3), and also show the results on cross domain datasets (Section 4.4). Using only a fraction of the labeled data available, we show competitive results on these datasets.

4.1. Training Details

Datasets We show the results of our approach on large scale urban driving datasets from various domains like Cityscapes [12] (CS), CamVid [4] (CVD) and Indian Driving Dataset (IDD) [29, 64].

Cityscapes [12] is a standard autonomous driving dataset consisting of 2975 training images collected from various cities across Europe finely annotated with 19 classes. CamVid [4] dataset contains 367 training, 101 validation and 233 testing images taken from video sequences finely labeled with 32 classes, although we use the more popular 11 class version from [3]. We also demonstrate results on IDD [29, 64] dataset, which is an in-the-wild dataset for autonomous navigation in unconstrained environments. It consists of 6993 training and 981 validation images finely annotated with 26 classes collected from 182 drive sequences on Indian roads, taken in highly varying weather and environment conditions. This is a challenging driving dataset

Method	Road	SideWalk	Building	Wall	Fence	Pole	Traff. lt.	Traff. Sgn.	Veg.	Train	Sky	Person	Rider	Car	Truck	Bus	Train	MotorCyc.	Bicycle	mIoU
CS only	91.76	**54.78**	80.02	3.70	16.58	29.84	22.31	33.74	**83.88**	32.89	**82.07**	52.67	**21.57**	81.11	**19.01**	3.87	0.0	**19.64**	**49.01**	40.97
Univ-basic	87.00	44.54	77.77	**10.21**	11.07	25.54	14.51	25.82	80.72	22.40	78.19	49.00	19.64	75.35	1.86	0.25	**10.98**	8.83	41.08	36.04
Univ-full	**92.18**	51.29	**80.07**	0.0	**24.01**	**33.73**	**26.16**	**38.71**	82.30	**36.39**	81.61	**54.38**	20.48	**81.71**	2.37	**22.79**	3.85	1.31	46.23	**41.03**

Method	Sky	Buil.	Pole	Road	Pave.	Tree	Sign	Fence	Car	Ped.	Bicy.	mIoU
Camvid only	85.58	75.15	8.17	84.86	52.34	69.68	27.11	20.48	73.1	24.36	29.42	50.02
Univ-basic	**87.04**	76.67	9.56	83.5	51.35	70.07	27.75	22.6	**73.22**	33.94	35.25	51.9
Univ-full	86.3	**77.23**	**17.13**	**84.99**	**53.35**	**70.57**	**31.99**	**32.45**	72.94	**36.61**	**37.22**	**54.62**

Table 2: Class-wise IoU values for the 19 classes in Cityscapes dataset and 11 classes in the CamVid dataset with various ablations of universal semantic segmentation models, for N=100 on Resnet-18. Note the improvement of our method (Univ-full) for smaller classes like *pole* and *sign* on Cityscapes and CamVid datasets.

Method	N=50			N=100		
	CS	CVD	Avg.	CS	CVD	Avg.
Train on CS	33.33	32.92	33.13	40.97	36.52	38.75
Train on CVD	19.47	42.81	31.14	22.20	50.02	36.11
Univ-basic ($\mathcal{L}_s$)	32.82	48.56	40.69	36.04	51.90	43.97
Univ-cross (+ $\mathcal{L}_c$)	33.86	52.57	43.22	37.82	49.31	43.57
Univ-full (+ $\mathcal{L}_c, \mathcal{L}_w$)	34.01	53.23	**43.62**	41.03	54.62	**47.83**

Table 3: mIoU values for universal segmentation using Cityscapes (CS) and CamVid (CVD) datasets with a Resnet-18 backbone. N is the number of supervised examples available from each dataset. Bold entries have the highest *average mIoU* across the datasets.

since it contains images taken from largely unstructured environments.

While these autonomous driving datasets typically offer many challenges, there is still limited variation with respect to the classes, object orientation or camera angles. Therefore, we also use SUN RGB-D [59] dataset for indoor segmentation, which contains 5285 training images along with 5050 validation images finely annotated with 37 labels consisting of regular household objects like chair, table, desk, pillow etc. We report results on the 13 class version used in [21], and use only the RGB information for our universal training and ignore the depth information provided.

Architecture Although the proposed framework is readily applicable to any state-of-the art encoder-decoder semantic segmentation framework, we use the openly available PyTorch implementation of dilated residual network [69] owing to its low latency in autonomous driving applications. We take the embedding dimension d to be 128, and use dot product for the pixel level similarity metric $\phi(.)$ as it can be implemented as a 1×1 convolution on most of the modern deep learning packages. More details for each experimental setting is presented in the supplementary section.

