TridentAdapt: Learning Domain-invariance via Source-Target Confrontation and Self-induced Cross-domain Augmentation

Fengyi Shen[1, 2]
fengyi.shen@tum.de

Akhil Gurram[*][2, 3]
akhil.gurram@hauwei.com

Ahmet Faruk Tuna[*†][2]
a.tuna@gmx.at

Onay Urfalioglu[†][2]
onay.urfalioglu@gmail.com

Alois Knoll[1]
knoll@in.tum.de

[1] Department of Informatics,
Technical University of Munich,
(TUM), 85748 Garching, Germany.

[2] Huawei Munich Research Center,
(MRC), 80992 Munich, Germany.

[3] Department of Computer Science,
Universitat Autònoma de Barcelona,
(UAB), 08193 Bellaterra, Spain.

Abstract

Due to the difficulty of obtaining ground-truth labels, learning from virtual-world datasets is of great interest for real-world applications like semantic segmentation. From domain adaptation perspective, the key challenge is to learn domain-agnostic representation of the inputs in order to benefit from virtual data.

In this paper, we propose a novel trident-like architecture that enforces a shared feature encoder to satisfy confrontational source and target constraints simultaneously, thus learning a domain-invariant feature space. Moreover, we also introduce a novel training pipeline enabling self-induced cross-domain data augmentation during the forward pass. This contributes to a further reduction of the domain gap. Combined with a self-training process, we obtain state-of-the-art results on benchmark datasets (e.g. GTA5 or Synthia to Cityscapes adaptation). Code and pre-trained models are available at https://github.com/HMRC-AEL/TridentAdapt

1 Introduction

Deep neural networks have shown tremendous potential in dealing with computer vision challenges such as semantic segmentation [4, 5, 28, 49], image recognition [11, 14, 16, 39, 40], etc. Semantic segmentation, a fundamental building block in autonomous driving systems, refers to the task of classifying each pixel in an image belonging to a certain semantic class. Unfortunately, the creation of datasets with pixel-wise labels is notably a laborious and high-cost procedure.

Benefiting from the development of modern computer graphics technology such as game simulators [10, 35, 37], the appearances of real-world objects can be well imitated via virtual imagery. Moreover, fine-grained semantic image labels can be acquired in large-scale for free
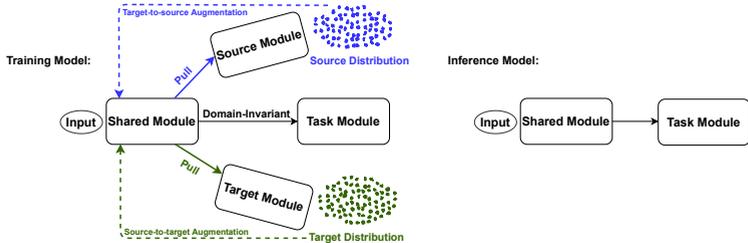

Figure 1: **Algorithmic Overview**. For any given input data (source or target domain), we leverage the source and target modules to put confrontational constraints to the shared module simultaneously, forcing it to produce a domain invariant representation of the feature map. The novel augmented views induced from source and target modules are then fed back to the shared module to further bridge the domain gap. The learning of feature domain-invariance will boost the performance of the task module on target domain data. Only the shared and task modules will be adopted for inference.

from those virtual environments, which opens a new research direction for computer vision applications. However, due to lighting and texture differences between objects in virtual and real-world images, direct knowledge transfer from virtual to real is limited. Well-trained models on virtual source datasets often experience drastic performance drop when they are applied on real-world target domain where labels are missing.

After showing promising results in pixel-level image style transfer, image-to-image translation [18, 27, 53] models have brought much attention to unsupervised domain adaption. A pioneer research is conducted in [15], which makes use of target-like images translated from source domain together with their labels to provide guidance when learning target domain segmentation. Despite the visually pleasing target-like translations, important features get lost during image translation, because a pure image-to-image translation model is not built for conveying semantic information, additionally it does not guarantee a perfect mapping from source to target domain. Thus, the performance gain for semantic segmentation is limited on target domain data.

Other interesting approaches such as curriculum domain adaptation [52], depth-aware domain adaptation [45] and frequency domain adaptation [48] also obtain state-of-the-art results for domain adaptive semantic segmentation. Unfortunately, adaptable knowledge is still not fully explored and transferred from source to target domain, and hence remains an open topic.

