Learning from Scale-Invariant Examples for Domain Adaptation in Semantic Segmentation

M. Naseer Subhani and Mohsen Ali

Information Technology University, Pakistan
{msee16021,mohsen.ali}@itu.edu.pk

Abstract. Self-supervised learning approaches for unsupervised domain adaptation (UDA) of semantic segmentation models suffer from challenges of predicting and selecting reasonable good quality pseudo labels. In this paper, we propose a novel approach of exploiting *scale-invariance property* of the semantic segmentation model for self-supervised domain adaptation. Our algorithm is based on a reasonable assumption that, in general, regardless of the size of the object and stuff (given context) the semantic labeling should be unchanged. We show that this constraint is violated over the images of the target domain, and hence could be used to transfer labels in-between differently scaled patches. Specifically, we show that semantic segmentation model produces output with high entropy when presented with scaled-up patches of target domain, in comparison to when presented original size images. These scale-invariant examples are extracted from the most confident images of the target domain. Dynamic class specific entropy thresholding mechanism is presented to filter out unreliable pseudo-labels. Furthermore, we also incorporate the focal loss to tackle the problem of class imbalance in self-supervised learning. Extensive experiments have been performed, and results indicate that exploiting the scale-invariant labeling, we outperform existing self-supervised based state-of-the-art domain adaptation methods. Specifically, we achieve 1.3% and 3.8% of lead for GTA5 to Cityscapes and SYNTHIA to Cityscapes with VGG16-FCN8 baseline network.

1 Introduction

Deep learning based semantic segmentation models [29, 3, 32, 31] have made considerable progress in last few years. Exploiting hierarchical representation, these models report state-of-the-art results over the large datasets. However, these models do not generalize well; when presented with out of domain images, their accuracies drops. This behavior is attributed to the shift between the source domain, one over which model has been trained, and target, over which its being tested. Most of semantic segmentation algorithms are trained in a supervised fashion, requiring pixel-level, labor extensive and costly annotations. Collecting such fine-grain annotations for every scene variation is not feasible. To avoid this pain-sticking task, road scene segmentation algorithm use synthetic but photo-realistic datasets, like GTA5 [20], Synthia [21], etc., for training. However, they

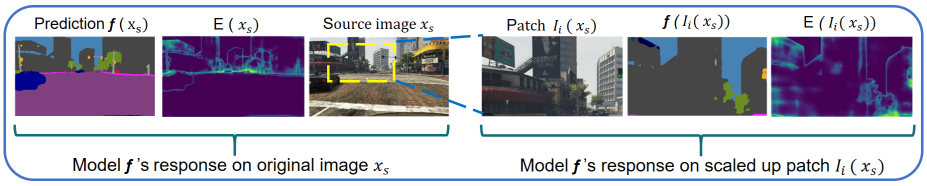

Fig. 1. Scale-invraince property of semantic segmentation model Original image and patch extracted from it and resized, are assigned same semantic labels by the model $\boldsymbol{f}$ at the corresponding locations. *Left:* An image x_s from the source domain, labels assigned to it by model $\boldsymbol{f}$. x_s belongs to the source domain. Self-entropy map E shows small values. Yellow box on x_s indicate patch location *Right:* Extracted patch resized to original image size. Assigned labels are similar to ones of original and self-entropy is similar that of original image.

are evaluated on the real datasets like Cityscapes [6], thus amplifying the domain shift.

Over the years, many unsupervised domain adaptation (UDA) methods have been proposed to overcome the domain shift, employing adversarial learning [4, 8, 22, 33], self-supervised learning [32, 34, 12], etc. or their combination. Where adversarial learning methods are dependent upon how good (input, feature or output) translation could be performed, self-supervised learning methods have to deal with challenges of generating so-called *good quality* pseudo-labels and selection of confident images for the learning from the domain.

In this paper we propose a novel method of generating pseudo-labels for self-supervised adaptation for semantic segmentation, by exploiting scale-invariance property of the model. Our proposed solution is based on an assumption that regardless of the size of an object in the image, the model's prediction should not be change, as shown in Fig. 1. To support our algorithm, we introduce three other novel components to be incorporated in the self-supervised method. A *class-based sorting mechanism* image selection process to identify images that should be used for the self-learning. To filter out pixels with non-confident pseudo-labels from learning process, we design an automatic process of estimating *class specific dynamic entropy-threshold* allowing "easy" classes to have tighter threshold than the ones that are "difficult" to adapt. To further reduce the effect of class imbalance over adaptation process, we also incorporate the focal loss [16] in our loss. Below we define the concept of scale-invariance.