Evaluation Metric We use the mean IoU (Intersection over Union) as the performance analysis metric. The IoU for each class is given by

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}}, \tag{6}$$

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.* [28]	58.80	-	-
Souly *et al.* [60]	-	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**

Table 4: Comparison of our approach with other semi-supervised approaches on the Resnet-101 backbone and CS+CVD dataset. Our approach (Univ-full) results in a single model across datasets unlike the previous semi-supervised approaches and deliver competitive performance on both the datasets.

where TP , FP , FN are the true positive, false positive and false negative pixels respectively, and mIoU is the mean of IoUs over all the classes. mIoUs are calculated separately for all datasets in a universal model. All the mIoU values reported are on the publicly available validation sets for the CS, IDD and SUN-RGB datasets, and on the test set for the CamVid dataset.

4.2. Ablation Studies

We perform the following ablation studies in our experiments to provide insights into the various components of the proposed objective function. (i) *Train on source*: We train a semantic segmentation network using *only* the limited training data available on one dataset, and provide results when tested on both the datasets. Since the label spaces do not directly overlap, we finetune a different classifier (decoder) for both the datasets and keep the feature extractor (encoder) as the same. (ii) *Univ-basic*: To study the effect of the unsupervised losses, we put $\alpha, \beta = 0$ and perform training using only the supervised loss term from Eq (1) and no entropy module at all. This is similar to plain joint training using the supervised data from each domain. (iii) *Univ-cross*: To study the effect of the cross dataset loss term from Eq (3), we conduct experiments by adding $\alpha = 1$ to the loss term. (iv) *Univ-full*: This is the proposed model, including all the supervised and unsupervised loss terms. We use $\alpha, \beta = 1$ in

Method	N=100 (Resnet-18)			N=1500 (Resnet-50)		
	CS	IDD	Avg.	CS	IDD	Avg.
Train on CS	40.97	14.64	27.81	64.23	32.50	48.37
Train on IDD	25.05	26.53	25.79	46.32	55.01	50.67
Univ-basic	37.94	25.21	31.58	63.55	53.21	58.38
Univ-full	36.48	27.45	**31.97**	64.12	55.14	**59.63**

Table 5: Universal segmentation results using IDD and CS datasets. Our approach (Univ-full) performs better across Resnet-18 and Resnet-50 CNN backbones.

the loss function in Eq (5). The best model is defined as the model having the highest *average mIoU* across the datasets.

Although many works on domain adaptation also provide results on Cityscapes dataset, we note that we cannot directly compare our result against them, since the problem setting is very different. While most of the domain adaptation approaches use large scale synthetic datasets as source dataset to improve performance on a specific target domain, we train our models on multiple resource constrained real world datasets directly.

Cityscapes + CamVid The results for training a universal model on Cityscapes and CamVid datasets is given in Table 3. For a setting of N=100 which corresponds to using 100 labeled examples from each domain, the proposed method gives the best mIoU value of 41.03% on Cityscapes and 54.62% on CamVid clearly outperforming the baseline approaches. Moreover, the universal segmentation method using the proposed unsupervised losses also performs better than using only supervised losses, which demonstrates the advantage of having unsupervised entropy regularization in domains with few labeled data and lots of unlabeled data.

Another observation from Table 3 is that for N=100, a model trained only on Cityscapes suffers a performance drop of 13.5% mIoU when tested on the CamVid dataset compared to a model trained on Camvid alone. Similarly, the performance drop in the case of Cityscapes is 18% mIoU for a model trained on Camvid. Therefore, it is evident that models trained on one dataset, like Cityscapes do not always perform well when deployed on a different dataset, like CamVid, due to domain shift and result in noisy predictions and poor output maps. This further brings out the necessity of training a single model which performs well on both the domains by using an entropy regularization based semantic transfer objective.

In the case of semantic segmentation datasets, very low values of N offers challenges like limited representation for many of the smaller labels, but we notice that the proposed model for N=50 still manages to perform consistently better on both the datasets.

