MLSL: Multi-Level Self-Supervised Learning for Domain Adaptation with Spatially Independent and Semantically Consistent Labeling

Javed Iqbal and Mohsen Ali
Information Technology University, Pakistan.
(javed.iqbal, mohsen.ali)@itu.edu.pk

Abstract

Most of the recent Deep Semantic Segmentation algorithms suffer from large generalization errors, even when powerful hierarchical representation models, based on convolutional neural networks, have been employed. This could be attributed to limited training data and large distribution gap in train and test domain datasets. In this paper, we propose a multi-level self-supervised learning model for domain adaptation of semantic segmentation. Exploiting the idea that an object (and most of the stuff given context) should be labeled consistently regardless of its location, we generate spatially independent and semantically consistent (SISC) pseudo-labels by segmenting multiple sub-images using base model and designing an aggregation strategy. Image level pseudo weak-labels, PWL, are computed to guide domain adaptation by capturing global context similarity in source and target domain at latent space level. Thus helping latent space learn the representation even when there are very few pixels belonging to the domain category (small object for example) compared to rest of the image. Our multi-level Self-supervised learning (MLSL) outperforms existing state-of-art (self or adversarial learning) algorithms. Specifically, keeping all setting similar and employing MLSL we obtain an mIoU gain of 5.1% *on GTA-V to Cityscapes adaptation and* 4.3% *on SYNTHIA to Cityscapes adaptation compared to existing state-of-art method.*

1. Introduction

With the evolution of deep learning methods during the last decade and the availability of densely labeled datasets [1–3], a considerable attention has been devoted to improving the performance of semantic segmentation [4–10]. Significant reliance of real-time applications like autonomous vehicles [11], bio-medical imaging [12], etc. over robust and accurate semantic segmentation step has also helped it gain prominence in current research. However, with the limited datasets for such a complex task (pixel-wise annotation), the state-of-the-art models have been reported to produce large generalization errors [13, 14]. This occurs naturally, because the train data may vary from test data (domain shift) in many aspects like illumination, visual appearance, camera quality, etc. It is time consuming and labor-intensive to densely label high resolution images covering all the domain variations. Modern computer graphics makes it easier to train deep models using synthetic images with computer generated dense labels [2,3]. However, these simulated-scene datasets are significantly different in visual appearance and object structures compared to real-life road-scene datasets, limiting the model performance. To overcome these domain shift issues, many techniques have been proposed to adapt the target data distribution [15–17]. Here our focus is to adapt the target domain dataset without labels in an unsupervised manner using self-supervised learning.

Due to large real-world applications, unsupervised domain adaptation (UDA) is a well-studied field in the current decade and aims to generalize to unseen data using only the labeled data of source domain. In UDA, most of the algorithms try to match the source and target data distribution using adversarial loss [18] either at structured output level [17] or latent space features level [19–21] respectively. Similarly, UDA based on adversarial learning augmented with other methods have recently produced good results on adaptation of semantic segmentation [14,22]. However, Zou et al. in [13] showed that a comparative performance can be achieved using an alternative method contrary to adversarial learning with less computational resources required compared to these complex methods. They introduced a class balanced self-supervised training method by generating pseudo-labels using the source-data trained model and tried to minimize a single loss function. However, they failed to capture the global context of the image referenced to categories and also the generated pseudo-labels had high uncertainty.

In this work, we propose a novel Multi-level Self-Supervised learning (MLSL) approach for UDA of semantic segmentation. The proposed approach consists of two complementary strategies. First, we propose *spatially independent and semantically consistent* (SISC) pseudo-labels generation process. We make reasonable assumption that an object should be segmented with same label regardless

of the location of the object. Same could be said about the stuff representing grass, road, sky, etc., given a reasonable context in surrounding. Using a base model, multiple sub-images (extracted from an image) are segmented independently and output probability volume is aggregated. This not only generates better pseudo-labels than single instance (SI) based ones, the assumption is more general than the spatial consistency assumption used by [13].

Secondly, we enforce the global context and small object information preservation while adaptation by attaching a category based image classification module at latent space level. For each target image, image level labels, called *pseudo weak-labels* (PWL) are generated using SISC pseudo-labels and size statistics collected from source domain. In summary, our main contributions are:

1. A Multi-level self learning strategy for UDA of semantic segmentation model by generating pseudo-labels at both fine-grain pixel-level and image level, helping identify domain invariant features at both latent and output level.
2. Designing a strategy, based on a reasonable assumption that for most categories, labels should be location invariant (given enough context) to generate *spatially independent and semantically consistent* pixel-wise pseudo-labels.
3. Using category wise size statistics to help build *pseudo weak-labels* (PWL) and train latent space.
4. State-of-the-art performance on benchmark datasets by augmenting the pseudo-labels with *class-wise spatial and image-level category distribution priors*.

2. Related Work

Due to the evolution of deep learning methods, most of the computer vision tasks including, but not limited to, object detection, semantic segmentation, etc., are shifted to deep neural networks based methods [23]. In [4], the authors proposed a fully convolutional network for pixel-level dense classification for the first time. Following them, many researchers proposed state-of-the-art methods for semantic segmentation taking the performance to an acceptable level for many computer vision tasks [5, 8, 9].

Domain adaptation is a widely studied area in computer vision for segmentation, detection, and classification tasks. With the emergence of semantic segmentation algorithms [4–6], availability of datasets [1–3] and modern applications demanding real-time constraints, e.g., self-driving cars, domain adaptation for semantic segmentation is in the spotlight. Many approaches exploited an appealing direction in semantic segmentation using domain adaptation from synthetic dataset to real-life datasets [17, 21, 24]. The underlying idea of UDA include matching target and source features using discrepancy minimization [20, 22], self-supervised learning with pseudo-labels [13, 25] and re-weighting source domain to look like target domain [16, 26]. This work thoroughly investigates the unsupervised domain adaptation for semantic segmentation with focus on self-supervised learning approach.