Scale-invariance: In general one can assume that depending on the camera location, pose and other parameters, objects in images will appear at varying sizes. In the road scene imagery, such as GTA5, Cityscapes, etc., due to movement of the vehicle and dynamic nature of environment, objects and other scene elements (like road, building) appear at multiple scales. These variations are readily visible in Fig. 2. Its reasonable to assume that the semantic segmentation model trained on such dataset that will assign objects and stuff with same semantic labels regardless of their size. This could be seen in Fig. 1, where when

Fig. 2. Objects and scene-elements exhibit the scale variations naturally in road scene images, as shown in the frames sampled from Cityscapes [6] and GTA5[20] datasets. As the vehicle moves, near by objects and other scene elements might become afar or vice-versa, resulting in scale changes. Matching color boxes highlight changing size of cars, buildings, and other regions as vehicle moves.

an image and a resized patch extracted from same image are presented to segmentation model we get similar semantic labels at (almost all) corresponding regions. For both, image and resized patch, self-entropy is also indicating that the decision was made with low uncertainty. Semantic segmentation model, when presented with an image, from the out of source but somewhat visually similar domain, and the patches extracted from that image, we see considerable difference between the labels assigned for patches and ones assigned to corresponding areas of original image. Comparative increase in the self-entropy indicates that labels assigned to patches are not reliable . In this work, as shown in Fig. 3, we propose to use semantic labels assigned to the image to create pseudo-labels of corresponding patches. Our objective is to preserve the scale-invariance property of the semantic segmentation model and use it to direct our adaptation process.

We summarize our contribution as bellow.

- We propose a novel approach of exploiting scale-invarince property of the model to generate pseudo-labels for the self-supervised domain adaptation of semantic segmentation model.
- Class specific dynamic entropy thresholding is introduced so that pixels belonging to classes at different adaptation stage could be judged differently when being made included in the loss function.
- To eliminate the effect of the class imbalance problem, we incorporate the focal loss to boost the performance of smaller classes. And Class-based target image sorting algorithm is proposed so that selected images have equal representation of all the classes.

Although, part of our algorithm is generic, we show our results on the adaptation from synthetic to real road scene segmentation. We report state-of-the-art results over the GTA to Cityscapes and Synthia to Cityscapes for the self-

supervised based domain adaptation algorithms. VGG16 [24] and ResNet101 [9] are used as our baseline architectures.

2 Related works

Semantic Segmentation: There is an intensive amount of research has been done in semantic segmentation due to its importance in the field of computer vision. State of the art methods in semantic segmentation have gained huge success for their contribution. Recently, many researchers have proposed algorithm for semantic segmentation such as DRN (Dilated Residual Network) [29], DeepLab [3] etc. [1, 32, 28]. [29] have proposed a dilated convolution neural network in semantic segmentation to increase the depth resolution of the model without effecting its receptive field. In this work, we have utilized FCN8s[17] with VGG16[24] and DeepLab [2] with ResNet101 [9] as our baseline architectures of semantic segmentation.

Domain Adaptation: Domain adaptation is a popular research area in computer vision, especially in classification and detection problems. The goal of domain adaptation is to minimize the distribution gap between source and target domain. Many of the algorithms have already developed for domain adaptation like [34, 27, 23, 10, 26, 30, 11, 33, 12]. In this paper, we are focused in self-supervised domain adaptation to tackle the problem of domain diversity. Previous methods have been applied Maximum Mean Discrepancy (MMA) [19] to minimize the distribution difference. Recently, there has been an enormous interest in developing domain adaptation methods with the help of unsupervised and semi-supervised learning.

Adversarial Domain Adaptation in Semantic Segmentation: Adversarial training for unsupervised domain adaptation is the most explored approach for semantic segmentation. [11] are the first ones to introduce domain adaptation in semantic segmentation. [27] have proposed an entropy minimization, based on domain adaptation in which they have minimized the self-entropy with the help of adversarial learning. In [26], they have applied adversarial learning at the output space to minimize the distribution at the pixel level between the source and the target domain. [5] presents Reality-Oriented-Adaptation-Network (ROAD) to learn invariant features of source and target domain by target guided distillation and spatial-aware adaptation. [18] has also introduced a categorical-level adversarial network (CLAN) in which they have aligned the features of each class by adaptive adjusting the weight on adversarial loss specific to each class. There are other methods with the generative part for adversarial training in semantic segmentation. In generative methods, they are trying to generate the target images with a condition of the source domain. [33] have proposed a pixel level adaptation to generate image similar in visual perception with target distribution. In [10], they have used pixel level and feature level adaptation to overcome the distribution gap between the source and the target domain. They

incorporate cycle consistency loss to generate the target image condition on the source domain. They have also utilized the feature space adaptation and generate target images from the source features and vice-versa.

Self-Supervised Domain Adaptation in Semantic Segmentation: The idea behind self-supervised learning is to adapt the model by the pseudo labels generated for unlabeled data from the previous state of the model. [14] proposed a method of self-supervised learning from the assembling of the output from different models and latter train the model by generating pseudo labels of unlabeled data. [25] developed an algorithm based on a teacher network where the model is adapted by averaging the different weights for better performance on the target domain. Recently, self-supervised learning has also gained popularity in the semantic segmentation task. [34] proposed a class-balanced-self-training (CBST) for domain adaptation by generating class-balanced pseudo-labels from images which were assigned labels with most confidence by last state of model. To help guide the adaptation, spatial priors were incorporated. [7] have also contributed their research in self-supervised learning by generating pseudo labels with a progressive reliable strategy. They have excluded less confident classes with a constant threshold and have trained the model on generated pseudo labels. In this research, we filter out the less confident classes by applying a dynamic threshold that is calculated for each class separately during the training process. [15] have proposed a self-motivated pyramid curriculum domain adaptation (PyCDA) for semantic segmentation. They have included the curriculum domain adaptation by constructing the pyramid of pixel squares at different sizes, which has included the image itself. The model trained on these pyramids of the pixel by capturing local information at different scales. Iqbal and Ali [12]'s spatially independent and semantically consistent (SISC) pseudo-generation method could be closest to our work. However, they only explore the spatial invariance by creating multiple translated versions of same image. Since they don't have knowledge of which version has results in better inference they aggregate inference probabilities from all to create a single version, leading to smoothed out pseudo-labels. We on the other hand, define a relationship between the scale of the image and the self-entropy; therefore instead of aggregating we use the inference for image of original scale to create pseudo-labels for the up-scaled patch extracted from same image. Along with it, we present a comprehensive strategy of overcoming class imbalance and selecting the reliable psuedo-labels.