Comparison of class-wise mIoUs of the universal segmentation approach for N=100 with CS+CamVid is given in Table 2. Entropy regularization clearly boosts performance in 9 of the 11 classes on the CamVid dataset, and for 10 out of 19 classes on the Cityscapes dataset. More importantly, it is the smaller classes like *pole*, *traffic sign*, *pedestrian* and *fence* which benefit greatly from universal training on both the Cityscapes and CamVid datasets, in spite of using only a small fraction of the labeled examples from these datasets.

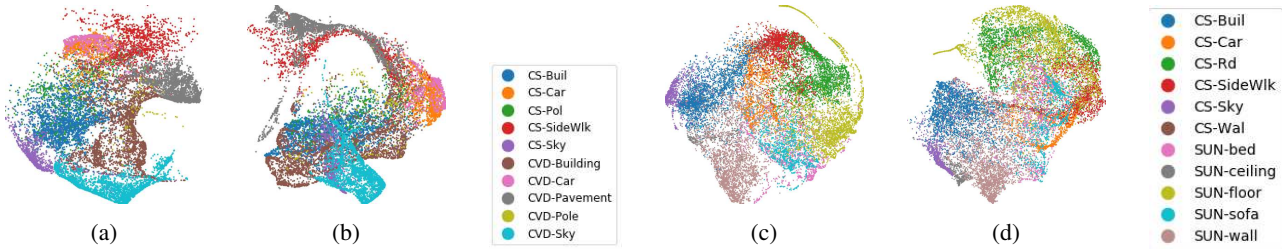
IDD + Cityscapes This combination is a chosen for validating the universal segmentation approach as the images are from widely dissimilar domains in terms of geography, weather conditions as well as traffic setup, and the datasets together capture the wide variety of road scenes one might encounter while training autonomous driving datasets for vision based navigation. The results for universal semantic segmentation using IDD and Cityscapes (CS) are shown in Table 5. Using 100 training examples from each domain, the proposed univ-full model gives an mIoU of 36.48% on Cityscapes (CS) and 27.45% on IDD using a Resnet-18 backbone, performing better than the univ-basic method on the average mIoU.

Similar to the CS+CamVid case, the features trained on Cityscapes dataset do not transfer directly to IDD, and shows a performance drop of 12% mIoU, demonstrating the necessity of learning universal representations for large scale datasets as well.

Furthermore, as an extreme case, we show the utility of the proposed approach even in the case of large number of labeled examples. We choose N=1500, which is a challenging setting since the number of supervised examples are already sufficient to train a joint model. However, from Table 5, the Resnet-50 based universal model still provides advantage over joint training method, which proves that adding unsupervised examples always helps the training, and more unsupervised examples can be added to these datasets to push the state of the art performance.

4.3. Comparison with state-of-the art

In addition to demonstrating the superiority of the proposed method over the baseline approaches, we also compare some of the existing semi supervised semantic segmentation works (which are targeted towards single dataset) with ours in Table 4, for similar amounts of labeled training data. Our model which uses dilated residual network gives competitive results on Cityscapes validation set when compared to [28] which uses a more complex DeepLab-V2 architecture. Similarly, without using any unsupervised video images unlike [60], we show superior results on the CamVid test set compared to them, in spite of the fact that our model is trained to perform well on multiple datasets at once. Most of the previous works optimize adversarial losses, and our results prove that entropy minimization is better suited for



(a) (b) (c) (d)

Figure 4: tSNE visualizations of the encoder output representations for majority classes from CS, CVD and SUN datasets. Plots (a) and (b) are for the Univ-basic and Univ-full model from CS-CVD datasets. Observe that the feature embeddings for large classes like *CS:Building-CVD:Building*, *CS:SideWalk-CVD:Pavement*, *CS:Sky-CVD:Sky* align a lot better with universal model. Plots (c) and (d) are for the Univ-basic and Univ-full model from CS-SUN datasets, and labels with similar visual features like *CS:Road - SUN:Floor* show better feature alignment. Best viewed in color and zoom.

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 [41]	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**

Table 6: mIoU values for universal segmentation across different task datasets with Resnet-50 backbone. While Cityscapes is an autonomous driving dataset, SUN dataset is mainly used for indoor segmentation. This demonstrates the effectiveness of universal segmentation even across diverse environments.

semi supervised approaches where limited supervision is available.

4.4. Cross Domain Experiment

A useful advantage of the universal segmentation model is its ability to perform knowledge transfer between datasets used in completely different settings, due to its effectiveness in exploiting useful visual relationships. We demonstrate this effect in the case of joint training between Cityscapes, which is a driving dataset with road scenes used for autonomous navigation and SUN RGB-D, which is an indoor segmentation dataset with household objects used for high-level scene understanding.