Adversarial learning is the most explored method for UDA of semantic segmentation [17, 19, 21]. Adversarial loss-based training is exploited for feature matching, structured output matching, and re-weighting processes frequently in UDA. The authors in [20] and [26] exploited latent space representations and used an adversarial loss to match the latent space features of source and target domains. Similarly, Chen et al. [19] used the adversarial loss for UDA of semantic segmentation augmented with class-specific adversaries to enhance the adaptation performance. The authors in [22] also proposed the latent space domain matching based on adversarial loss augmented with appearance adaptation network at the input. They tried to combine the latent space adaptation and re-weighting process and observed a significant gain in performance. In [16] the authors adapted similar approach to first transform the fully labeled source images to target images, train the segmentation model using the labeled source data, and then adapt further to target data. Rui et al. [27] devised a domain flow approach to transfer source images to new domains using adversarial learning at intermediate levels. In [28], the authors leveraged the spatial structure of source and target domain dataset, and working in latent space, proposed domain independent structure and domain specific texture based composite architecture for UDA. However, due to high dimensional representation at latent space, it is hard to adapt to new data distributions using adversarial loss because of the instability of the adversarial learning process.

In [17], the authors proposed a structured output space UDA approach based on adversarial learning. They avoided the curse of high dimensionality at latent space by exploiting the defined structure of road scene imagery at the output and provided a baseline solution along with state-of-the-art results in comparison to previous methods. The authors in [24] proposed a curriculum domain adaptation by addressing the easy examples first. They introduced a superpixels based-loss at the output space in conjunction with image level loss. Similarly, DADA [29] tried to exploit depth information for UDA of urban scene segmentation. Zou et al. [13] proposed a comparative performance method based on iterative learning. They proposed a class balanced self-training mechanism and obtained state-of-the-art performance using spatial priors in the pseudo-labels generation process. A tri-branch UDA model for semantic segmentation is proposed in [25], where they generate pseudo-labels from two branches and train the third branch on that pseudo-labels alternatively. The authors in [14] stated that, only adversarial learning at latent space or output space is not enough to learn the target distribution. They used a

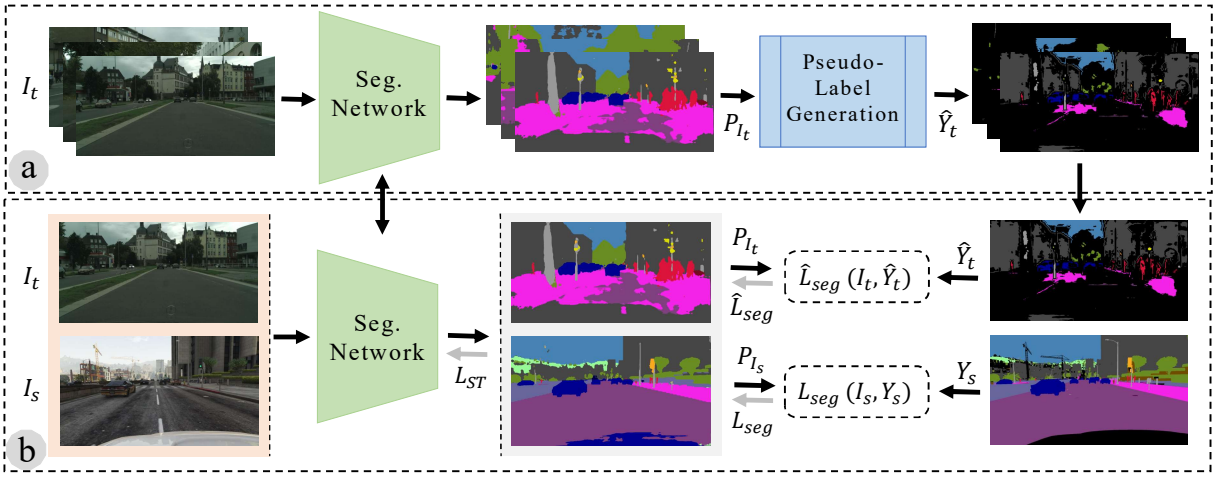

Figure 1. An illustration of the alternating self-supervised learning method for UDA of semantic segmentation. (a) shows pseudo-label generation and (b) shows segmentation network training on source and target images. (a) and (b) are repeated iteratively.

direct entropy minimization algorithm augmented with an entropy-based adversarial loss for UDA of semantic segmentation.

In summary, the existing solutions are suffering due to various problems e.g. latent space adaptation suffers from high dimensional feature representation, output space adaptation struggles with small and thin objects, re-weighting independently is not enough to achieve the goal. Similarly, the existing iterative methods are not capable to generate good pseudo-labels and cannot capture the global image context. In this work we propose category-based image classification using PWL and SISC based self-supervised learning for domain adaptation of semantic segmentation.

3. Approach

In this section, we present the proposed self-supervised and weakly-supervised learning approaches based on SISC pseudo-labels and PWL for domain adaptation of semantic segmentation. We start with existing state-of-the-art networks in semantic segmentation [30] and self-training for domain adaptation [13] as baseline methods and plugin additional modules for proposed approaches. Fig. 1 illustrates, iterative self-supervised learning technique for UDA.