3 Methodology

In this section, we briefly describe our propose domain adaptation algorithm by learning from self-generated scale-invariant examples for semantic segmentation. In this work, we assume that the predictions of these confident images on target data are the approximation of their actual labels.

3.1 Preliminaries

Let X_S be set of images belonging to the source domain, such that for each image $x_s \in \mathbb{R}^{H \times W \times 3}$, in the source domain we have respective ground-truth one-hot encoded matrix $y_s \in \mathbb{R}^{H \times W \times C}$. Where C is the number of classes and $H \times W$ is the spatial size of the image. Similarly, let X_T be set of images belonging target domain. We train a fully convolution neural network, $\boldsymbol{f}$, in a supervised fashion over the source domain for the task of semantic segmentation. Let $P = \boldsymbol{f}(x)$ be soft-max output volume of size $H \times W \times C$, representing predicted semantic class probabilities for each pixel. The segmentation loss for any image x with the given ground-truth labels y and predicted probabilities P is given by

$$\mathcal{L}_{seg}(x, y) = -\sum_{h,w,c}^{H,W,C} y^{h,w,c} log(P^{h,w,c}) \tag{1}$$

In later cases to increase readability we just use h, w, c with summation sign, to indicate the summation over total height, width and channels. Source model $\boldsymbol{f}$ has been trained by minimizing $\mathcal{L}^S_{seg} = \sum_s^S \mathcal{L}_{seg}(x_s, y_s)$.

For target domain, since we do not have ground-truth labels, self-supervised learning method requires us to generate *pseudo-labels*. Let $x_t \in X_T$ be an image in the target domain, $P_t = \boldsymbol{f}(x_t)$ be output probability volume, one hot encoded pseudo-labels $\hat{y}_t$ could be generated by assigning label at each pixel to the class with maximum predicted probability. Since, source model is not accurate on the target domain, therefore a binary map $F_t \in \mathbb{B}^{H \times W}$ is defined to select the pixels whose prediction loss has to be back-propagated.

$$\mathcal{L}_{seg}(x_t, \hat{y}_t) = -\sum_{h,w,c}^{H,W,C} F_t^{h,w} \hat{y}_t^{h,w,c} log(P_t^{h,w,c}) \tag{2}$$

In general, for self-supervised learning, we minimize the loss in Eq. 2 over the selected subset of images from the target domain.

3.2 Class-Based Sorting for Target Subset Selection

To train the model with self-supervised learning, we need to extract the pseudo-labels which are reliable. A binary filter defined in Eq. 2, helps select pixels with so-called ǵoodṕseudo-labels, however, does not give us global view of how good are predictions in the whole image. Calculating an average of maximum probability per location of $\hat{y}_t$ can help us define the confidence of predictions on the x_t, for readability we call it *confidence of image* x_t. A subset selected on the base of the above defined confidence can lead to a class-imbalance with more images with pseudo-labels belonging to large and frequently appearing classes. That in turn leads to adaptation failing for the smaller objects or infrequent classes. We design a class based image subset selection process from the target domain (Algo. 1) to mitigate this effect.

Algorithm 1: Class-Based Sorting

Input : Model $f(\mathbf{w})$, Target data X_{t}, portion p
Output: Confident images X'_{t} of target domain , Entropy threshold h_{c}

for $t = 1$ *to* T **do**
  $P_{x_t} = \boldsymbol{f}\,(w, x_t)$
  $M_{P_{x_t}} = max\,(P_{x_t},\, axis = 0)$
  $A_{P_{x_t}} = argmax\,(P_{x_t},\, axis = 0)$
  for $c = 1$ *to* C **do**
    $M_{P_{x_t,c}} = M_{P_{x_t}}[\,A_{P_{x_t}} == c\,]$
    $U_c = [\,U_c,\ mean\,(M_{P_{x_t,c}})\,]$
    $X_{t,c} = [\,X_{t,c},\, x_t\,]$
  end
end
for $c = 1$ *to* C **do**
  $X_{t,c,sort} = sort\,(X_{t,c}\ w.r.t\ U_c,\, descending\ order)$
  $len_{th} = length\,(X_{t,c,sort}) \times (p/C) \quad \rightarrow (p/C)$is the portion of class c
  $X^{'}{}_{\mathrm{t}} = [\,X^{'}{}_{\mathrm{t}},\, X_{t,c,sort}\,[\,0 : len_{th}\ -\ 1\,]\,]$