In this work, we provide a new perspective for solving the domain adaptive semantic segmentation problem. We hypothesize that a shared feature map which can help produce a source as well as a target domain output simultaneously, should be domain-invariant. To this end, as Fig. 1 illustrates, we design our TridentAdapt framework such that the shared module is forced to satisfy both source and target constraints, resulting in learning domain-invariant representations from input data. With this design, the framework further benefits from its back-fed cross-domain augmented views that are self-induced by source and target modules during training to bridge the domain gap.

We summarize our contributions as follows:

- We put forward an intuitive yet effective trident-like architecture in which source and target distributions adopt a confrontational stance, compelling the shared encoder to produce feature maps that are indistinguishable in terms of domain.

- We incorporate cross-domain augmented views which are self-induced from our framework into the training pipeline. Thus, we introduce semantic consistency losses on feature level and extra segmentation losses on output level, which not only encourage semantic information to be conveyed into the generators for better output quality, but also boost the learning of domain-invariance for the encoder and segmentation network.

- Our trained models demonstrate superior results over state-of-the-art methods on challenging benchmark datasets.

2   Related Work

Input Level Adaptation. Since image-to-image translation models [18, 27, 53] based on GANs [1, 13, 20, 21, 29] can be trained following an unsupervised manner, in [15, 50] target-like images translated from source domain are involved to train a target segmentation model for cross-domain improvement. [6] employs for source and target domains separate segmentation networks, to which source data and their target-like translations are fed, respectively, to force the segmentation maps to be consistent. In [41], other than adversarial feature alignment, all possible cross-domain outputs based on CycleGAN training pipeline are taken into consideration to reduce the domain gap. [50] improves source-to-target translation by incorporating SPADE layers [33]. In another work [22], semantic segmentation is learned by receiving target-like translations together with the stylized source images carrying various texture changes to prevent the segmentation network from overfitting on one specific source texture. SG-GAN [24] employs a gradient-sensitive loss and a semantic-aware discriminator to improve structural contents after image translation. However, the above approaches seek to build separate networks for image translation and semantic segmentation purposes, where domain transfer modules and segmentation encoder do not place sufficient constraints on each other. Therefore, the potential of image translation is not fully explored to support domain adaptive semantic segmentation. Our proposed method tackles this by sharing the learned semantic knowledge across all networks and looping back cross-domain outputs to collaboratively learn a reinforced domain-agnostic feature space.

Output Level Adaptation. Aligning segmentation outputs between source and target domains is considered an effective way to narrow the domain gap. Adversarial learning is adopted in [42, 44] by connecting source and target segmentation outputs to a discriminator which learns structured output space. [47] proposes to refine segmentation by connecting a image reconstruction network to the output label maps. Self-training [3, 54] such as pseudo-label generation, which enhances the confidence level of knowledge learned from source data, has become a widely adopted concept in domain adaptive semantic segmentation. In [7, 12, 22, 25, 52, 56, 57], self-training is conducted to acquire pseudo-labels which further enhance segmentation performance on target domain.

Feature Level Adaptation. In semantic segmentation models, deeper layer features often convey rich semantic information. Therefore, exploring a domain-invariant feature space, which can be shared by both source and target input data, will theoretically bring significant effect on minimizing cross-domain discrepancy. However, this has always been a key challenge in domain adaptive semantic segmentation. Inspired by adversarial learning, a feature discriminator is introduced in [15] to force the encoder of segmentation network to extract similarly distributed feature maps from both domains. Taking one step forward, [12]
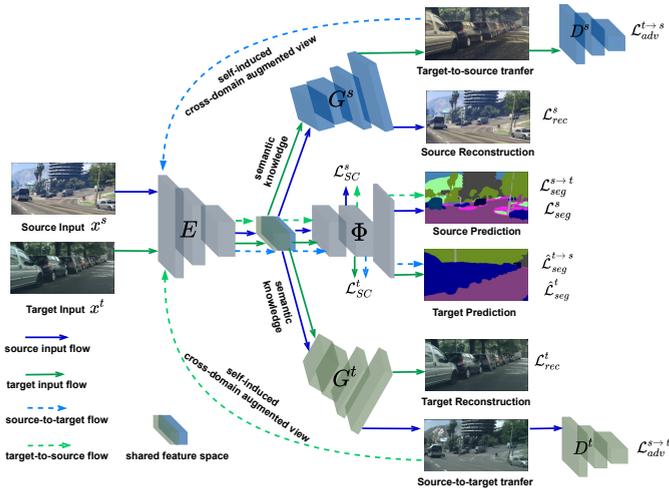
Figure 2: A pictorial overview of our proposed TridentAdapt framework in detail.