3.1. Preliminaries

Let $I_s \in \mathbb{R}^{H\times W\times 3}$ and $Y_s \in \mathbb{R}^{H\times W\times C}$ where, I_s corresponds to RGB images of source dataset with resolution $H \times W$ and Y_s are ground truth labels as C-classes one-hot vectors with same spatial resolution as I_s. Let G be a fully convolutional network which predicts softmax outputs $G(I) = G(I^{H\times W\times 3}) = P_I^{H\times W\times C} = P_I$ for an input image I. One needs to learn the parameters w_g of G by minimizing the cross-entropy loss given in Eq. 1 on source domain images.

$$L_{seg}(I_s, Y_s) = -\sum_{H,W,C} Y_s^{H\times W\times C} log(P_{I_s}^{H\times W\times C}) \quad (1)$$

If ground truth labels for target dataset are available, the most direct strategy would be to use Eq. 1 and fine-tune the source trained model to target dataset. However, labels for target dataset are not always available, especially in case of real-time applications, e.g., self-driving cars. Therefore, an alternate way for unsupervised domain adaptation is to fine-tune the source trained model on the most confident outputs called "pseudo-labels", which the model produces on target domain images. The pseudo-labels have exactly the same dimensions as Y_s. The loss function for the target domain images is formulated as follows:

$$\hat{L}_{seg}(I_t, \hat{Y}_t) = -\sum_{H,W,C} d^{H\times W} \hat{Y}_t^{H\times W\times C} log(P_{I_t}^{H\times W\times C}) \quad (2)$$

where $\hat{L}_{seg}(I_t, \hat{Y}_t)$ in Eq. 2 is self-training loss with $\hat{Y}_t$ as the pseudo-labels one-hot vectors with C classes, and $d^{H\times W}$ is a binary map, obtained from pseudo-labels $\hat{Y}_t$ e.g., $d_{ij} = 1$ if any pseudo-label is there at $\hat{Y}_{t_{ij}}$, and $d_{ij} = 0$ if there is no pseudo-label assigned at $\hat{Y}_{t_{ij}}$, where $i = 1, ..., H$ and $j = 1, ..., W$. d allows to back propagate loss for those pixel locations only, which are assigned pseudo-labels. We name the training method as "self-supervised learning" or "self-training".

3.2. Semantically consistent pseudo-labels

Training a network using single inference (SI) generated pseudo-labels only, misleads the training process as there is no guarantee over the quality of pseudo-labels. An initial strategy is to jointly train the segmentation network using the ground truth labels of source images and the generated pseudo-labels of target images. The joint loss function is given by Eq. 3.

$$\min_{w_g} L_{ST}(I_s, Y_s, I_t, \hat{Y}_t) = L_{seg}(I_s, Y_s) + \hat{L}_{seg}(I_t, \hat{Y}_t) \quad (3)$$

where, $L_{seg}(I_s, Y_s)$ is the loss of source images and $\hat{L}_{seg}(I_t, \hat{Y}_t)$ is the loss of target images given in Eq. 1 and

