Cross-Domain Transfer Learning with CoRTe: Consistent and Reliable Transfer from Black-Box to Lightweight Segmentation Model

Cross-Domain Transfer Learning with CoRTe: Consistent and Reliable Transfer from Black-Box to Lightweight Segmentation Model

Claudia Cuttano Antonio Tavera Fabio Cermelli Giuseppe Averta
Barbara Caputo
name.surname@polito.it
Politecnico di Torino, Corso Duca degli Abruzzi, 24 — 10129 Torino, ITALIA

Abstract

Many practical applications require training of semantic segmentation models on unlabelled datasets and their execution on low-resource hardware. Distillation from a trained source model may represent a solution for the first but does not account for the different distribution of the training data. Unsupervised domain adaptation (UDA) techniques claim to solve the domain shift, but in most cases assume the availability of the source data or an accessible white-box source model, which in practical applications are often unavailable for commercial and/or safety reasons. In this paper, we investigate a more challenging setting in which a lightweight model has to be trained on a target unlabelled dataset for semantic segmentation, under the assumption that we have access only to black-box source model predictions. Our method, named CoRTe, consists of (i) a pseudo-labelling function that extracts reliable knowledge from the black-box source model using its relative confidence, (ii) a pseudo label refinement method to retain and enhance the novel information learned by the student model on the target data, and (iii) a consistent training of the model using the extracted pseudo labels. We benchmark CoRTe on two synthetic-to-real settings, demonstrating remarkable results when using black-box models to transfer knowledge on lightweight models for a target data distribution.

1. Introduction

In the last few years, semantic segmentation models have achieved impressive performances for various applications. The mainstream approach to increase their performance is to design deeper and wider neural networks, promoting accuracy at the expense of computational time, memory consumption, and hardware requirements. However, the continuous development of smart tiny devices, together with the wide diversity of edge applications has underlined the need of delivering models suitable for real-world applications: reduced model footprint and limited inference time.

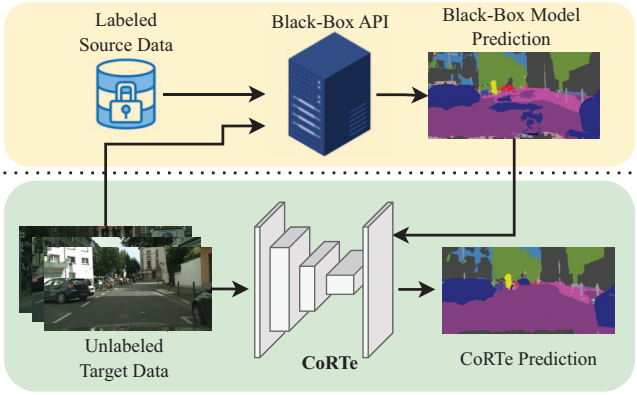

Figure 1: With CoRTe we can train a low-resources model with unlabelled target data extracting knowledge from a pre-trained source model accessible via input-output API. During the knowledge transfer, neither the *source data* nor the *source model* is accessible.

Under these regards, a crucial task is to develop methods to transfer large pre-trained models into efficient networks ready for real-world applications. However, most of the commercially available architectures are kept as black-box and secured under APIs running on cloud services to minimize model misuse and safeguard against white-box attacks. Furthermore, it is reasonable to expect that the source data used to train the model are confidential or commercially valuable and thus, not released along with the model.

This task entails two major problems: (i) transferring knowledge from a black-box teacher to an efficient target model, and (ii) addressing the domain gap that exists between the pre-training (source) and the application (target) datasets. Although the two challenges have been already studied independently, their coupled solution is not straightforward, since most of the model distillation methods assume that teacher and student models share the same data distribution, while unsupervised domain adaptation ap-

proaches disregard model efficiency and assume to have access to the source data during the alignment process.

In this paper, we fill this gap and propose a new setting, illustrated in Fig. 1, for learning a lightweight semantic segmentation model using an unlabelled target dataset and transferring knowledge from a black-box model trained on a source dataset that is not provided. The black-box model only provides the probabilities associated with the target classes that can be used to supervise the efficient model within the target domain. However, merely optimizing the target network using the predictions generated by the teacher model is vulnerable to inaccuracies that arise as a result of the domain shift between the source and target domains. Specifically, the predictions generated by the source model for the target samples are characterized by a high degree of noise. This limitation strongly vouches for the need of a more sophisticated approach to effectively address the domain shift and enhance the transfer of knowledge between the black-box predictor and the target model.

Motivated by the hypothesis that (i) trained source models yield numerous highly confident yet inaccurate predictions within the target domain and (ii) the efficient model progressively acquires valuable knowledge about the target domain, we propose CoRTe, that performs a **Co**nsistent and **R**eliable **T**ransf**e**r from a black-box source model to train a lightweight semantic segmentation model using unsupervised data. Specifically, CoRTe extracts reliable pseudo-supervision from the black-box model filtering uncertain pixels based on their relative confidence. Furthermore, to exploit the knowledge learned on the target domain, it refines the pseudo-supervision using the target model. Finally, it trains the student model introducing strong augmentations to improve its generalization abilities.

The contributions of this paper can be summarized as:

- We study the task of learning a lightweight semantic segmentation model using an unsupervised target dataset and transferring knowledge from a black-box model without being provided with the source dataset.
- We propose CoRTe, a new method able to extract reliable supervision from the black-box source model and refine it using the knowledge of the target domain.
- Comprehensive experiments demonstrate the effectiveness of our proposed method on two challenging synthetic-to-real semantic segmentation protocols.