makes use of the downsampled pseudo-labels to design a semantic-wise separable feature discriminator which allows class level adversarial learning to improve the domain invariance of deep features. In [58] and [55], the segmentation encoder is connected by a shared generator to produce domain transferred outputs alternately to reduce domain gap on feature level. However, the generated images are not utilized to further interact with the segmentation network for mutual improvement. DCAN [46] seeks to align channel-wise statistics of bottleneck features between source and target domains via AdaIn [17]. Inspired by [18], [2] is trained to split input data into domain-specific texture and domain-invariant structure, and the learned structural features are used to train the segmentation network. In our framework, however, we implicitly put source and target distribution in confrontation to help search for a domain-invariant feature space, obtaining better performance than existing approaches.

3 Proposed Method

In this section, we introduce TridentAdapt framework for domain adaptive semantic segmentation. Let $\{X^s, Y^s\}$ and X^t denote the source and target domain datasets respectively, where $x^s \in X^s$ stands for a source training RGB image with corresponding source label map $y^s \in Y^s$, and $x^t \in X^t$ stands for a target training image whose label $y^t \in Y^t$ is missing. The goal is to train a model that is able to predict correct per-pixel label for X^t by the assistance of $\{X^s, Y^s\}$. As depicted in Fig. 1, we achieve this by leveraging a confrontation between source and target distributions to learn a domain-invariant feature space for the segmentation network Φ. The learning is reinforced by introducing self-induced cross-domain augmented data into a backward loop, bridging the domain gap further. Detailed pictorial description is provided in Fig. 2.

3.1 Source-target confrontation

To achieve our design purpose, effective data distribution modelling for both domains is a key step. Although classical image-to-image translation approaches [18, 27, 53] exhibit some clues on how image distribution can be modelled using GANs, yet they focus more

on altering image appearances for style transfer, ignoring the semantic level information. Therefore, for modelling source and target distributions we propose semantic-aware generators, to which semantic knowledge from the encoder is incorporated. As shown in Fig. 2, TridentAdapt training pipeline contains mainly four modules: shared encoder E initialized by weights of ImageNet [9] pretrained backbone network, source generator G^s, target generator G^t and segmentation network Φ.

If E receives an image x^s from source domain dataset during training and outputs a feature map $E(x^s)$, based on which G^s will produce a source reconstruction image $G^s(E(x^s))$ measured by L1-loss $\mathcal{L}^s_{rec}$ in Eq. (1). Note that for $\mathcal{L}^s_{rec}$, pixels belonging to image edges are prioritized in order to compensate the omitted pixels due to the max pooling operations in the encoder.

$$\mathcal{L}^s_{rec}(E,G^s;X^s) = \mathbb{E}_{x^s \sim X^s}[||\Omega^s \odot (G^s(E(x^s)) - x^s)||_1] \tag{1}$$

$\Omega^s \in \mathcal{R}^{c^{im} \times h \times w}$, is a weight matrix where $\omega^s_{ijk} = \begin{cases} 1 + \eta^s, & \text{if } x^s_{ijk} \in edges \\ 1, & \text{otherwise} \end{cases}$

(i stands for the channel index and j,k the spatial indices). $\odot$ denotes element-wise multiplication. Empirically we set η^s to 0.5. Here the $edges$ are computed based on a sobel operator, where h, w and c^{im} are height, width and the number of channels of an image x^s respectively.

In parallel, G^t takes the same feature map $E(x^s)$ and generates a source-to-target transferred image $G^t(E(x^s))$ following target distribution with help of the target discriminator D^t, computing adversarial loss $\mathcal{L}^{s \to t}_{adv}$ based on LSGAN [29] for target domain,

$$\mathcal{L}^{s \to t}_{adv}(E,G^t,D^t;X^t,X^s) = \mathbb{E}_{x^t \sim X^t}[(D^t(x^t))^2] + \mathbb{E}_{x^s \sim X^s}[(1 - D^t(G^t(E(x^s))))^2] \tag{2}$$

Here the domain-agnostic feature is learned by the confrontational constraints coming from intra-domain reconstruction (pulling towards source) and cross-domain transfer (pulling towards target) simultaneously.