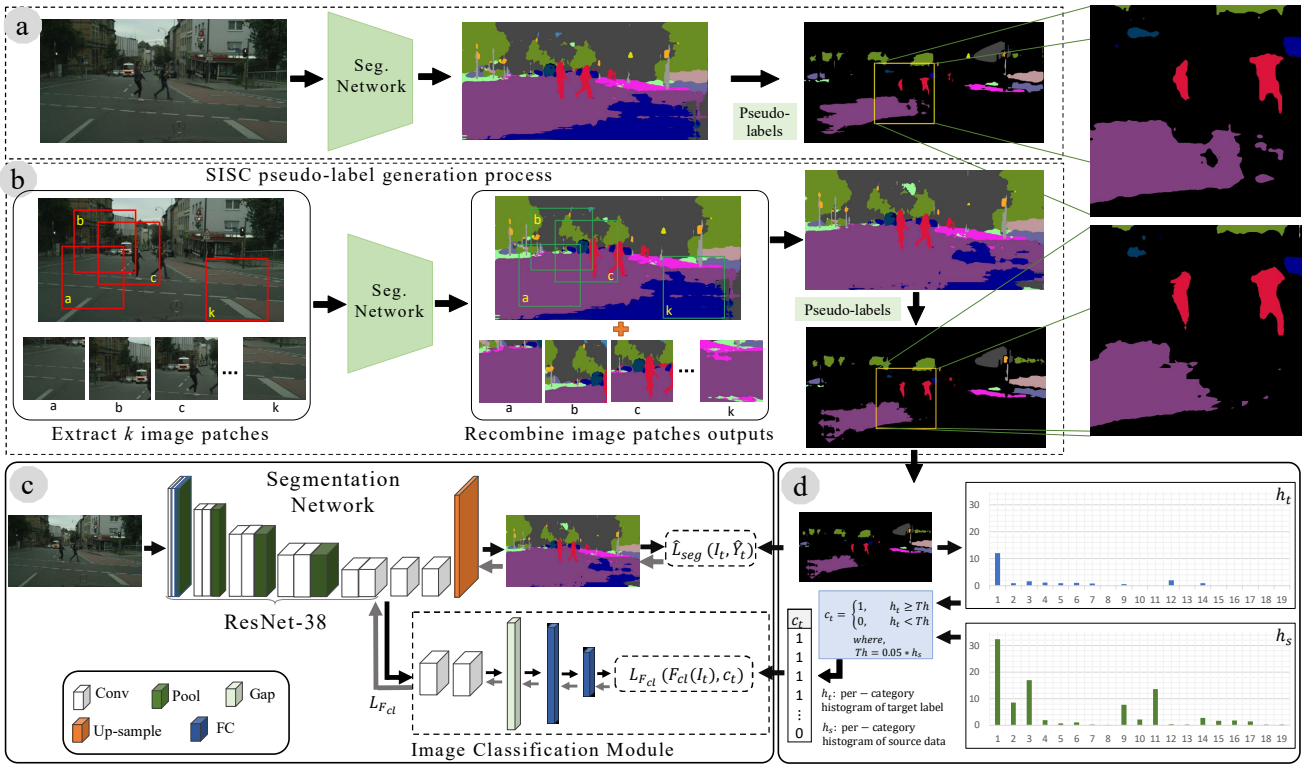

Figure 2. (a) Single-inference pseudo-label generation, (b) SISC pseudo-labels generation where, from left to right: patches are extracted randomly, segmented, recombined, normalized and pseudo-labels are generated. (c) shows the semantic segmentation and category-based image classification model, and (d) describes the PWL generation process.

Eq. 2 respectively. To minimize the loss in Eq. 3, we follow the two stage alternating process given below:

1. Generate pseudo-labels by fixing the parameters w_g.
2. Minimize the loss in Eq. 3 with respect to w_g by fixing the pseudo-labels $\hat{Y}_t$ generated in the previous step.

In this work, Step-1 and Step-2 are executed alternatively and repeated for multiple iterations. A work-flow of the proposed algorithm is shown in Fig. 1. Step-1 generate pseudo-labels using the output softmax probabilities of the target images based on the more confident examples. Once the pseudo-labels are generated, Step-2 updates the model parameters w_g using stochastic gradient descent (SGD) by minimizing the loss function given in Eq. 3.

Spatially independent and semantically consistent pseudo-labels: Instead of generating pseudo-labels using SI, (e.g., segmenting the whole image simultaneously), we generate "spatially independent and semantically consistent (SISC)" pseudo-labels. We leverage the spatial independence of our baseline semantic segmentation model to generate spatially independent and semantically consistent predictions. To quantitatively show the contribution of semantic consistency, we evaluate the softmax predictions based on different spatial context and select the most consistent ones. For each target image I_t, we select K partially overlapping patches $[p_1, p_2, ..., p_K]$ of size $h \times w$ each. Each patch p_i is passed through the segmentation algorithm to assign pixel-wise confidence vectors using softmax outputs. The output softmax probabilities for each patch are added to an empty matrix $P_{I_c} \in \mathbb{R}^{H \times W \times C}$ in specific locations where each patch belongs, and generate the composite output. Each pixel in P_{I_c} has an associated count based on the number of occurrences in different patches during inference. We normalize P_{I_c} with associated counts to obtain a normalized probability map and forward it to pseudo-label selection step which chooses the most confident outputs as pseudo-labels. The whole process of patch-based and single inference based pseudo-label generation is shown in Fig. 2.

Unlike simple pseudo-labels generation methods which suffer from category distribution imbalance problem, we use the category-balanced pseudo-label selection similar to the method used in [13]. Using the obtained normalized probability map, we further normalize the category-wise probabilities and select the pixels having high probability within a specific category. For example, we select all pixels locations which are assigned to be "road", normalize probabilities on that locations and then select the most confident ones. This process balances the inter-category pseudo-labels ratio and avoids the training process to adapt simple examples only. The obtained pseudo-labels belong to the more consistent pixels inferred without the global view. The loss function given in Eq. 3 is minimized using the origi-

nal labels for source domain and SISC pseudo-labels for the target domain.

3.3. Pseudo weak-labels guided domain adaptation

The cross-entropy loss for an input image/label pair defined in Eq. 1 calculates the sum of independent pixel-wise entropies, dealing with each pixel and label at the location independently. Thus ignoring any spatially global information, prone to effected by sparse erroneous pseudo-labels. Due to unbalanced pixels per category distribution, minimizing the summation of independent pixels entropies ignores the global data distribution. Even balancing the labels [13], the low-density classes fades (for target domain) as self-training proceeds.

We employ the pseudo weak-labels (PWL), guided multi-task weakly-supervised learning to regularize the pixel wise cross-entropy loss. The PWL based category level cross-entropy loss is attached at the encoder level while adapting. This forces the latent space to learn to represent target categories, even for the small objects whose latent space representation might be faded if only pixel-wise cross-entropy loss is used.

3.3.1 PWL Filtering

The pixel-wise pseudo-labels are too noisy to generate the image level pseudo-labels. Assuming that source and target have similar objects and their instances, we build a naive model for the category's size relationship with the image. From the source dataset we calculate $h_s = \{m_1, m_2, \ldots, m_c\}$, to represent mean size of each class, where

$$m_i = \frac{1}{(\sum_{j=1}^{N} 1_i^j) \times H \times W} \sum_{j=1}^{N} 1_i^j \{\sum_{x=1}^{H} \sum_{y=1}^{W} Y_s^j(x, y, i)\} \quad (4)$$

N stands for total images, and indicator function $\mathbf{1}_i^j$ is 1 if j^{th} image has class i, otherwise zero. For each target image I_t, we compute SISC pseudo-labels $\hat{Y}_t$ and use it to compute array h_t. PWL vector for image I_t is an indicator vector c^{pwl}, s.t. $c_i^{pwl} = 1$ if $h_t(i) > \eta h_s(i)$ otherwise zero. η is a small value chosen by the user. However, we are determined to make this process learn-able in future.