  Calculate h_{c} for each class
  $x_l = X_{t,c,sort}[\,len_{th}\ -\ 1\,]$
  $P_{x_l} = \boldsymbol{f}\,(w, x_l)$
  $A_{P_{x_l}} = argmax\,(P_{x_l},\, axis = 0)$
  $E_{P_{x_l}} = entropy\,(P_{x_l}) \quad \rightarrow$ normalized to $[0, 1]$
  $h_c = mean\,(\ E_{P_{x_l}}[\,A_{P_{x_l}} == c\,])$
end
return $X^{'}{}_{\mathrm{t}}, h_{\mathrm{c}}$

Instead of calculating confidence for each image globally, we calculate the *confidence* with respect to each class c for every image in target data X_{T}. For each class, X_T is sorted with respect to the class specific confidence U_c and a subset, of size p, is selected. Union of these subsets form our confident target images subset $X_t^{'}$, note that repeated entries are removed. The algorithm of class-based sorting shown in Algorithm 1. For $X_t^{'}$ the model prediction are relatively of more confidence than rest of the set and can be utilized to adapt the model by self-supervised learning.

3.3 Dynamic entropy threshold for class dependent filter selection

The class based sorting takes in consideration all the pixels and does not make distinction between pixel-wise reliable and unreliable predictions. We define reliable or good predictions as by how low is the self-entropy of the prediction. If the entropy is low the prediction is more confident, if its high it means that the model is undecided which semantic label should be assigned to the pixel. Let, $P(x_t^{'}) = \boldsymbol{f}(x_t^{'})$ be the predicted probability volume, and $E_{x_t^{'}} = -\sum_c P_c(x_t^{'})\ log(P_c(x_t^{'}))$

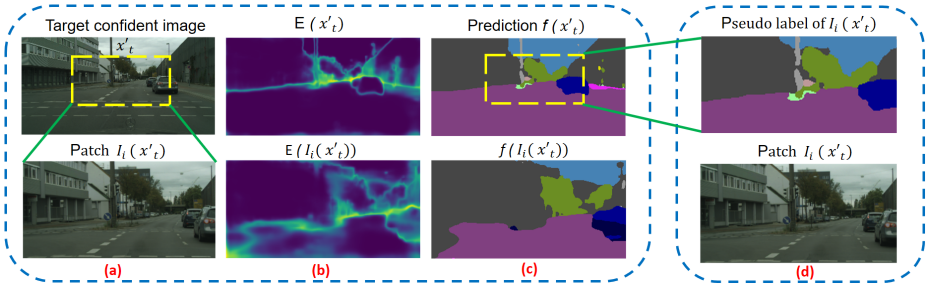

Fig. 3. Exploiting Scale-Invariance property for generated pseudo labels: For an image x^t belonging to target domain and its zoomed-in version scale-invariance property is violated. (a) Image x^t and its extracted patch $\boldsymbol{I}_\text{i}$. (b) High self-entorpy values computed from the output probabilities indicate source model $\boldsymbol{f}$ is not confident about the labels assigned to resized patch. (c) comparison of the labels indicate violation of scale-invariance property (d) Since original image exhibit low self-entropy we can use predictions over it as the pseudo-labels for the resized patch.

be entropy computed at each location. A binary filter map $F_{x'_t}$ is generated by thresholding the entropy at every location, by a class specific threshold.

$$F_{x'_t}(h,w) = \begin{cases} 1 \; E_{x'_t}(h,w) \leq h_c \;\; ; \;\; \text{where } c = argmax(P(x'_t)(h,w)) \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

Instead of being h_c a global and constant hyper-parameter, h_c is different for every class and depends upon predicted probabilities pixels belonging to that class in the selected confident set X'_t. As the adaptation for that class improves the filter selection for that class becomes more tighter (Algo. 1).

3.4 Self-Generated Scale-Invariant Examples

Based on a reasonable assumption that a source domain consists of images with scene elements and objects of same class appearing in scale variations, we claim that model trained on such dataset should label same object with same semantic label regardless of its size in the image. We define this as *scale-invariance property* of the model. As shown in Fig. 3 such a property is violated when target domain images are presented to the source model and could be used to guide the domain adaptation process. Specifically, lets assume $x'_t \in X'_\text{t}$ be the one of the selected images, $F_{x'_t}$ be the binary mask, and $P(x'_t) = \boldsymbol{f}(x'_t)$ is the output probability volume. Let $\boldsymbol{R}(x'_t, rec_i)$ be the operation applied on x'_t to extract i^{th} patch from location $rec_i = (r_i, c_i, w_i, h_i)$ and resized to spatial size of $H \times W$. Then we can define, $I^t_i = \boldsymbol{R}(x'_t, rec_i)$, $F^i_{x'_t} = \boldsymbol{R}(F_{x'_t}, rec_i)$ and $P^i_{x'_t} = \boldsymbol{R}(P_{x'_t}, rec_i)$ be the corresponding extracted and resized versions. We compute $\hat{y}^i_t$ is the one-