At the same time, in the middle path of TridentAdapt, Φ is the task module which takes encoded feature maps for semantic segmentation. By default we denote the output of Φ as an upsampled probability map which has been processed through softmax operation. As ground truth is known for each x^s, during training we minimize the cross entropy loss $\mathcal{L}^s_{seg}$ using source prediction $\Phi(E(x^s))$ under the supervision of y^s:

$$\mathcal{L}^s_{seg}(E,\Phi;X^s,Y^s) = -\mathbb{E}_{(x^s,y^s) \sim (X^s,Y^s)} \sum_{h,w,c} y^s_{(h,w,c)} \log(\Phi(E(x^s)))_{(h,w,c)} \tag{3}$$

Here c is the number of semantic classes. Since E also serves as feature extractor for Φ, semantic information will thus be incorporated into G^t to make it semantic-aware while generating $G^t(E(x^s))$.

On the other way round, if E receives an image x^t from target domain dataset (unlabelled), following a symmetric data flow, the confrontational constraints on feature map $E(x^t)$ can be computed similar to Eq. (1) and Eq. (2) to obtain $\mathcal{L}^t_{rec}$ and $\mathcal{L}^{t \to s}_{adv}$ respectively.

In the above procedure, based on input data domain, G^s and G^t are updated according to a role-switching mechanism, e.g., during training G^s is always an image decoder from perspective of x^s but is adopted as image translator for x^t. By switching the role of the generators according to the switch of input data, our framework guarantees that G^s only produces source images, G^t purely outputting target images regardless of the input data domains. In

this way, for example, when G^s is used to reconstruct x^s, it learns how 'real' source data distribution should be reflected on its output, and this knowledge is expected to provide weak guidance to fine-tune itself when it takes the next role as image translator for x^t to generate a 'fake' source image $G^s(E(x^t))$, contributing to reinforced target to source transfer.

3.2 Self-induced Cross-domain Augmentation

The shared trident-like design of our framework allows us to utilize its own outputs to enhance the learning of domain-invariance in a self-served manner. We introduce a backward loop where the self-induced cross-domain augmented views $G^t(E(x^s))$ and $G^s(E(x^t))$ are fed to E. As those views never appear in training set but they resemble cross-domain data distributions, such that E receives a broader coverage of input data. In this way, the augmented views that are self-induced on-the-fly during each iteration can provide a smooth transition to bridge the domain gap between source and target domains. To enable this, we introduce semantic consistency(SC) loss $\mathcal{L}_{SC}$ to force the encoder to consider each input and its cross-domain version semantically identical.

Specifically, we take the semantic features from an intermediate layer of the segmentation network Φ, which is divided into 2 blocks $\Phi = \{\Phi_f, \Phi_s\}$. Now for each domain we obtain one feature pair, $i.e.$, $\Phi_f(E(x^s))$ and $\Phi_f(E(G^t(E(x^s))))$ for source, $\Phi_f(E(x^t))$ and $\Phi_f(E(G^s(E(x^t))))$ for target. Minimizing the feature distance of each pair, we compute:

$$\mathcal{L}_{SC}^s(E,\Phi_f,G^s;X^s) = \mathbb{E}_{x^s \sim X^s}[||\Phi_f(E(G^t(E(x^s)))) - \Phi_f(E(x^s))||_1] \quad (4)$$

$$\mathcal{L}_{SC}^t(E,\Phi_f,G^t;X^t) = \mathbb{E}_{x^t \sim X^t}[||\Phi_f(E(G^s(E(x^t)))) - \Phi_f(E(x^t))||_1] \quad (5)$$

In addition, for each input and its generator outputs, we add a VGG-based perceptual loss [19] which is widely adopted for training image-to-image translation models, maintaining feature-level structure between input and output images. We get $\mathcal{L}_{percep}^{s_{rec},s}$, $\mathcal{L}_{percep}^{s \rightarrow t,s}$, $\mathcal{L}_{percep}^{t_{rec},t}$ and $\mathcal{L}_{percep}^{t \rightarrow s,t}$ based on L1-metric following exactly [19].

Since E is shared across G^s, G^t and Φ, with $\mathcal{L}_{SC}^s$ and $\mathcal{L}_{SC}^t$ we also improve G^s and G^t to carry higher-level semantic information. We choose Φ_f instead of E bottleneck features, as it contains richer task-specific information.