3.3.2 PWL Loss

Given any image I, an image classification module F_{cl} is designed to input the latent space representation (in this case of ResNet-38), and predict labels (Fig. 2(c)). Instead of softmax, we use sigmoid so that it can predict multiple labels for the image and use binary cross-entropy loss function given in Eq.5 .

$$L_{F_{cl}}(I, c) = -\frac{1}{C} \sum_{i=1}^{C} (c_i) log(F_{cl}(I)) + (1 - c_i) log(1 - F_{cl}(I)) \quad (5)$$

For the source images I_s, indicator vector c represents image level label crated from ground truth segmentation labels. For the images in target domain I_t, image level weak-labels c^{pwl} are created as detailed in Sec. 3.3.1.

3.4. Final Loss Function

The overall loss function for segmentation network and category-based image classification network for source domain is the composition of both 1 and 5, and is given by

$$L_{cmp}(I, Y, c) = L_{seg}(I, Y) + \lambda_{F_{cl}} L_{F_{cl}}(I, c) \quad (6)$$

where $\lambda_{F_{cl}}$ is the scaling factor and c is image level label. The combined loss function for self-supervised and weakly-supervised learning is given by;

$$L_{STWL}(I_s, Y_s, I_t, \hat{Y}_t, c) = L_{cmp}(I_s, Y_s, c) + \hat{L}_{cmp}(I_t, \hat{Y}_t, c) \quad (7)$$

Eq. 7, is minimized using criteria described in Sec. 3.2.

4. Experiments

In this section, we present experimental details and discuss the main results of our proposed UDA methods.

4.1. Experimental setup

4.1.1 Datasets

We follow the *synthetic-to-real* setup for UDA. We use GTA-V [3] and SYNTHIA [2] as our source domain synthetic datasets and Cityscapes [1] as real-world target domain dataset. GTA-V consist of 24966 synthetic frames of spatial resolution 1052×1914 extracted from a video game. All the 24966 frames have pixel level labels available for 33 categories, but we used 19 categories compatible with real-world Cityscapes dataset. Similarly, we use SYNTHIA-RAND-CITYSCAPES set having 9400 synthetic frames of size 760×1280 from SYNTHIA dataset. We train and evaluate our baseline and proposed models with 16 common classes in SYNTHIA and Cityscapes. We also report the 13 classes evaluation as described in [14] and [13].

In both the experiments, we use the Cityscapes training set without labels for unsupervised domain adaptation and evaluate the adapted models on Cityscapes separate validation set having 500 images. We use standard mean Intersection-over-Union (mIoU) as our evaluation metric.

4.1.2 Model architecture

We use ResNet-38 [30] as our baseline semantic segmentation model. The pre-trained ResNet-38 (trained on ImageNet [33]) is trained for semantic segmentation on GTA-V and SYNTHIA datasets. The ResNet-38 contains 7-blocks (convolutional, residual and pooling), followed by two segmentation layers and an upsampling layer. We also call the ResNet-38 as encoder for segmentation network and refer

Table 1. Semantic segmentation performance when the model trained on GTA-V dataset is adapted to Cityscapes dataset. We present the results of our proposed SISC pseudo-labels based self-supervised learning and PWL augmented self-training. We use the competitive baseline model and show a thorough comparison with existing state-of-the-art methods. The abbreviations "ST" and "Adv" indicates the self-training (self-supervised learning) and adversarial learning respectively.

		GTA-V → Cityscapes																			
Methods	Appr.	Road	Sidewalk	Building	Wall	Fence	Pole	T. Light	T. Sign	Veg.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.cycle	Bicycle	mIoU
ResNet-38 [30]	-	70.0	23.7	67.8	15.4	18.1	40.2	41.9	25.3	78.8	11.7	31.4	62.9	29.8	60.1	21.5	26.8	7.7	28.1	12.0	35.4
AdaptSetNet [17]	Adv	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
Saleh et al [31]	ST	79.8	29.3	77.8	24.2	21.6	6.9	23.5	44.2	80.5	**38.0**	76.2	52.7	22.2	83.0	32.3	41.3	27.0	19.3	27.7	42.5
MinEnt [14]	ST	86.2	18.6	80.3	27.2	24.0	23.4	33.5	24.7	**83.3**	31.0	75.6	54.6	25.6	85.2	30.0	10.9	0.1	21.3	37.1	42.3
DLOW [27]	Adv	87.1	33.5	80.5	24.5	13.2	29.8	29.5	26.6	82.6	26.7	81.8	55.9	25.3	78.0	33.5	38.7	0.0	22.9	34.5	42.3
CLAN [32]	Adv	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2
All Structure [28]	Adv	**91.5**	47.5	**82.5**	31.3	25.6	33.0	33.7	25.8	82.7	28.8	**82.7**	62.4	**30.8**	85.2	27.7	34.5	6.4	25.2	24.4	45.4
CBST-SP [13]	ST	88	**56.2**	77	**27.4**	22.4	**40.7**	**47.3**	40.9	82.4	21.6	60.3	50.2	20.4	83.8	35	**51**	15.2	20.6	37	46.2
Ours (SISC)	ST	91.0	49.3	79.9	24.4	**27.9**	37.9	45.1	**45.1**	81.3	19.0	61.7	**63.9**	28.0	**86.5**	23.9	42.3	**41.9**	33.1	44.4	48.7
Ours (SISC-PWL)	ST	89.0	45.2	78.2	22.9	27.3	37.4	46.1	43.8	82.9	18.6	61.2	60.4	26.7	85.4	**35.9**	44.9	36.4	**37.2**	**49.3**	**49.0**

its output as latent space representation. The two convolution layers comprises of 3×3 filters with depth of 512 and C (number of classes to segment). At the end, the upsampling layer up-scales the output using bi-linear interpolation.