2. Related works

Semantic Segmentation. It aims to classify each image pixel with its category label. FCN [33] was the first to efficiently learn to make dense predictions by replacing the fully connected layers with convolutional layers. This approach evolved in the encoder-decoder architecture [3, 36, 40]. To overcome low spatial resolution at the output, solutions such as dilated convolutions [7, 8, 54] and skip connections [40] were proposed. Further improvements have been achieved by harnessing context information [16, 20, 55, 7, 60, 8, 37]. Recently, with the growing popularity of attention-based Transformers, they have been effectively adapted to semantic segmentation [61, 49, 6, 32], where they have shown promise in capturing long-range dependencies and context information in images. Motivated by the assumption that traditional frameworks are often computationally demanding and complicated, [50] designed an efficient framework that combines Transformers with lightweight multilayer perceptron decoder, and scale the approach to obtain a range of models with varying levels of complexity. [18] extends [50] by leveraging the context across features from different encoder levels in order to provide additional information in the decoder, increasing decoder complexity in favor of performance.

Unsupervised Domain Adaptation (UDA). It is a form of Transfer Learning that uses labelled source data to execute new tasks in an unlabelled target domain. In adversarial learning, a discriminator is introduced to reach domain invariance by acting as source-target domain classifier at either intermediate feature level [9] or output level [47, 34, 5, 46]. While generative-based approaches aim at learning a function to map images across domains [53, 62, 26], other methods minimize the entropy to force the over-confident source behavior on the target domain [47, 53] or use a curriculum learning approach to gradually infer useful properties about the target domain [59, 27]. Self-training strategies leverage confident predictions inferred from the target data to reinforce the training. To regularize the training, approaches such as confidence thresholding [64, 63], pseudo-label prototypes [57, 58], prediction ensembling [12], consistency regularization [45, 35, 38] and multi-resolution input fusion [19] have been proposed.

Source-Free DA. Traditional UDA strategies assume the source data is available during training. The concept of *source-free* was introduced by [11], motivated by the belief that source data are often subject to commercial or confidentiality constraints between data owners and customers. In the last few years, due to the mounting concerns for data privacy, the source-free setting for UDA is receiving increasing attention. Recent works [28, 23, 52] fine-tune the model trained on the source domain with the unlabelled target data. Specifically, [28] adapt the source model with pseudo-labelling and information maximization, which is extended to multi-source in [1, 14]. Other works leverage the source knowledge of the network to synthesize target-style samples [25] or source-like samples [24, 31] based on the statistics learned by the source model. In the methods above, the details of the source model are exposed (e.g.

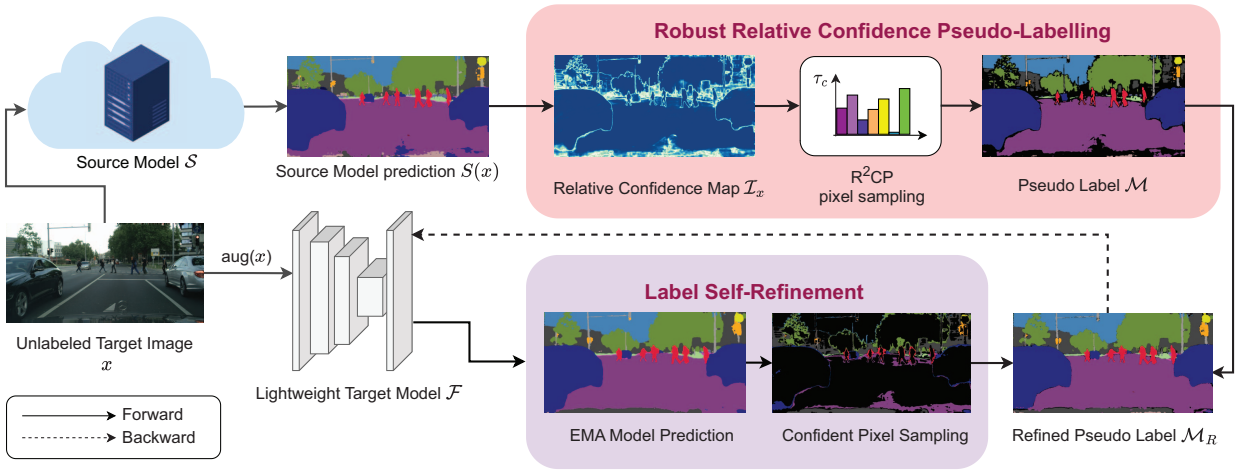

Figure 2: An overview of the proposed CoRTe framework. Our approach involves using a teacher model $\mathcal{S}$ to predict labels for an unlabelled target image x. To filter the predictions, we introduce a **Robust Relative Confidence Pseudo-Labelling** method that preserves pixels where the *relative confidence* of the model is above a threshold. The resulting pseudo label $\mathcal{M}$ is then further refined using the **Label Self-Refinement** technique, which leverages the knowledge gained by the student lightweight model. Finally, the refined pseudo label $\mathcal{M}_\mathcal{R}$ is used as the ground truth for training.

white-box source model). However, the exposition of the trained source model may be subject to white-box attacks [29]. Also, the white-box source model could be unavailable for commercial or safety reasons [56]. The task of *black-box unsupervised domain adaptation* [56], where the trained model is fixed and accessible only through an API, has been investigated mostly in image classification. [30] divides the target data into two portions to perform self-supervised learning on the uncertain split, while [56] proposes iterative learning with the noisy labels obtained from the black-box model. [29] reduces the noise through label smoothing during distillation and fine-tunes the model on the target domain. Differently from these works, we are the first to investigate the challenges of transferring the knowledge of a black-box to a lightweight semantic segmentation model using only an unsupervised target dataset.