Augmented Semantic Segmentation. For further fine-tuning the segmentation head using source-to-target augmented view, we let $G^t(E(x^s))$ share the same ground truth y^s, computing another supervised segmentation loss $\mathcal{L}_{seg}^{s \rightarrow t}$ similar to Eq. (3). Thus, Φ gradually adapts itself to images which display target domain characteristics.

Self-training. Although ground-truths for target domain are missing, the above training steps can still provide large support for learning domain-invariance to improve segmentation performance on target domain. Therefore, we first train our framework for a warming-up stage which is referred to as 'stg1'. Thereafter for target domain dataset we perform pseudo-labelling [25] offline based on predictions of Φ (class-wise median probability as threshold) to acquire a set of pseudo-labels $\hat{Y}^t$, with which we go for a self-training phase 'stg2'.

Upon the acquisition of $\hat{Y}^t$, we are able to enhance segmentation for target domain data simply by adding loss $\hat{\mathcal{L}}_{seg}^t$ during each training iteration,

$$\hat{\mathcal{L}}_{seg}^t(E,\Phi;X^t,\hat{Y}^t) = -\mathbb{E}_{(x^t,\hat{y}^t) \sim (X^t,\hat{Y}^t)} \sum_{h,w,c} \hat{y}_{(h,w,c)}^t \log(\Phi(E(x^t)))_{(h,w,c)} \quad (6)$$

Similarly, we also fine-tuned the segmentation head further by computing $\hat{\mathcal{L}}_{seg}^{t \rightarrow s}$ using the self-induced target-to-source view $G^s(E(x^t))$ which enjoys the same pseudo-label.

3.3 Full objective

We denote $\mathbb{N}$ as the set of names for all above losses, summing up all of which formulates our full objective for training TridentAdapt:

$$\mathcal{L}_{\text{TridentAdapt}} = \sum_{i \in \mathbb{N}} \lambda_i \mathcal{L}_i \qquad (7)$$

4 Experiments and Discussion

In this section, we provide our experimental setups and report our results on benchmark datasets for comparison with state-of-the-art methods.

4.1 Datasets and Implementation Details

For target domain we adopt Cityscapes dataset [8] containing 2975 annotated street scene images with 2048×1024 resolution (annotations excluded during training) and 500 images for validation, and we consider for source domain the GTA5 dataset [35] consisting of 24,966 annotated images with 1914×1052 resolution taken from game engine, as well as SYNTHIA-RAND-CITYSCAPES dataset [37] consisting of 9,400 images of 1280×760 resolution with fine-grained segmentation labels.

We implement TridentAdapt with Pytorch [34] on an NVIDIA Quadro RTX 8000 with 48 GB memory, of which around 37 GB will be taken when running our training scripts. For all experiments we use pretrained ResNet-101 [14] as backbone feature extractor to initialize encoder E and adopt Deeplab-V2 [4] for segmentation network Φ. During training, we first resize source input images to 1280×720 resolution and target images to 1024×512 resolution but take 512×256 random crops for both domains in each training iteration. We train our framework using batch size 4 and set the max training iteration number to 2.5×10^5 (Though the model performs equally well even before 2×10^5 iterations). We use the SGD [36] optimizer with a default learning rate of 2.5×10^{-4} for E and Φ, and Adam [23] optimizers for G^s, G^t with default learning rate 1.0×10^{-3} but 1.0×10^{-4} for D^s, D^t. Polynomial decay policy is applied to all learning rates. We set momentum to 0.9 and 0.99. For evaluation on target validation set we upsample the segmentation predictions to the full resolution of Cityscapes [8] dataset.

Notice that for the first 3500 iterations in 'stage 1' we detach feature maps from E before passing to G^s and G^t, aiming to let them warm up and not to deteriorate the pretrained encoder E in the initial training stage. Starting from the 5000th iteration we take source-to-target transferred images to compute segmentation loss and all cross-domain augmented images to compute semantic consistency losses as they look realistic enough.

We adopt multi-scale discriminator architecture for D^s and D^t following [18]. During training, we weight the losses with different hyperparameters, here empirically we set $\lambda_{rec}^s = \lambda_{rec}^t = 1$, $\lambda_{adv}^{s \to t} = \lambda_{adv}^{t \to s} = 0.1$, $\lambda_{sc}^s = \lambda_{sc}^t = 0.1$, $\lambda_{seg}^s = \lambda_{seg}^{s \to t} = 1$, $\hat{\lambda}_{seg}^t = \hat{\lambda}_{seg}^{t \to s} = 0.75$, $\lambda_{percep}^{s_{rec},s} = \lambda_{percep}^{t_{rec},t} = 0.5$, $\lambda_{percep}^{s \to t,s} = \lambda_{percep}^{t \to s,t} = 0.25$.