Similarly, the image classification part discussed in Section 3.3 is a category (object/stuff) based image classification module augmented with ResNet-38. The image classification module consist of two convolution layers with filters $[1 \times 1, 3 \times 3]$ with depth 2048 each. A global average pooling (GAP) layer is applied to capture the global nature of the feature map channels. The output of GAP is passed through two fully connected layers of depth 512 and C respectively. Relu activation function is applied except the last layer where sigmoid is used.

4.1.3 Implementation and training details

We use MxNet [34] deep learning framework and a single Core-i5 machine with 32GB RAM and a GTX 1080 GPU with 8GB of memory to implement the proposed methods for domain adaptation of semantic segmentation. Our proposed model uses SGD optimizer for training with an initial learning rate of 1×10^{-4}. To generate SISC pseudo-labels, $K = 50$ is chosen (e.g. 50 sub-images of a target image are selected randomly). For SISC pseudo-labels based self-supervised learning, a batch size of 2 is chosen while the weakly-supervised setup described in section 3.3 processes a single image only. To optimize the joint loss function given in Eq. 7, the value of $\lambda_{F_{cl}}$ is investigated thoroughly (as shown in Section 4.3) and chosen as 0.025 to limit the image classification loss to back propagate large gradients. $\lambda_{F_{cl}}$ also controls the speed of adaptation with trade-off to segmentation performance, so the mentioned nominal value is used for all followed experiments. The iterative process of MLSL is repeated for 6 rounds where each rounds is composed of 2 epochs.

4.2. Experimental results

The experimental results of our proposed approaches compared to baseline ResNet-38 and existing state-of-the-art UDA methods are presented in this section. Our proposed approaches perform superior to other methods for domain adaptation and produce state-of-the-art results on two benchmark datasets. We also describe in detail, the behaviour of proposed approaches when exploited with different settings and different source datasets.

GTA-V to Cityscapes: Table 1 details the experimental results of 19 categories when adapted from GTA-V to Cityscapes. We use standard mIoU as semantic segmentation performance measure and report results on Cityscapes validation set. Our proposed approach of self-supervised learning with SISC pseudo-labels, shows state-of-the-art performance with ResNet-38 segmentation model. The SISC approach outperforms the latest approaches for UDA of semantic segmentation. Compared to MinEnt [14] which tries to minimize the self-entropy using direct entropy minimization, our SISC approach shows 13.1% improvement in overall mIoU. Similarly, compared to the self-training approach presented in [13], the proposed SISC method outperforms it with a margin of 5.1% in mIoU.

Our weak-labels guided UDA approach tries to capture the global image context by category (object/stuff) based image classification. This model helps improving the overall performance, and especially boost the performance for small and less occurring objects as shown in Table 1. The consistency and accuracy of pseudo weak-labels for image classification enable this approach to help the segmentation model for better performance. With ResNet-38 baseline, pseudo weak-labels when combined with CBST [13] provides 2.3% boost in mIoU compared to simple CBST. Similarly, when SISC is augmented with PWL based image classification, the mIoU performance increases by 5.7% from existing stat-of-the-art CBST-SP [13] as shown in Table 1. The ensemble of the two proposed approaches for UDA achieve 49.0 mIoU on Cityscapes validation set, which sets a new benchmark. The high boost in performance shows that both the approaches are capable to extract domain independent representations and produce better segmentation results comparatively.

Table 2. Semantic segmentation performance of Cityscapes validation set when adapted from SYNTHIA dataset. We present mIoU and mIoU* (13-categories) comparison with existing state-of-the-art methods for Cityscapes validation set.

SYNTHIA → Cityscapes																			
Methods	Appr.	Road	Sidewalk	Building	Wall	Fence	Pole	T. Light	T. Sign	Veg.	Sky	Person	Rider	Car	Bus	M.cycle	Bicycle	mIoU	mIoU*
ResNet-38 [30]	-	32.6	21.5	46.5	4.81	0.03	26.5	14.8	13.1	70.8	60.3	56.6	3.5	74.1	20.4	8.9	13.1	29.2	33.6
Road [21]	Adv	77.7	30.0	77.5	9.6	0.3	25.8	10.3	15.6	77.6	79.8	44.5	16.6	67.8	14.5	7.0	23.8	36.2	41.8
AdaptSetNet [17]	Adv	81.7	39.1	78.4	11.1	0.3	25.8	6.8	9.0	79.1	80.8	54.8	21.0	66.8	**34.7**	13.8	29.9	39.6	45.8
MinEnt [14]	ST	73.5	29.2	77.1	7.7	0.2	27.0	7.1	11.4	76.7	**82.1**	57.2	21.3	69.4	29.2	12.9	27.9	38.1	44.2
CLAN [32]	Adv	81.3	37.0	**80.1**	-	-	-	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	-	47.8
All Structure [28]	Adv	**91.7**	**53.5**	77.1	2.5	0.2	27.1	6.2	7.6	78.4	81.2	55.8	19.2	82.3	30.3	17.1	34.3	41.5	48.7
CBST [13]	ST	53.6	23.7	75.0	12.5	0.3	36.4	23.5	26.3	84.8	74.7	67.2	17.5	**84.5**	28.4	15.2	**55.8**	42.5	48.4
Ours (SISC)	ST	73.7	34.4	78.7	13.7	**2.9**	**36.6**	28.2	22.3	86.1	76.8	65.3	20.5	81.7	31.4	13.9	47.3	44.4	50.8
Ours (SISC+PWL)	ST	59.2	30.2	68.5	**22.9**	1.0	36.2	**32.7**	**28.3**	**86.2**	75.4	**68.6**	**27.7**	82.7	26.3	**24.3**	52.7	**45.2**	**51.0**

For a more fair comparison with other UDA methods, in Table 3, we show the mIoU gain with respect to specific baselines methods used. Compared to more complex models with very deep backbones, our approaches produces a higher gain of +13.6 points to source model surpassing the existing methods by a minimum margin of 20%. A comparison between upper-bound and our results is shown in supplementary document. Fig. 3 shows some examples of semantic segmentation before and after domain adaptation. As illustrated in the figure, the segmentation results improves significantly with SISC and SISC+PWL based approaches compared to source and CBST-SP methods.