3. Methodology

3.1. Problem Definition

The goal of Source-Free Black-Box Unsupervised Domain Adaptation for Semantic Segmentation is to learn a lightweight semantic segmentation model $\mathcal{F}_\theta$ using only an unlabelled target dataset $\mathcal{D}_\mathcal{T}$. To effectively learn from it, it is possible to exploit a black-box model $\mathcal{S}$ that is accessible only through an API, that has been trained to perform semantic segmentation on a similar yet different source domain. Formally, it is given a source model $\mathcal{S} : \mathcal{X}_\mathcal{S} \rightarrow [0,1]^{|\mathcal{Y}|\times H\times W}$, where $\mathcal{S}$ is fixed and only the per-pixel class probabilities are accessible, $\mathcal{X}_\mathcal{S}$ is the source domain image space, and $\mathcal{Y}$ the label space, with $|\mathcal{Y}|$ the number of classes. In addition, it is provided a target dataset containing n_t unlabelled images from the target domain $\mathcal{D}_\mathcal{T} = \{x | x \in \mathcal{X}_\mathcal{T}\}$, where $\mathcal{X}_\mathcal{T}$ is the target domain image space. Our goal is to train a lightweight segmentation target model $\mathcal{F}_\theta : \mathcal{X}_\mathcal{T} \rightarrow [0,1]^{|\mathcal{Y}|\times H\times W}$ to infer pixel-wise labels $\{y\}_{i=1}^{H\cdot W}$, where H, W are respectively the height and width of the images, exploiting the dataset $\mathcal{D}_\mathcal{T}$ and the supervision coming from the source model $\mathcal{S}$.

3.2. Extracting Supervision in the Target Domain

A natural solution to transfer knowledge from a pre-trained larger model to a smaller one is Knowledge Distillation (KD) [17], where a small student network is forced to mimic the teacher's predictions. A common technique employed by previous works [29, 48] is to penalize the Kullback-Leibler divergence between the output predictions of the target model (the *student*) and the source model (the *teacher*) on the target domain. However, Kim *et al*. [22] argue that when the teacher is trained on a dataset with noisy labels, it may transfer corrupted knowledge to the student. We posit that employing KD in our setting would cause the same issue due to the domain shift. Hence, as suggested by [22], we focus on *label matching* instead of *logit matching*, forcing the student to neglect the noisy information by only relying on the hard pseudo label produced by the teacher.

However, forcing the student model to perfectly fit the source predictions may lead to reproducing the same inaccuracies of the source model on the target domain, limiting its performance. To effectively address the domain shift between the source and target domain, we argue that it is essential to filter the noisy information coming from the source model by retaining only the reliable pixels to supervise the target model. In addition, the pseudo label may

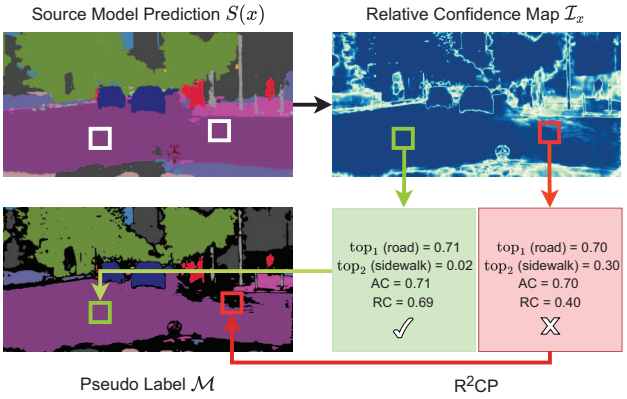

Figure 3: In this example, we demonstrate the Robust Relative Confidence Pseudo-Labelling strategy (R^{2}CP) which begins by extracting the Relative Confidence Map $\mathcal{I}_x$ from the source model's prediction on the target image $\mathcal{S}(x)$. The Relative Confidence (RC) is computed using Eq. (1) for each pixel. Finally, we apply Eq. (2) to identify the set of reliable pixels to be retained in the Pseudo Label $\mathcal{M}$.

be further refined by exploiting the new knowledge learned by the student model on the target domain.

In the following, we show how to obtain a reliable pseudo label to supervise the student model. First, we illustrate how to filter the noise when extrapolating the pseudo labels from the source model. Thereafter, we refine the pseudo labels directly using the knowledge of the target domain of the model itself. An illustration of the method is provided in Fig. 2.

Robust Relative Confidence Pseudo-Labelling. To determine the pixels that are favorable for domain transfer, a trivial technique involves applying a pixel-level filter to the pseudo labels on the basis of the model *confidence*, named *absolute confidence* (AC), which serves as an indicator of the network's reliability on each prediction. However, in the case of a noisy teacher model, this approach can produce several confident but incorrect predictions, resulting in an unreliable filtering process. This phenomenon is due to the presence of visually distinct categories in the source domain that are more difficult to distinguish in the target domain, making the teacher highly uncertain between the top two predicted classes (i.e. indecision between *bicycle* and *rider* or *sidewalk* and *road*).