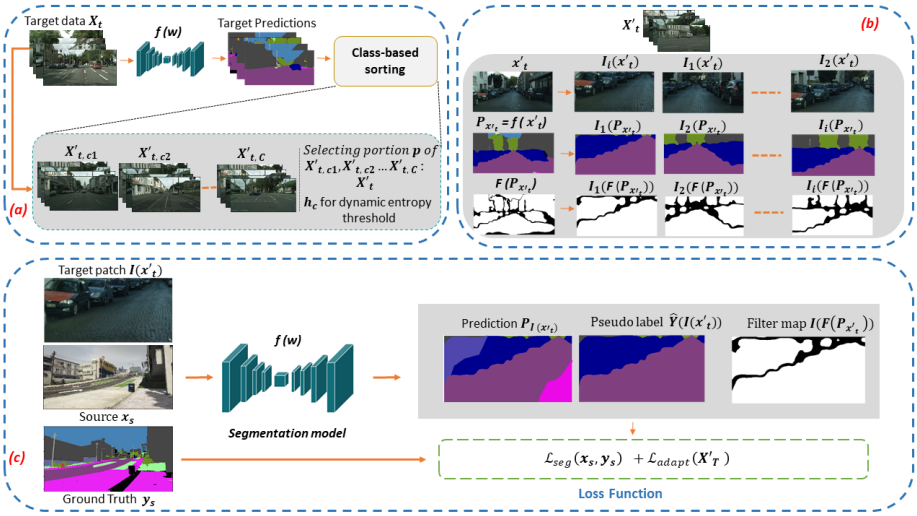

Fig. 4. Algorithm Overview: Our algorithm consists of three main steps. (a) First, we have calculated the confidence of each target images X_t with reference to each class c. We have sort out these images $X'_{t,c}$ of each class c in descending order on the basis of their confidence value. After that, we have selected the top portion from these sorted images $X'_{t,c}$ to form confident target data X'_t. (b) Second, we have extracted the random patches I_i from each confident images x'_t of target domain X_t. These patches are the scale-invariant with full-sized image. The model performs inconsistent on these patches and predict an output with high entropy prediction. To filter out the less confident pixels we have generated a filter map for each confident images x'_t by calculating their entropy with the help of threshold h_c for each class c. (c) Third, we have trained the model by given loss function on these scale-invariant examples with their pseudo labels that are generated from the previous state of the model.

hot encoded pseudo labels created from $P^i_{x'_t}$. Then loss for violating the scale invariance could be computed by Eq. 4.

$$\mathcal{L}_{seg}(I^t_i, \hat{y}^i_t) = -\sum_{h,w,c}^{H,W,C} F^{i,h,w}_t \hat{y}^{i,h,w,c}_t log(\boldsymbol{f}(I^t_i)^{h,w,c}). \tag{4}$$

3.5 Leveraging Focal Loss for Class-Imbalance

Self-supervised approach for domain adaptation highly dependent on information represented in selected confident images of the target domain. Biased distribution, i.e. number of pixels per class, in the road scenes creates a class imbalance problem. Even after the class based sorting (Sec. 3.2) and class dependent entropy thresholding, classes with high volume of pixels in target dataset (such as road, building, vegetation, etc.) end up having more contribution towards loss

function. Classes which appear infrequently and/or have less volume of pixels per image will contribute less and hence adaptation will be slow. To eliminate the effect of class imbalance problem, we incorporate the focal loss [16], so that cross-entropy function of each pixel is weighted by the based on pixel confidence. Focal loss balanced the loss for each pixel based on their confident level. This approach of applying focal loss balance the learning process of self-supervised learning equally to each class. In this work, we apply focal loss during the training of scale-invariant examples. Eq. 5 shows the formulation of focal loss.

$$\mathcal{L}_{FL}(I_i^t, \hat{y}_t^i) = -\sum_{h,w,c}^{H,W,C} \hat{y}_t^{i,h,w,c} \; \boldsymbol{log}(\boldsymbol{f}(I_i^t)^{h,w,c})(1-\boldsymbol{f}(I_i^t)^{h,w,c})^{\gamma} \tag{5}$$

Where γ is the hyperparameter that controls the focus and generally have value between 0 to 5. Low value bring it closer cross-entropy and high value focusing only on the hard examples. We set γ to middle value,3.

3.6 Adaptation

During adaptation, for each round r, we perform class based sorting of target dataset to create subset $X_T^{'}$. For each $x_t^{'} \in X_T^{'}$, k patches are extracted. Out total loss is defined as

$$\mathcal{L}_{LSE} = \sum_{x_s \in X_S} \mathcal{L}_{seg}(x_s, y_s) + \mathcal{L}_{adapt}(X_T^{'}) \tag{6}$$

where first term is cross entropy loss over source domain $\mathrm{X_s}$ to prevent the model from forgetting the previous knowledge. Second term, is adaptation loss computed as summation of focal loss Eq. 5 and segmentation loss (Eq. 4), trying to minimize loss of violating scale-invariance.

$$\mathcal{L}_{adapt}(X_T^{'}) = \sum_{x_t^{'} \in X_T^{'}} \sum_{i}^{k} \beta \mathcal{L}_{FL}(I_i^t, \hat{y}_t^i) + \mathcal{L}_{seg}(I_i^t, \hat{y}_t^i), \tag{7}$$

β is a hyperparameter that controls the effect of focal loss on self-supervised domain adaptation. In the end, we adapt the model with an iterative process for each rounds r. Fig. 4 shows complete model.

4 Experiments and Results

In this section, we provide implementation details and experimental setup of our proposed approach. We evaluate the proposed self-supervised learning strategy on standard synthetic to real domain adaptation setup and present a detailed comparison with state-of-the-art methods.