The label sets in Cityscapes and SUN-RGBD dataset are completely different (non overlapping), so the simple joint training techniques generally give poor results. However, from Table 6, our model outperforms the baselines and provides a good joint model across the domains making use of the unlabeled examples. We also compare our work against SceneNet [41], which uses large scale synthetic examples with RGB and depth data for pre-training, as well as all of the 5.3k available labeled examples for training. Using *only 28%* of the training data from the SUN-RGB dataset, and limited supervision from Cityscapes instead of synthetic examples, we achieve upto 88% of the mIoU reported in [41].

4.5. Feature Visualization

A more intuitive understanding of the feature alignment performed by our universal model is obtained from the tSNE embeddings [38] of the visual features. The pixel wise output of the encoder module is used to plot the tSNE of selected labels in Figure 4. For the universal training between CS and CVD in Figures 4a and 4b, we can observe that classes like *Building-CS* and *Building-CVD*, as well as *Sidewalk-CS* and *Pavement-CVD* align with each other better when trained using a universal segmentation objective. For the universal training between CS and SUN from Figure 4c and Figure 4d, labels with similar visual attributes such as *Road* and *Floor* align close to each other in spite of the label sets themselves being completely non overlapping.

5. Conclusion

In this work, we demonstrate a simple and effective way to perform universal semi-supervised semantic segmentation. We train a joint model using the few labeled examples and large amounts of unlabeled examples from each domain by an entropy regularization based semantic transfer objective. We show this approach to be useful in better alignment of the visual features corresponding to different domains. We demonstrate superior performance of the proposed approach when compared to supervised training or joint training based methods over a wide variety of segmentation datasets with varying degree of label overlap. We hope that our work would address the growing concern in the deep learning community over the difficulty involved in collection of large number of labeled examples for dense prediction tasks such as semantic segmentation. We also believe that other computer vision tasks like object detection and instance aware segmentation can benefit greatly from the ideas discussed in this work.

Acknowledgement M. Chandraker was supported by NSF CAREER 1751365.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, June 2013. 3

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015. 3

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 1, 3, 5

[4] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. 2, 5

[5] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[6] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[7] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 3

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 1, 3

[9] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018. 3

[10] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2011–2020. IEEE, 2017. 1, 3

[11] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008. 3

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 5

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 1

[14] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. corr abs/1310.1531 (2013), 2013. 3

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 3

[16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1180–1189, 2015. 3

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 3

[18] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41 – 65, 2018. 1

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[20] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 4

[21] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4077–4085, 2016. 6

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[24] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2, 3

[25] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 1, 3

[26] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in neural information processing systems*, pages 1495–1503, 2015. 3

[27] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018. 3

[28] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018. 2, 3, 4, 6, 7

[29] CV Jawahar, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandrakar, and Srikumar Ramalingam. AutoNUE workshop and challenge at ECCV'18. 5

[30] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017. 2, 3

[31] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, volume 2, page 8, 2017. 2, 3

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[33] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, March 2014. 3

[34] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2018. 3, 4

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 3

[36] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 4

[37] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems*, pages 165–177, 2017. 3

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[39] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018. 2, 3

[40] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[41] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687, 2017. 8

[42] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 1

[43] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 3

[44] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *International Conference on Learning Representations (ICLR)*, 2014. 3

[45] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 3

[46] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 3

[47] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013, 2018. 3

[48] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3

[50] E. Romera, J. M. lvarez, L. M. Bergasa, and R. Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, Jan 2018. 3

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[52] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2, 3

[53] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[54] Soham Saha, Girish Varma, and CV Jawahar. Class2Str: End to end latent hierarchy learning. In *International Conference on Pattern Recognition (ICPR)*, 2018. 3

[55] Michael L Seltzer and Jasha Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6965–6969. IEEE, 2013. 3

[56] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 3

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[58] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 2, 3

[59] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576. IEEE, 2015. 2, 6

[60] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, Oct 2017. 2, 3, 4, 6, 7

[61] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *arXiv preprint arXiv:1802.10349*, 2018. 1, 3

[62] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015. 3

[63] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[64] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019. 1, 2, 5

[65] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *arXiv preprint arXiv:1811.12833*, 2018. 4

[66] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 3

[67] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016. 3

[68] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[69] Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017. 2, 3, 6

[70] Amir R Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722, 2018. 2, 3

[71] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3

[72] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3