A better way to consider the source model confidence is to relate the probability of the predicted class with the other classes. In particular, we consider the *relative confidence* (RC) as the difference between the probability associated with the top-first predicted class and the one associated with the top-second predicted class. Formally, given a target image x, we first get the source model probabilities $q = \mathcal{S}(x)$, and we then compute the relative confidence $\mathcal{I}_x$ as:

$$\mathcal{I}_x^i = \text{top}_1(q^i) - \text{top}_2(q^i), \tag{1}$$

where $\text{top}_1(\cdot)$ and $\text{top}_2(\cdot)$ indicate the probability value of the first and second predicted classes and q^i is the probability distribution for the i-th pixel. Intuitively, the *relative confidence* is a measure of the certainty of the prediction of the teacher network. When top_1 prevails over top_2 the difference is high, meaning that the teacher is certain about the assigned category for the given target pixel i.

Therefore, given a target image x, we obtain a reliable pseudo label $\mathcal{M}$ through the *robust RC-driven Pseudo-Labelling* (R^{2}CP) function:

$$\mathcal{M}(x)_c^i = \begin{cases} 1 & \text{if } I_x^i \geq \tau_c \text{ and} \\ & \quad c = \arg\max_{k \in \mathcal{Y}} q_k^i, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where q_k^i is the probability of the pixel i for class k, $\mathcal{M}(x)_c^i$ indicates the value of $\mathcal{M}$ for the pixel i and class c. The threshold τ_c is the average relative confidence $\mathcal{I}$ of the teacher for each class c on the whole target domain:

$$\tau_c = \frac{1}{N_T^c} \sum_{x \in \mathcal{D}_t} \sum_{i=1}^{H \cdot W} \mathbb{1}(c = \arg\max_{k \in \mathcal{Y}} q_k^i) \mathcal{I}_x^i, \tag{3}$$

where N_T^c is the number of target samples predicted as c and $\mathbb{1}(c = \arg\max_{k \in \mathcal{Y}} q_k^i)$ is the indicator function that equals 1 when the model predicts the class c for the pixel i and 0 otherwise. Fig. 3 illustrates how the Robust Relative Confidence Pseudo-Labelling method works, providing an example of its operation.

Label Self-Refinement. With respect to the source model, which is static, the target model dynamically evolves as training proceeds and gradually learns valuable knowledge about the target domain. Inspired by [21], we propose to use the target-aware predictions of the target network to refine the pseudo labels, forcing the student to become the teacher model itself over the pixels over which the source model is more uncertain.

In particular, during training, we add the supervision of a second teacher, named $\mathcal{F}_\Theta$, with Θ indicating its parameters, which is obtained as the temporal ensemble derived via exponential moving average (EMA) [43] of the target network $\mathcal{F}_\theta$. The EMA model is updated based on $\mathcal{F}_\theta$ during training following:

$$\Theta_{t+1} = \alpha \Theta_t + (1 - \alpha)\theta, \tag{4}$$

where α is a parameter controlling the update momentum, Θ_t e Θ_{t+1} are, respectively, the weight of the EMA model network before and after the update at the timestep t, and we recall θ are the target model parameters.

This secondary teacher is used to refine the pseudo labels generated in Eq. (2) by including the valuable knowledge provided on the target domain. Our goal is to refine the supervision provided by the source model by introducing a pseudo-supervision on the uncertain pixels by exploiting the confident pixels extracted from the EMA model. Formally, given an image x and denoting the predictions of the EMA model as $\hat{p} = \mathcal{F}_\Theta(x)$, we refine $\mathcal{M}$ as:

$$\mathcal{M}_R(x)_c^i = \begin{cases} 1 & \text{if } \mathcal{M}(x)_c^i = 1, \\ \lambda_t & \text{if } \hat{p}_c^i \geq \beta, \mathcal{M}(x)_c^i = 0, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where β is a confidence threshold applied to the EMA model's probabilities, $\mathcal{M}_R(x)_c^i$ and $\hat{p}_c^i$ indicate respectively the refined mask and the probability value of the EMA model at pixel i for class c, λ_t is a hyper-parameter that controls the contribution of the loss during the training. In particular, λ_t linearly increases during the training as the reliability of $\mathcal{F}_\Theta$ increases.

3.3. Consistent Training of the Target Model

The refined pseudo labels $\mathcal{M}_R$ provide direct supervision and enable the knowledge transfer between source and target models on the unlabelled target domain. However, the limited size of the target dataset and the label-matching objective between source and target models may have an impact on the generalization capability of the target network. Several works [4, 15, 51, 42, 10, 44, 2] leverage consistency regularization to make predictions on unlabelled samples invariant to perturbations. Inspired by these works, we propose to improve the generalization capability of the student by enforcing consistency regularization between the prediction of the teacher model on the original target sample and the prediction of the student on its augmented version. Specifically, for each training image x, we first compute the refined pseudo label $\mathcal{M}_R(x)$ as defined in Eq. (5). Then, instead of computing the target model's probability on x, we augment the image, such that $p(\text{aug}(x)) = \mathcal{F}_\theta(\text{aug}(x))$, where $\text{aug}(\cdot)$ is a function that strongly augments the images without introducing geometric distortions.

Finally, to train the target segmentation model, we use the refined pseudo supervision $\mathcal{M}_R(x)$ and the target model's probability obtained on the augmented image $p(\text{aug}(x))$, and we minimize the following loss function:

$$\ell(x) = -\frac{1}{H \cdot W} \sum_{i=1}^{H \cdot W} \sum_{c=1}^{C} \mathcal{M}_R(x)_c^i \log p(\text{aug}(x))_c^i, \quad (6)$$

where $p(x)_c^i$ indicates the probability of the target model on the i-th pixel and the c-th class, and H, W are the height and width of the image. Note that, when $\mathcal{M}_R(x)_c^i = 0$ for all c, the pixel i does not contribute to the loss function. Differently, if $\mathcal{M}_R(x)_c^i \neq 0$, the objective reduces to a weighted cross-entropy loss, where the supervision coming from the source model is weighted 1, while the supervision coming from the EMA model is weighted λ_t.