4.1 Experimental Details

Network Architecture: For a fair comparison we follow the standard practice of using FCN-8s [17] with VGG16 and DeepLab-v2 [2] with ResNet-101 [9] as our baseline approaches. We have used pretrained models for further adaptation towards the target domain

Datasets and Evaluation Metric: To evaluate the proposed approach, we have used benchmark synthetic datasets, e.g., GTA5 [20] and SYNTHIA-RAND-CITYSCAPES [21] as our source domain datasets and real imagery Cityscapes[6] as our target domain dataset. The GTA5 dataset consists of 24966 high resolution (1052 x 1914) densely annotated images captured from the GTA5 game. Similarly, SYNTHIA contains 9400 labeled images with a spatial resolution of 760 x 1280. The Cityscapes datasets has 2975 training images and 500 validation images. We use mean intersection over union (mIoU) as the evaluation metric and evaluate the proposed approach on compatible 19 and 16 classes for GTA to Cityscapes and SYNTHIA to Cityscapes adaptation respectively. Due to GPU memory limitations we use the highest spatial size of 512×1024.

Implementation Details: We have used PyToch deep learning framework to implement our algorithm with a Tesla k80 GPU having 12GB of memory. To select number of high confident images for each class, we choose $p = 0.1$ and after each round increment it with 0.05. $k = 4$ number of patches, of spatial size of 256×512, are chosen randomly and resized to 512×1024. For focal loss, we use $\gamma = 3$ and $\beta = 0.1$ in-order to focus on hard examples. We used Adam optimizer [13] with learning rate and momentum of $1\text{x}10^{-6}$ and 0.9 respectively.

4.2 Comparisons with state-of-the-art Methods

To compare with existing methods, we perform experiments of adapting to Cityscapes from two different synthetic datasets, GTA5 and SYNTHIA. All experiments were done under the standard settings.

GTA5 to Cityscapes: Table 1 shows the comparison of our result with existing state of the art domain adaptation methods in semantic segmentation from GTA5 to Cityscapes respectively. Proposed approach reports state-of-the-art results on VGG16-FCN8 [17] and ResNet101 [9], for self-training based adaptation methods. It outperforms most of the non self-training methods and complex methods too, and is comparative to state-of-the-art. We report the results with and without the focal loss to see the effect on the model regarding class balance adaptation. Due to focal loss, the small/infrequent objects benefit specifically.

SYNTHIA to Cityscapes: Table 2 describes the quantitative results of LSE and a detailed comparison with existing methods. Like previous methods [12], we report both the mIoU (16 classes) and mIoU* (13 classes) for the classes compatible with Cityscapes. The LSE+FL performs comparative to other complex methods based on adversarial learning, however, in self-training setting LSE+FL shows 4.1% mIoU gain over state-of-the-art PyCDA [32].

GTA5 to Cityscapes																					
	Arch.	Meth.	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
FCN wild [11]	V	AT	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
CyCADA [10]	V	AT	85.2	**37.2**	76.5	21.8	15.0	23.8	22.9	**21.5**	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
ROAD [5]	V	AT	85.4	31.2	78.6	27.9	22.2	21.9	23.7	11.4	80.7	29.3	68.9	48.5	14.1	78.0	19.1	23.8	9.4	8.3	0.0	35.9
	R	AT	76.3	36.1	69.6	28.6	22.4	28.6	29.3	14.8	82.3	35.3	72.9	54.4	17.8	78.9	27.7	30.3	4.0	24.9	12.6	39.4
CLAN [18]	V	AT	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	**34.5**	72.0	45.8	7.9	80.5	**26.6**	29.9	0.0	10.7	0.0	36.6
	R	AT	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
Curr. DA [30]	V	AT	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
AdvEnt [27]	V	AT,ST	86.9	28.7	78.7	28.5	25.2	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	81.5	26.0	17.2	**18.9**	11.7	1.6	36.1
	R	AT,ST	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	**38.5**	44.5	1.7	31.6	32.4	45.5
SSF-DAN [7]	V	ST,AT	**88.7**	32.1	79.5	29.9	22.0	23.8	21.7	10.7	80.8	29.8	72.5	49.5	16.1	**82.1**	23.2	18.1	3.5	**24.4**	8.1	37.7
	R	ST,AT	90.3	38.9	81.7	24.8	22.9	30.5	37.0	21.2	84.8	**38.8**	76.9	58.8	30.7	**85.7**	30.6	38.1	5.9	28.3	36.9	45.4
CBST [34]	V	ST	66.7	26.8	73.7	14.8	9.5	28.3	25.9	10.1	75.5	15.7	51.6	47.2	6.2	71.9	3.7	2.2	5.4	18.9	**32.4**	30.9
PyCDA[15]	V	ST	86.7	24.8	**80.9**	21.4	**27.3**	**30.2**	26.6	21.1	**86.6**	28.9	58.8	**53.2**	17.9	80.4	18.8	22.4	4.1	9.7	6.2	37.2
	R	ST	**90.5**	36.3	**84.4**	**32.4**	**28.7**	**34.6**	36.4	31.5	**86.8**	37.9	78.5	**62.3**	21.5	85.6	27.9	34.8	**18.0**	22.9	**49.3**	47.4
LSE	V	ST	80.2	26.6	78.1	28.4	17.3	19.8	27.6	12.2	78.6	23.6	72.0	50.8	14.8	81.2	22.5	20.3	4.0	20.1	14.5	36.4
LSE + FL	V	ST	86.0	26.0	76.7	**33.1**	13.2	21.8	**30.1**	16.5	78.8	25.8	**74.7**	50.6	**18.7**	81.8	22.5	**30.5**	12.3	16.9	25.4	**39.0**
LSE + FL	R	ST	90.2	**40.0**	83.5	31.9	26.4	32.6	**38.7**	**37.5**	81.0	34.2	**84.6**	61.6	**33.4**	82.5	32.8	**45.9**	6.7	29.1	30.6	**47.5**

Table 1. Results from GTA5 to Cityscapes. We report the results of our algorithm by presenting IoU of each class and also overall mIoU. 'V' and 'R' represents VGG-FCN8 and ResNet101 as our baseline network. 'ST' and 'AT' represents self-training and adversarial training respectively. We report the best results in **bold**.