4.2 Comparison with State of the Art

Our TridentAdapt approach shows leading performance among the state-of-the-art methods on GTA5-to-Cityscapes adaptation presented in Table 1, achieving 53.3 mIoU. Moreover, our approach outperforms other methods by considerable margins on many challenging

classes (e.g., pole, traffic light, trafic sign, person, rider, motorcycle, etc.). Fig. 3 visually presents examples of the segmentation results by TridentAdapt. On Synthia-to-Cityscapes adaptation in Table 2, we also achieved state-of-the-art results, being superior to other methods in segmenting cars, buses, etc. Our framework is compatible with other data augmentation techniques such as [51] and [31] to achieve better results (See Supplementary Sec.4).

4.3 Ablation study

In this section, we investigate the effectiveness of TridentAdapt components. In rows 1-8 of Table 3, we add each component at a time to show the performance of segmentation task on GTA5-to-Cityscapes adaptation. The 1st row stands for training with source only data, whereas the 2nd row is our baseline which is taken from [42]. The 3rd row shows that training with G^s & G^t will bring confrontational constraints (described in Sect. 3.1) to E to learn feature domain-invariance, yielding a performance increase of 0.7 over the baseline, and adding VGG perceptual losses in row 4 gives 0.3 improvement. In the 5th row, involving self-induced cross-domain augmentations (described in Sect. 3.2) to compute $\mathcal{L}_{sc}^s$ & $\mathcal{L}_{sc}^t$ leads to a performance boost to 44.6 mIoU, which suggests that domain gap can be bridged by our self-induced augmentations, widening input coverage in domain level. By introducing $\mathcal{L}_{seg}^{s \to t}$ to the previous setup (5th row) for fine-tuning ϕ brings another performance gain of 1.9 mIoU as can be seen in 6th row. Note that the result seen in 6th row, which totally relies on adversarial learning for domain adaptation, already outperforms some existing self-training based approaches [7, 12, 32, 56]. Finally, adding pseudo-labelling by computing $\hat{\mathcal{L}}_{seg}^t$, $\hat{\mathcal{L}}_{seg}^{t \to s}$ leads to the best performance in 8th row, which increases mIoU from 46.5 to 53.3. This is conceivable since E is shared across G^s, G^t and Φ, any improvement on E means other modules can benefit as well, thus enhancing the learning objectives among each other. Row 9 indicates that training without looping back our self-induced augmentations will lead to a performance drop from 53.3 to 47.6. Overall, this analysis shows how all the considered components are relevant in our proposal.
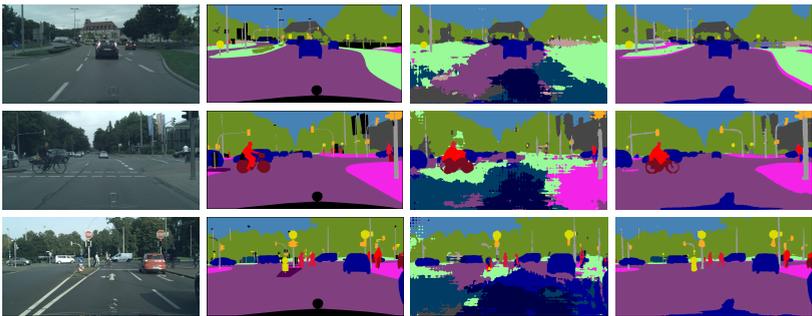
Figure 3: Qualitative results of GTA5-to-Cityscapes adaptation on Cityscapes validation set. Columns from left to right are: target domain inputs; ground-truth labels; segmentation predictions of source-only model; segmentation predictions of TridentAdapt.