Table 3. Performance (mIoU, mIoU*) gain comparison between the GTA-V and SYNTHIA trained source models and the respective adapted models from GTA-V and SYNTHIA to Cityscapes.

Dataset	GTA → Cityscapes			SYN → Cityscapes		
Methods	Source only	UDA Algo.	mIoU gain	Source only	UDA Algo.	mIoU* gain
FCN in the wild [15]	21.2	27.1	5.9	23.6	25.4	1.8
Curriculam DA [35]	22.3	28.9	6.6	28.4	34.82	6.42
AdaptSetNet [17]	36.6	42.4	5.8	38.6	46.7	8.1
MinEnt [14]	36.6	42.3	5.7	38.6	44.2	5.6
CLAN [32]	36.6	43.2	6.6	38.6	47.8	9.2
All Structure [28]	36.6	45.4	8.8	38.6	48.7	10.1
CBST [13]	35.4	46.2	10.8	33.6	48.4	14.8
Ours (SISC)	35.4	48.7	13.3	33.6	50.8	17.2
Ours (SISC+PWL)	35.4	**49**	**13.6**	33.6	**51.0**	**17.4**

SYNTHIA to Cityscapes: SYNTHIA is a more diverse dataset with multiple viewpoints and different spatial constraints compared to GTA-V and Cityscapes. In Table 2, we present the unsupervised adaptation results on Cityscapes validation set when adapted from SYNTHIA. The categories in SYNTHIA and Cityscapes do not fully overlap, so we have selected the common 16 classes as done in [13,15,24,35] for evaluation. We have also reported the performance (mIoU*) over the 13 common classes as used in [13, 17, 32]. With ResNet-38 as baseline netwrok, our proposed SISC based sef-supervised learning method performs superior to existing state-of-the-art methods as shown in Table 2. Compared to MinEnt [14] which uses similar entropy minimization technique, our SISC based UDA approach achieves 14.2% gain in mIoU and 13.3% gain in mIoU*. Similarly, compared to CBST presented in [13], our SISC based approach gains 4.3% and 4.7% points in mIoU and mIoU* respectively. Our proposed PWL guided UDA approach combined with SISC based self-supervised learning provides 6.0% and 5.1% boost in mIoU and mIoU* respectively when compared with CBST. Compared to an ensemble method (adversarial training and self-training) [14], our composite UDA method achieves 9.8% and 7.1% gain in mIoU and mIoU* respectively.

To make a more fair comparison with existing methods, Table 3 shows the baseline, after adaptation, and gain in terms of mIoU*. It is fair to say, that our proposed methods outperforms the existing state-of-the-art methods achieving the gain over baseline with a minimum margin of 16.3%. In Fig. 4, some examples of semantic segmentation before and after UDA are shown. As illustrated, the segmentation results improves significantly with SISC and SISC+PWL based approaches compared to source and CBST methods.

4.3. Ablation experiments

Relative frequency based pseudo-labels: Besides the adapted methodology in Section 3.2, we also generated pixel classification relative frequency based pseudo-labels. The randomly selected patches like SISC are segmented and recombined in the large output map. A count is made for each pixel with respect to assigned category in each patch, and then relative frequency is calculated. This relative frequency is used as prediction probability and incorporated in pseudo-labels generation. Due to hard decision, the pseudo-labels generated were not effective and lead to a decline in the performance.

Table 4. Influence of $\lambda_{F_{cl}}$ and η on overall performance.

GTA-V → Cityscapes				
$\lambda_{F_{cl}}$	0.1	0.05	0.025	0.001
SISC+PWL	46.0	48.1	49.0	48.24
η	0.0	0.1	0.05	0.025
SISC+PWL	45.5	46.0	49.0	47.33

Patch size selection: Our base models for semantic segmentation in both cases are trained on 500×500 random patches selected from the whole image randomly. Following that nominal size, we have chosen 512×512 as our patch size for pseudo-label generation. We also tried with 256×256 patch size but on high resolution Cityscapes images, these small image patches were not contributing. For

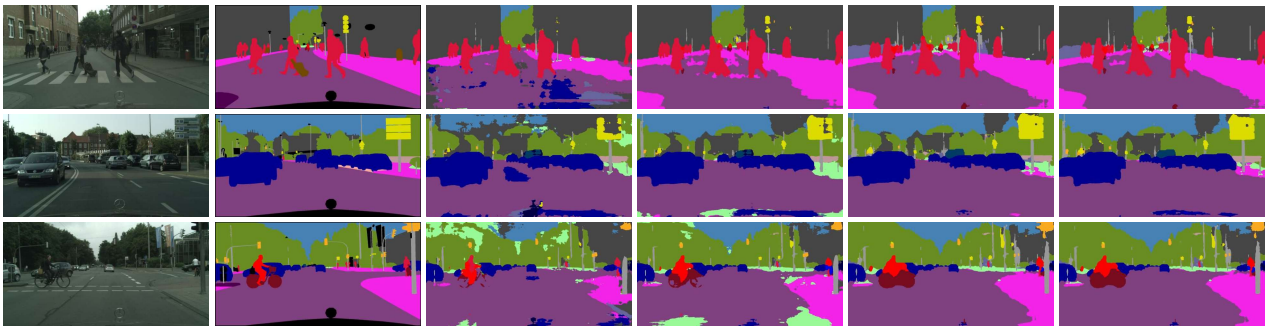

Figure 3. Segmentation results on Cityscapes validation set when adapted from GTA to Cityscapes.