4.3 Analysis

To demonstrate the reasoning of the working principle for the proposed algorithm, we evaluate different aspect of our algorithm.

Effect of Focal Loss: To verify the effect of focal loss on each class equally, we calculate the number of images selected for each class after a few rounds. Focal loss can affect the smaller classes for each class on different rounds, as shown in Figure 5. The graph demonstrates the effect on different classes to balance the effect of learning for self-supervised domain adaptation. For each class, the Figure 5 shows three bars, red shows the number of images selected on the first round of adaptation, whereas the orange and green are the corresponding values of selected images after fourth round and with and without focal loss respectively. It can be seen that the focal loss balances the selection process especially for infrequent classes, by maximizing their prediction probabilities.

Performance Gap: We also compare the performance of our algorithm using the performance gap with other state-of-the-art methods of domain adaptation. Table 3 shows the performance gap of different algorithms with their oracle values. Our algorithm clearly shows the best results with a gap **-21.3** as compared to other algorithms we mentioned.

SYNTHIA to Cityscapes																				
	Arch.	Meth.	road	sidewalk	building	wall	fence	pole	light	sign	veg	sky	person	rider	car	bus	mbike	bike	mIoU	mIoU*
ROAD [5]	V	AT	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	-
CLAN [18]	V	AT	80.4	30.7	74.7	-	-	-	1.4	8.0	77.1	79.0	46.5	8.9	73.8	18.2	2.2	9.9	-	39.3
	R	AT	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
Curr. DA [30]	V	AT	65.2	26.1	74.9	0.1	0.5	10.7	3.7	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	-	34.8
AdvEnt [27]	V	AT,ST	67.9	29.4	71.9	6.3	0.3	19.9	0.6	2.6	74.9	74.9	35.4	9.6	67.8	21.4	4.1	15.5	31.4	36.6
	R	AT,ST	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
SSF-DAN [7]	V	ST,AT	**87.1**	36.5	**79.7**	-	-	-	13.5	7.8	81.2	76.7	50.1	12.7	**78.0**	**35.0**	4.6	1.6	-	43.4
	R	ST,AT	**84.6**	41.7	80.8	-	-	-	11.5	14.7	80.8	**85.3**	57.5	21.6	82.0	36.0	19.3	**34.5**	-	50.0
CBST [34]	V	ST	69.6	28.7	69.5	**12.1**	0.1	25.4	11.9	13.6	**82.0**	**81.9**	49.1	14.5	66	6.6	3.7	**32.4**	35.4	36.1
PyCDA[15]	V	ST	80.6	26.6	74.5	2.0	0.1	18.1	13.7	14.2	80.8	71.0	48.0	**19.0**	72.3	22.5	**12.1**	18.1	35.9	42.6
	R	ST	75.5	30.9	**83.3**	**20.8**	**0.7**	**32.7**	**27.3**	**33.5**	**84.7**	85.0	**64.1**	25.4	**85.0**	**45.2**	21.2	32.0	**46.7**	**53.3**
LSE	V	ST	82.2	38.4	79.0	2.2	0.5	25.3	9.6	20.7	78.6	77.4	51.7	18.0	72.9	21.7	11.1	22.2	38.2	44.9
LSE + FL	V	ST	83.6	**39.6**	79.3	3.6	**0.9**	25.3	**14.1**	**26.1**	79.4	76.5	**51.0**	18.1	75.7	22.5	12.0	32.1	**40.0**	**47.0**
LSE + FL	R	ST	82.9	**43.1**	78.1	9.3	0.6	28.2	9.1	14.4	77.0	83.5	58.1	**25.9**	71.9	38.0	**29.4**	31.2	42.6	49.4

Table 2. mIoU (16-categories) and mIoU* (13-categories) results from SYNTHIA to Cityscapes. 'V' and 'R' represent VGG-FCN8 and ResNet101 as our baseline network. 'ST' and 'AT' represent self-training and adversarial training, respectively. We have reported the highest results in **bold**.

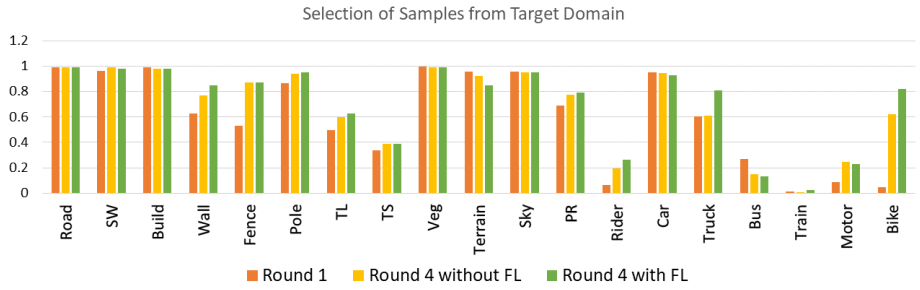

Fig. 5. Effect of focal loss on each class after the first and the fourth round of domain adaptation with self-supervised learning for semantic segmentation, evaluated for GTA5 to Cityscape with VGG16-FCN8 baseline network.