4.4 Learning Domain-invariance

We show empirically that the shared encoder E reduces the domain gap at feature-level by visualizing the Φ_f feature-map distribution (described in Sect. 3.1) and projecting the

Method	road	sdwlk	bldng	wall	fence	pole	light	sign	veg	trrn	sky	psn	rider	car	truck	bus	train	moto	bike	mIoU
Source-only [ ]	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
AdaptSeg [ ]	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
ADVENT [ ]	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	45.4
SSF-DAN [ ]	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
CRST [ ]	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
PyCDA [ ]	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
BDL [ ]	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
IntroDA [ ]	90.6	36.1	82.6	29.5	21.3	27.6	31.4	23.1	85.2	39.3	80.2	59.3	29.4	86.4	33.6	53.9	0.0	32.7	37.6	46.3
TIR [ ]	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2
SAIT [ ]	91.2	43.3	85.2	38.6	25.9	34.7	41.3	41.0	85.5	46.0	86.5	61.7	33.8	85.5	34.4	48.7	0.0	36.1	37.8	50.4
TridentAdapt	91.3	51.5	86.4	38.8	36.4	42.3	45.4	42.0	86.6	36.4	84.3	67.7	42.8	89.1	41.7	38.2	20.6	40.3	30.7	53.3

Table 1: GTA5-to-Cityscapes adaptation results. we compare our model performance with state-of-the-art methods which are trained with `ResNet-101` [14] and `Deeplab-V2` [4] based models. In all the tables of Sect. 4, bold stands for **best**, and underline for second-best.

Method	road	sdwlk	bldng	wall*	fence*	pole*	light	sign	veg	sky	psn	rider	car	bus	mcycl	bcycl	mIoU	mIoU*
Source-only [ ]	55.6	23.8	74.6	-	-	-	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	-	38.6
AdaptSeg [ ]	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
ADVENT [ ]	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 [ ]	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
CRST [ ]	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
PyCDA [ ]	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
BDL [ ]	86.0	46.7	80.3	-	-	-	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	-	51.4
IntroDA [ ]	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
TIR [ ]	92.6	53.2	79.2	-	-	-	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	-	49.3
SAIT [ ]	87.7	49.7	81.6	-	-	-	19.3	18.5	81.1	83.7	58.7	31.8	73.3	47.9	37.1	45.7	-	55.1
TridentAdapt	89.5	51.9	79.1	7.3	1.1	34.3	15.2	25.8	80.4	88.0	57.3	19.2	87.5	52.2	18.6	42.1	46.8	54.4

Table 2: Synthia-to-Cityscapes adaptation results. mIoU, mIoU⋆ refer to 16-class and 13-class experiment settings, respectively.

high dimensional features class-wise on a 2D-space using the t-SNE [43] algorithm. We compare the class-wise feature distributions before adaptation (source-only model) and after adaptation (TridentAdapt). In addition to the visualization we compute for each class the Cluster-Center-Distance (CCD) bewteen source and target feature vectors. Fig. 4 illustrates the analysis on GTA5-to-Cityscapes adaptation for four different classes (road, sidewalk, wall, and person). From road, sidewalk and wall classes we can clearly see the source and target domain features of the source-only model are building two different clusters, which are perfectly separable. In comparison, however, we can see that our TridentAdapt model is able to narrow the feature-distribution gap for these classes and can reduce the CCD drastically. For person class, although the features of source-only model are not as fully separable as in previous examples, nevertheless by comparing the CCD we can also confirm that TridentAdapt reduces the domain gap at feature level. This analysis implies that the proposed TridentAdapt approach enforces the shared encoder to produce domain-invariant features, such that the cross-domain discrepancy between source and target input data is reduced.

4.5 Modelling source and target distributions

We demonstrate experimentally the visibility of employing our proposed semantic-aware generators to model source and target distributions, therefore placing strong domain-specific constraints to the shared encoder in achieving domain-invariance and providing substantial support for bridging domain gap when introduced into a backward loop. We first evaluate the target-to-source image translation performance on Cityscapes validation set. Specifically, we compare CycleGAN [53] and TridentAdapt translated images by passing them to a source-

configuration	mIoU
1. X^s, Y^s [12]	36.6
2. $+X^t$ [12]	42.4
3. $+G^s \& G^t$	43.1
4. $+\mathcal{L}_{percep}$	43.4
5. $+\mathcal{L}_{sc}^s \& \mathcal{L}_{sc}^t$	44.6
6. $+\mathcal{L}_{seg}^{s \to t}$ (**stg1**)	46.5
7. $+\hat{\mathcal{L}}_{seg}^t$	51.8
8. $+\hat{\mathcal{L}}_{seg}^{t \to s}$ (**stg2**)	**53.3**
9. $-\mathcal{L}_{sc}^s \& \mathcal{L}_{sc}^t$ $-\hat{\mathcal{L}}_{seg}^{t \to s} - \mathcal{L}_{seg}^{s \to t}$	47.6