4. Experiments

4.1. Dataset and Evaluation Protocols

Following [45, 19, 53, 63], we demonstrate the efficacy of the proposed method on the synthetic-to-real unsupervised domain adaptation tasks, where the synthetic source labelled data comes from either GTA5 [39] or SYNTHIA [41], and the unlabelled target data from Cityscapes [13].

GTA5: consists of 24,966 training images captured in a video game with resolution 1914×1052. We resize the images to 1280×720 and randomly crop them to 512×512.

SYNTHIA: we use the SYNTHIA-RAND-CITYSCAPES subset consisting of 9,400 training images with resolution 1280×760. We randomly crop the images to 512×512.

Cityscapes: consists of real-world images collected from a car in urban environments. We use the 2,975 images from the training set as target data during training. Previous works resize the training images to 1024×512. To maintain higher resolution, we resize the training images to 1280×640 and randomly crop them to 512×512. For a fair comparison, we test on the 500 annotated images from the validation set resized to 1024×512.

We evaluate our method using the standard segmentation evaluation metrics: classwise Intersection over Union scores (**IoU**) and mean IoU (**mIoU**).

4.2. Baselines

Black-box unsupervised domain adaptation for Semantic Segmentation is fairly new. Therefore, we implemented several baselines. **Source only**: evaluates the performance of the trained source model on the target images. **No adapt**: the target network is trained on the annotated source domain without any adaptation. **DACS** [45] and **HRDA** [19]: provide adaptation during the target network training. **Naive transfer**: the target images are pseudo-labelled by the source model and used to train the target network. **KL-DIV**: we train the target model by penalizing the KL-divergence between the output predictions of the student and the teacher. **Target-only**: the target network is directly trained with the annotated target domain.

Implementation Details We employ the Transformers-based architectures tailored for semantic segmentation proposed in [50]. The source model is based on DAFormer [18]. It consists of a MiT-B5 encoder [50] and a context-aware feature fusion decoder [18]. As the target model we employ the lightweight SegFormer-B0 [50]. We pre-train the target network on the ImageNet-1K and randomly initialize the decoder. Following [18], we train the network

Method	SF	T→S	Road	Sidewalk	Building	Wall	Fence	Pole	T.Light	T.sight	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	mIoU
Source-only	✗	✗	78.6	22.2	80.0	34.6	27.0	36.8	45.7	27.0	86.5	37.2	84.2	67.2	36.1	80.0	47.8	50.4	42.4	38.7	25.6	49.9
No adapt	✗	✗	4.9	12.2	62.6	9.0	10.8	11.3	12.8	8.2	83.1	28.0	61.6	44.7	1.5	54.0	30.2	14.1	0.0	4.9	1.8	24.0
DACS[45]	✗	✗	85.4	0.0	83.7	5.5	4.0	26.5	25.9	40.3	85.7	39.9	87.9	45.5	0.1	80.5	28.7	21.4	0.0	0.1	1.0	34.9
HRDA[19]	✗	✗	94.0	63.7	85.8	40.8	14.2	31.2	36.1	46.0	89.4	46.0	92.4	58.7	2.7	85.6	33.1	44.2	0.0	39.7	57.3	50.6
Naive Transfer	✓	✓	82.0	26.7	81.2	37.0	25.8	32.6	43.5	23.2	87.6	45.0	83.3	65.4	34.7	84.2	44.0	49.2	34.1	37.0	32.6	49.9
KL-DIV	✓	✓	81.5	26.2	80.8	36.7	25.0	32.4	43.1	23.5	87.6	45.0	83.5	65.3	34.7	84.1	44.6	49.8	33.8	37.8	32.9	49.9
CoRTe	✓	✓	87.0	37.5	84.6	44.6	29.0	31.0	41.5	25.4	88.0	46.4	88.3	62.2	33.9	86.4	54.2	61.6	52.0	44.2	56.5	**55.5**
Target-only	✓	✗	97.6	80.9	90.4	54.7	53.4	50.9	58.4	68.6	90.3	58.6	93.0	73.9	52.8	92.3	60.3	76.1	53.6	52.4	69.1	69.9

Table 1: mIoU for GTA5→Cityscapes. **SF** denotes the *Source-Free* methods, whereas **T→S** refers to the methods that leverage the black-box model for training the target network. All methods use B0 as encoder, while *Source-only* uses B5.

Method	SF	T→S	Road	Sidewalk	Building	Wall	Fence	Pole	T.Light	T.sight	Vegetation	Sky	Person	Rider	Car	Bus	Motorbike	Bicycle	mIoU
Source-only	✗	✗	58.9	22.2	79.3	22.9	1.0	40.6	34.7	21.2	81.8	80.5	58.2	20.8	78.5	28.6	16.8	20.5	41.6
No adapt	✗	✗	12.9	14.8	55.9	3.6	0.0	20.8	3.8	6.2	66.1	63.5	46.7	8.9	37.4	5.9	1.6	9.6	22.3
DACS[45]	✗	✗	69.2	14.3	72.8	3.5	0.2	32.2	7.2	29.6	84.4	83.4	58.0	12.3	78.2	0.3	0.1	10.0	34.7
HRDA[19]	✗	✗	70.2	29.0	83.3	0.8	0.2	39.0	34.8	41.6	85.6	92.4	66.8	5.6	80.0	0.0	0.0	59.7	43.1
Naive transfer	✓	✓	65.2	25.1	80.8	10.0	1.3	40.5	36.5	19.1	83.7	83.1	57.6	20.6	80.5	40.2	19.5	31.4	44.0
KL-DIV	✓	✓	65.2	25.1	80.6	17.8	1.3	40.0	35.0	19.0	83.7	83.0	57.5	21.0	79.4	39.5	20.0	31.5	43.7
CoRTe	✓	✓	68.4	26.7	83.1	22.9	1.7	38.7	38.3	20.3	85.8	85.4	56.7	18.9	85.0	47.9	21.7	31.3	**45.8**
Target-only	✓	✗	97.6	80.9	90.4	54.7	53.4	50.9	58.4	68.7	90.3	93.0	73.9	52.8	92.2	60.3	52.4	69.1	71.2