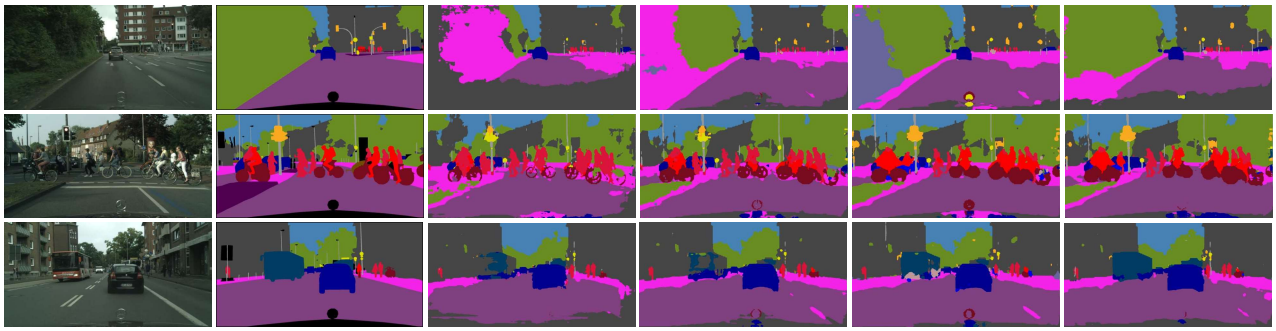

Figure 4. Segmentation results on Cityscapes validation set when adapted from SYNTHIA to Cityscapes.

patch size greater than 512×512 there were GPU memory limitations. Similarly, we selected 25, 50 and 100 patches per image randomly for SISC pseudo-labels generation. 25 patches were not enough to capture the high resolution Cityscapes images and 100 patches were taking the process very slow with negligible gain over 50 patches. Therefore, for all experiments, we have chosen 50 random patches per image. *Category based image classification loss weight:* Since image classification is added as a supporting module to segmentation network, the loss contribution by this module should also be limited. We tried multiple weight factors, and selected $\lambda_{F_{cl}} = 0.025$ (Table 4). *Pseudo-weak-label generation:* For category based image classification loss, the PWL are generated from segmentation pseudo-labels. Since it is difficult to set a minimum number of pixels limit for a category to be labeled as present in an image. Therefore, we exploited the category distribution of source datasets and assigned pseudo weak-labels to present categories based on source data distribution. For GTA-V to Cityscapes, we select a category to be labeled as present in an image if, it has more pixels compared to the 5% of mean category pixels of the same category in the source dataset. A detailed comparison along with respective mIoU is shown in Table 4.

5. Conclusions

In this paper, we have proposed, Multi-level self learning strategy (MLSL) for UDA of semantic segmentation by generating pseudo-labels at fine-grain pixel-level and image level, helping identify domain invariant features at both latent and output space. Using a reasonable assumption that labels of objects and stuff should be same regardless of their location, we generate Spatially independent but Semantically Consistent Labels. Image level labels, called pseudo weak-label (PWL) are generated and used as consistency check over SISC pseudo-labels. Pixel-wise object label distribution in the source domain images is used to regularize PWL. Binary cross-entropy loss using PWL enforces latent space to preserve the information about the objects, helping domain adapt for small objects. This multi-level pseudo-label generation for self-supervised learning, allows the network to learn domain-invariant features at different hierarchical levels. The rigorous experimentation demonstrates that the proposed SISC based self-supervised method alone outperforms the existing state-of-the-art algorithms on benchmark datasets: mIoU* improves from 46.2 to 48.7 and 48.4 to 50.8 on GTA-V & SYNTHIA to Cityscapes respectively. This includes both, ones using self-supervision or adversarial learning. Augmented with a PWL based image classification module, our proposed method further improves the performance, especially in the small objects. Effectiveness of SISC and PWL is highlighted by the substantial improvement of mean IOU over the base model, which is significantly more than previous state-of-methods.

References

[1] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5

[2] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 5

[3] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1, 2, 5

[4] 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, 2

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1, 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA*, 2015. 1, 2

[7] 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. 1

[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, 2

[9] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 1, 2

[10] Gabriela Csurka and Florent Perronnin. A simple high performance approach to semantic segmentation. In *BMVC*, pages 1–10, 2008. 1

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[12] 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

[13] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[14] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 1, 2, 5, 6, 7

[15] 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, 7

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1994–2003, 2018. 1, 2

[17] 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. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 6, 7

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[19] 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, 2

[20] Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Boosting domain adaptation by discovering latent domains. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[21] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 7

[22] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018. 1, 2

[23] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[24] 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. 2, 7

[25] Junting Zhang, Chen Liang, and C-C Jay Kuo. A fully convolutional tri-branch network (fctn) for domain adaptation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3001–3005. IEEE, 2018. 2

[26] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[27] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 6

[28] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 6, 7

[29] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. Dada: Depth-aware domain adaptation in semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[30] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 3, 5, 6, 7, 8

[31] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *European Conference on Computer Vision*, pages 86–103. Springer, 2018. 6

[32] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6, 7

[33] 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. 5

[34] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *LearningSys Workshop, NIPS 2015*, abs/1512.01274, 2015. 6

[35] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7