Performance Table			
GTA5 to Cityscapes (VGG16-FCN8)			
Method	Oracle	mIoU %	gap (%)
FCN wild [11]	64.6	27.1	-37.5
CyCADA [10]	60.3	35.4	-24.9
ROAD[5]	64.6	35.9	-28.7
CLAN[18]	64.6	36.6	-28.0
AdvEnt [27]	61.8	36.1	-25.7
SSF-DAN[7]	65.1	37.7	-27.4
CBST [34]	65.1	30.9	-34.2
PyCDA[15]	65.1	37.2	-27.9
Ours	60.3	39.9	**-21.3**

Table 3. Comparisons of performance gap of adaptation algorithms vs oracle scores

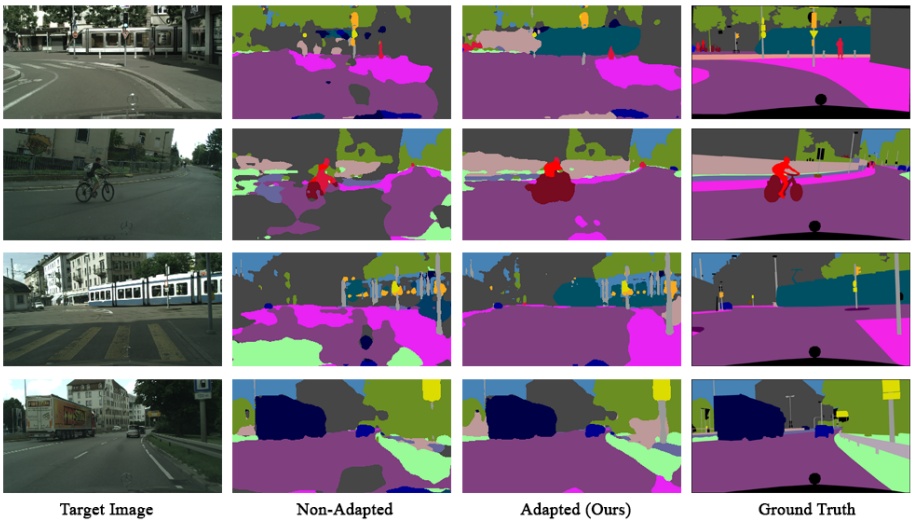

Fig. 6. Qualitative results of our algorithm with self-supervised domain adaptation for GTA5 to Cityscapes. For each example, we show images without adaptation and with adaptation as our result. We also show the ground truth for each image.

5 Conclusion

In this paper, we have proposed a novel approach of self-supervised domain adaptation method by exploiting the scale-invariance properties of the semantic segmentation model. In general images in dataset, especially road-scene dataset, contains objects in varying sizes and scene elements closer and far away from the. The scale invarance property of the model is defined as ability to assign same semantic labels to scaled instance of the image or parts of image as it will assign to the original image. In simple words regardless of size variation of object it should be similarly semantically labeled. We show that for the target domain this property is violated and could be used to direct the adaptation label by using the pseudo-labels for the original size images as pseudo-labels for the zoomed in region. Multiple strategies were employed to counter the class imbalance problem and pseudo-label selection problem. Class specific sorting algorithm is desinged to select images from target dataset such that all classes are equally represented at image level. Dynamic class dependent entropy threshold mechanism is presented to allow classes at different levels of adaptation have different threshold. Finally, a focal loss is introduced to guide the adaptation process. Our experimenal results are competitive to state-of-the-ar algorithms and outpeform state-of-the-art self-training methods.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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 (2017)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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)
4. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1992–2001 (2017)
5. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7892–7901 (2018)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
7. Du, L., Tan, J., Yang, H., Feng, J., Xue, X., Zheng, Q., Ye, X., Zhang, X.: Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 982–991 (2019)
8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
11. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
12. Iqbal, J., Ali, M.: Mlsl: Multi-level self-supervised learning for domain adaptation with spatially independent and semantically consistent labeling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (March 2020)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
15. Lian, Q., Lv, F., Duan, L., Gong, B.: Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6758–6767 (2019)

16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
18. Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y.: Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2507–2516 (2019)
19. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.: Covariate shift and local learning by distribution matching (2008)
20. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European Conference on Computer Vision. pp. 102–118. Springer (2016)
21. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
22. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Unsupervised domain adaptation for semantic segmentation with gans. arXiv preprint arXiv:1711.06969 **2** (2017)
23. Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3752–3761 (2018)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems. pp. 1195–1204 (2017)
26. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
27. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. arXiv preprint arXiv:1811.12833 (2018)
28. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.: Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1451–1460. IEEE (2018)
29. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 472–480 (2017)
30. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2020–2030 (2017)
31. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 405–420 (2018)

32. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)
34. Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 289–305 (2018)