Table 2: mIoU for SYNTHIA→Cityscapes. **SF** denotes the *Source-Free* methods, whereas **T→S** refers to the methods that leverage the black-box model for training the target network. All methods use B0 as encoder, while *Source-only* uses B5.

with AdamW optimizer using a learning rate of 6×10^{-5} for the encoder and 6×10^{-4} for the decoder, weight decay of 0.01, and linear learning rate warm up. We use a batch size of 8 and train the model for 80k iterations. We set $\alpha = 0.99$, $\beta = 0.60$, λ in the range [0,5]. To enforce consistency regularization, we apply Color jittering, Gaussian blur, and random flipping. Following previous works [18], the source network is optimized in a supervised manner by minimizing the cross-entropy loss on the source domain for 40k iterations using batches of 2 images.

4.3. Results

GTA5 → Cityscapes

In Tab. 1 we show the results of our proposed framework when the source model is trained on GTA5. When evaluated on the target domain, the source model achieves an overall mIoU of 49.9%, performing well on simple classes (*car, sky, road*), but failing on other classes (*sidewalk*, *bicycle*) where the discrepancy between the source and target domains has a significant impact. The domain shift is more pronounced on the lightweight target model, which achieves an overall mIoU of 24% (*No adapt*) indicating lower generalization capability on the new domain. Moreover, the *Target-only* upper bound demonstrates a high domain shift between GTA5 and Cityscapes: it achieves 69.9%, which is 45.9% higher than *No adapt*. Standard UDA techniques improve the model's performance significantly, with DACS [45] achieving 34.9% and HRDA [19] achieving 50.6%. These results show that UDA techniques help in adapting the network, improving its performance. However, they rely on source data, which are not available in our setting.

When directly trained with the knowledge generated by the source model (*Naive transfer* and *KL-DIV*), the student converges to the performance of the teacher (*Source-only*), obtaining very similar results. The target network achieves satisfying results on the unlabelled target domain (-20% w.r.t. *Target-only*), yet it copies the source model behavior,

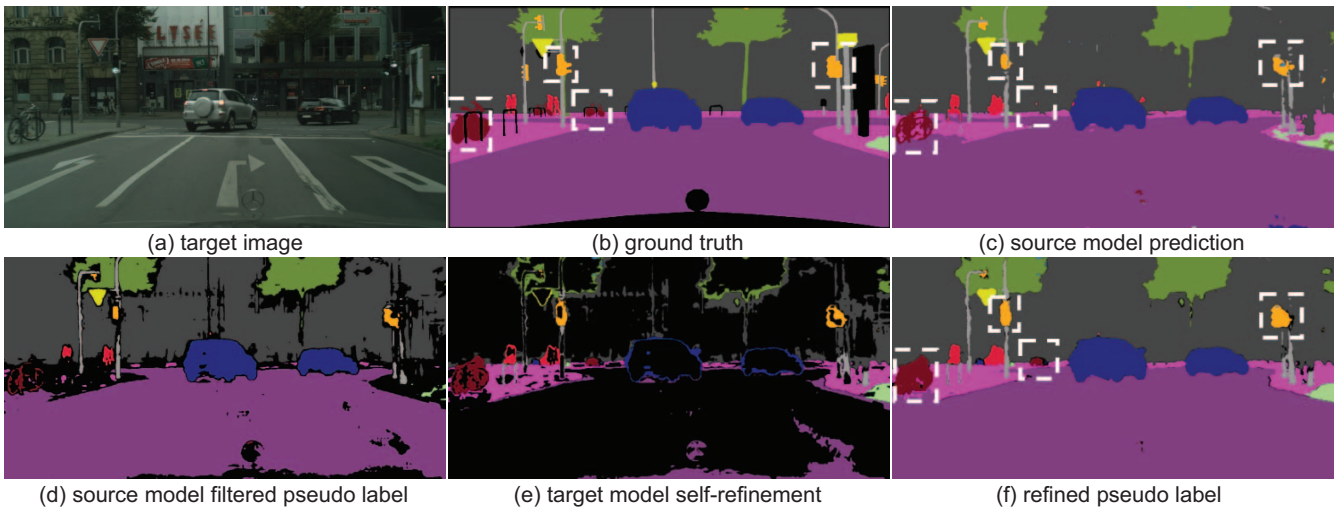

Figure 4: *Graphical interpretation of the label generation process.* At the top left, a target image from Cityscapes (a) and its corresponding label (b). We query the teacher model and obtain its prediction for the target image (c). Our robust pseudo-labelling module exploits the relative confidence of the teacher model to filter out the uncertain pixels (black in d). During the training process, the increasing knowledge of the student on the target data is used to automatically refine the pseudo label (e) resulting in the final label for training the model (f).

fitting also its noisy predictions. CoRTe outperforms them by nearly +5.6%, obtaining a gain in almost all the classes. Significant improvements are observed particularly in the classes over which the teacher is more uncertain, namely *sidewalk* (+11%) which is typically misclassified as *road*, and *bicycle* (+24%) which is often confused with *rider*. Additionally, a substantial improvement is obtained for the typically hard-to-transfer class *train* (+18%), where it almost reaches the *Target-only* upper bound (52% vs. 53.6%). Compared with *HRDA*[19] for UDA, CoRTe enables competitive performance (+4.9%), while training with images at lower resolution (512×512 vs. 1024×1024) and with no access to the annotated source dataset.