Table 3: Ablation Study on GTA5-to-Cityscapes adaptation
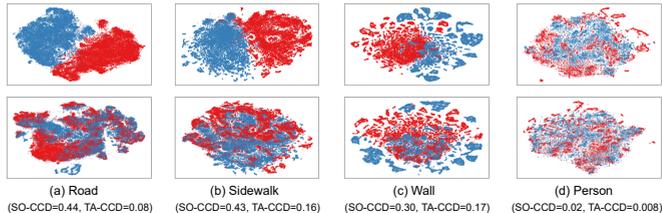
Figure 4: Class-wise t-SNE [44] visualization: GTA5 features (red) vs. Cityscapes features (blue). From top to bottom: source only (SO-CCD) before adaptation and TridentAdapt model (TA-CCD) after adaptation.

(a) Road (SO-CCD=0.44, TA-CCD=0.08) (b) Sidewalk (SO-CCD=0.43, TA-CCD=0.16) (c) Wall (SO-CCD=0.30, TA-CCD=0.17) (d) Person (SO-CCD=0.02, TA-CCD=0.008)

Translation Model	mIoU	Δ
Source-only (ours)	31.5	-
CycleGAN (T2S 'validate')	35.8	+4.3
TridentAdapt (T2S 'validate')	**38.6**	+7.1
CycleGAN (S2T 'train')	39.3	+7.8
TridentAdapt (S2T 'train')	**44.5**	+13.0

Table 4: Quantitative comparison of image translation results for semantic segmentation.

Figure 5: Visual comparison of image translation results. From left to right: input, Cycle-GAN and TridentAdapt output.

only (GTA5) segmentation model, which is, for fair comparison, trained using our configurations (e.g., batch size, crop size, augmentations) as in TridentAdapt. We observe in Table 4 that our source-like images achieve a 7.1 gain in mIoU over non-translated Cityscapes validation images, outperforming CycleGAN result also by a large margin. To evaluate our target generator G^t, we train a segmentation model sorely on our source-to-target translated GTA5 images, computed mIoU of the trained model on Cityscapes validation set, and compared with the result of CycleGAN [53]. Interestingly, this improves the source-only model result by 13.0 in mIoU and outperforms CycleGAN [53] by 5.2. Fig. 5 visually reveals that TridentAdapt better preserves semantic contents (e.g., distant vegetation and vehicles) during translation. Therefore we conclude that generators of TridentAdapt are effective for modelling domain data distributions, which is beneficial to the subsequent tasks for learning domain-invariance.

5   Conclusion

We propose TridentAdapt, a trident-like architecture for domain adaptation including a source module and a target module which simultaneously imposes confrontational constraints on the shared feature encoder. We present a novel framework which produces self-induced cross-domain augmentations during the forward pass to further reduce domain gap. Experimental results show SOTA semantic segmentation performance on target domain data.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1909, 2019.

[3] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[4] 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, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1791–1800, 2019.

[7] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019.

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

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] Liang Du, Jingang Tan, Hongye Yang, Jianfeng Feng, Xiangyang Xue, Qibao Zheng, Xiaoqing Ye, and Xiaolin Zhang. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 982–991, 2019.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 2014.

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

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[16] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[20] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *International Conference on Learning Representations*, 2019.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[22] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12975–12984, 2020.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[24] Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv preprint arXiv:1801.01726*, 2018.

[25] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[26] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6758–6767, 2019.

[27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 700–708, Red Hook, NY, USA, 2017.

[28] 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.

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[30] Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*, 2020.

[31] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

[32] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3764–3773, 2020.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

[35] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[36] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[37] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[38] 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*, pages 3752–3761, 2018.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[41] Marco Toldo, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for mobile semantic segmentation based on cycle consistency and feature alignment. *Image and Vision Computing*, 95:103889, 2020.

[42] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[44] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[45] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019.

[46] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018.

[47] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European Conference on Computer Vision*, pages 480–498. Springer, 2020.

[48] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[49] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision – ECCV 2020*, pages 173–190, 2020.

[50] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019.

[51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[52] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[54] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.

[55] Xinge Zhu, Hui Zhou, Ceyuan Yang, Jianping Shi, and Dahua Lin. Penalizing top performers: Conservative loss for semantic segmentation adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–583, 2018.

[56] Yang Zou, Zhiding Yu, BVK 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.

[57] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.