SYNTHIA → Cityscapes

In Tab. 2 we report the results of training the source model on SYNTHIA. Following the protocol of UDA, we report mIoU on the 16 classes in common with Cityscapes.

The results are remarkably coherent with the previous setting. Specifically, the source model achieves an overall mIoU of 41.6%, yet it suffers from the domain shift on hard classes such as *bicycle* or *motorbike*. In comparison, the B0 model trained on the source data (*No adapt*) achieves largely worse performance (22.3%), showing the larger model has better generalization capabilities. Moreover, the *Target-only* upper bound performance confirms the domain shift between SYNTHIA and Cityscapes: it achieves 71.2%, which is 48.9% higher than the *No Adapt* baseline. Employing standard UDA techniques, the performances improve sensibly: DACS [45] achieves 34.7% and HRDA [19] 43.1%. Differently, methods relying on the knowledge of the source model do not use source images while achieving comparable performance. In particular, *Naive Transfer* and *KL-DIV* achieve results slightly better than HRDA [19] (respectively 44% and 43.7%). Finally, we show that CoRTe outperforms all the baselines, achieving an overall IoU of 45.8%. Specifically, it improves HRDA [19] of +2.7%. In addition, it outperforms the *Naive transfer* baseline of 1.8%, showing the benefits of filtering the pseudo labels coming from the source model and refining them using the target model knowledge.

4.4. Ablation Study

Influence of each component. In this paragraph, we dissect the contributions of each component to the overall performance. In Tab. 3, we report the results when the source model is trained on GTA5. We initially show the baseline performance of the target model (line 1) trained under the supervision of the noisy pseudo labels produced by the source model (*Naive Transfer*). The addition of our Robust RC Pseudo-Labelling module (line 2) yields an improvement in terms of performance of 1.5%. Combined with Consistency Regularization (line 4), the mIoU increases up to 52%, enabling a further gain of 0.6%. The most significant contribution, however, is granted by the Label Self-Refinement (line 5), which ensures a further improvement of 3.5% in the final mIoU. In addition, we also evaluate the contribution of our Robust RC Pseudo-Labelling function with respect to the filtering function based on Absolute Confidence (AC Filtering* in Tab. 3) (line 3), proving the effectiveness of our certainty-driven filtering approach.

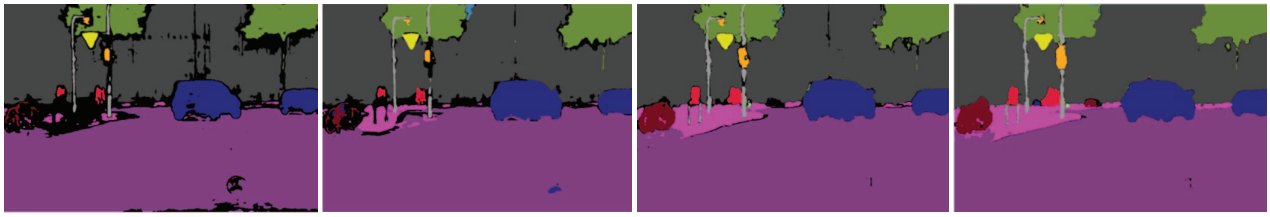

Figure 5: *Self-label refinement*. Visual representation of our refined pseudo label used at different steps of the training (from left to right at 0, 1.5k, 5k, and 80k steps respectively). Intuitively, at the very beginning of the training, the teacher's prediction is filtered from the uncertain pixels ad used to train the student. During training, this latter gradually increases its confidence and its own predictions can be used to refine the pseudo label with our Label Self-Refinement module.

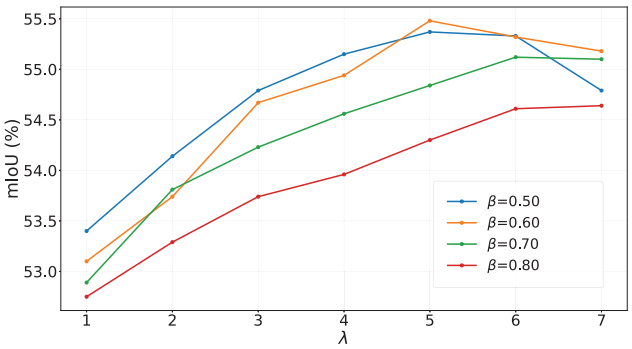

Figure 6: Parameters selection for λ_t and β.

R^{2}CP	AC Filtering*	Consistency Regularization	Label self-Refinement	mIoU
✗	✗	✗	✗	49.9
✓	✗	✗	✗	51.4
✗	✓	✗	✗	50.5
✓	✗	✓	✗	52.0
✓	✗	✓	✓	**55.5**

Table 3: Ablation study on GTA5→Cityscapes.

Parameter Sensitivity Analysis. Our student-driven label refinement involves two hyperparameters β and λ_t. To investigate their impact on the training process, we conduct experiments on GTA5→ Cityscapes. In Fig. 6 we report the resulting mIoU while changing λ_t and β. The experimental results demonstrate that our proposed model achieves the highest mIoU with the values of $\lambda_t = 5$ and $\beta = 0.60$.

The performance of the model is comparatively lower when β is set to 0.80 as a substantial number of pixels are filtered out. The model's performance is observed to improve upon decreasing the threshold value, as a greater number of informative pixels are included in the supervision process. When increasing the parameter λ_t, the mIoU rapidly increases until it reaches a plateau within the range [5, 6, 7]. This result proves that increasing the contribution of the target network in the training process significantly enhances the final performance of the network.

Qualitative Analysis. In Fig. 4 we report a qualitative interpretation of the label generation process. A comparison between the refined pseudo label (Fig. 4f) and the prediction obtained from the source model (Fig. 4c) reveals the benefit of the self-refinement process, which enables a higher level of detail (e.g. the *traffic light* in the foreground) and even the identification of entire objects (e.g. the *bicycle* in the background). The process of self-refinement gradually includes valuable knowledge during training, as shown in Fig. 5, where progressively larger portions of images are added to the supervision.

5. Conclusion

In this paper, we explore the challenging scenario of learning a compact and efficient neural network for semantic segmentation by leveraging a black-box model without access to any source data or target annotations. To address this novel setting, we propose CoRTe that reliably transfers the knowledge from the black-box predictor and provides valuable pseudo-supervision from the target model itself during training. We assess the benefits of CoRTe on two synthetic-to-real benchmarks, showing it is able to outperform all the considered transfer learning baselines.

Limitations CoRTe enables efficient knowledge transfer between a black-box source predictor and a lightweight target model, allowing it to operate on unlabelled target domains. However, it has limitations in dealing with unknown classes present in the target domain that were not learned by the source model. Additionally, to reach state-of-the-art results, CoRTe requires a robust pre-trained source model.

Acknowledgements This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] Sk Miraj Ahmed, Dripta S. Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K. Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10103–10112, June 2021. 2

[2] Edoardo Arnaudo, Antonio Tavera, Carlo Masone, Fabrizio Dominici, and Barbara Caputo. Hierarchical instance mixing across domains in aerial segmentation. *IEEE Access*, 11:13324–13333, 2023. 5

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

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 5

[5] Matteo Biasetton, Umberto Michieli, Gianluca Agresti, and Pietro Zanuttigh. Unsupervised domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[7] 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. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[9] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7892–7901, 2018. 2

[10] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[11] Boris Chidlovskii, Stephane Clinchant, and Gabriela Csurka. Domain adaptation in the absence of source domain data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 451–460, 2016. 2

[12] 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. 2

[13] 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 (CVPR)*, June 2016. 5

[14] Haozhe Feng, Zhaoyang You, Minghao Chen, Tianye Zhang, Minfeng Zhu, Fei Wu, Chao Wu, and Wei Chen. Kd3a: Unsupervised multi-source decentralized domain adaptation via knowledge distillation. In *ICML*, pages 3274–3283, 2021. 2

[15] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 5

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 2

[17] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 3

[18] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9924–9935, June 2022. 2, 5, 6

[19] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 372–391. Springer, 2022. 2, 5, 6, 7

[20] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv preprint arXiv:1907.12273*, 2019. 2

[21] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6547–6556, 2021. 4

[22] Taehyeon Kim, Jaehoon Oh, Nak Yil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2628–2635. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 3

[23] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2(6):508–518, 2021. 2

[24] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 615–625, 2021. 2

[25] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9641–9650, 2020. 2

[26] 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 (CVPR)*, June 2019. 2

[27] 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 (ICCV)*, October 2019. 2

[28] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 2

[29] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8003–8013, June 2022. 3

[30] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2022. 3

[31] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[33] 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. 2

[34] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2

[35] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021. 2

[36] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 2

[37] Rudra PK Poudel, Ujwal Bonde, Stephan Liwicki, and Christopher Zach. Contextnet: Exploring context and detail for semantic segmentation in real-time. *arXiv preprint arXiv:1805.04554*, 2018. 2

[38] Viraj Uday Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 2

[39] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 102–118, Cham, 2016. Springer International Publishing. 5

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[41] 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 (CVPR)*, June 2016. 5

[42] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 5

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4

[44] Antonio Tavera, Edoardo Arnaudo, Carlo Masone, and Barbara Caputo. Augmentation invariance and adaptive sampling in semantic segmentation of agricultural aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1656–1665, June 2022. 5

[45] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 2, 5, 6, 7

[46] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. 2

[47] 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. 2

[48] Liu X, Yoo C, Xing F, Kuo CJ, El Fakhri G, Kang JW, and Woo J. Unsupervised domain adaptation for segmentation with black-box source model. In *Proc SPIE Int Soc Opt Eng. 2022 Feb-Mar;12032:1203210.*, April 2022. 3

[49] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. *arXiv preprint arXiv:2101.08461*, 2021. 2

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090. Curran Associates, Inc., 2021. 2, 5

[51] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 5

[52] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 1(2):5, 2020. 2

[53] 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. 2, 5

[54] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[55] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 2

[56] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. *arXiv preprint arXiv:2101.02839*, 2021. 3

[57] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 2

[58] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in neural information processing systems*, 32, 2019. 2

[59] 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. 2

[60] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[61] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[62] 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. 2

[63] 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. 2, 5

[64] Yang Zou, Zhiding Yu, Xiaofeng Liu, B.V.K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2