IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Counterfactual Balancing Feature Alignment for Few-Shot Cross-Domain Scene Parsing

Cody Mingzhan Sun[1], Ziliang Chen[2,3], Liang Lin[3]

[1]Plan Cloud Research; Block B, SYSU Technology Park, Building 705, West District, No.135, Xingang West Road, Haizhu District, Guangzhou (e-mail: codysun@trussan.com)

[2]Jinan University; Huangpu Avenue, Guangzhou, China (e-mail: c.ziliang@yahoo.com)

[3]Sun Yat-sen University; 135 XinGangxi Road, Haizhu District, Guangzhou, China (e-mail: linlng@mail.sysu.edu.cn)

Corresponding author: Ziliang Chen (e-mail: c.ziliang@yahoo.com).

ABSTRACT Scene parsing becomes a key step to develop a visual autonomous driver. Real-world images are too expansive to annotate at scale, while few-shot cross-domain scene parsing (CSP) approaches only require a few labeled target images to train a model with source virtual data, thus, attracting more attention in the community. However, since the target training images are too few to support the cross-domain measures in statistics, it is inappropriate of resembling the spirit of conventional domain adaptation. In this paper, we reconsider this imbalance transfer learning demand as a covariate balancing issue regularly found in Rubin causal framework. We first consider the domain adaptation in pixels in the view of the average treatment effect (ATE), in which data are categorized into a treatment group or a control group in terms of the domain identity taken as the treatment. In this manner, the pair of domains could be perfectly aligned if the ATE converges to zero. It motivates Counterfactual Balance Feature Alignment (CBFA) to mitigate the cross-domain imbalance in the categories. CBFA revises existing adversarial adaptation techniques by modeling the propensity score for all pixels in their contexts, for the sake of predicting which groups they belong to. The propensity score for a pixel refers to its output of the domain discriminator and can be applied to balancing the adversarial adaptation objective. We evaluate our method on two suites of virtual-to-real scene parsing setups. Our method has obtained the new state of the art across 1-5 shot scenarios (in particular, 1-shot 56.79 in SYNTHIA-to-CITYSCAPES and 51.56 in GTA5-to-CITYSCAPES), demonstrating our motivation of building the connection between ATE and domain gap.

INDEX TERMS Computer vision, Few-shot domain adaptation, Urban scene parsing, Semantic segmentation, Rubin Causality.

I. INTRODUCTION

Resurrecting with huge-scale databases with labels, deep learning has already become the dominative method to predict structured labels (*e.g.*, semantic masks) in diverse machine vision areas, *e.g.*, generic semantic segmentation [3], [22], [55], [58], human body and scene parsing [10], [57], [58], *etc*. Among these pixel-level visual prediction tasks, scene parsing attracts an increasing amount of attention due to its potential in autonomous driving [31], [45]. However, building an urban scene large-scale labeled database exhausts labor efforts and can be quite expensive. To this end, plenty of current research works resort to virtual scene images, which can be handily generated by the computer-graphic programs within urban scene simulators [35], with free machine-annotated semantic masks. Incorporating the synthetic labeled source images, cross-domain scene parsing (CSP) [4], [38], [47] employs the real-world scene parser by minimizing the domain discrepancy between the virtual-world and real-world domains. The CSP models learn to transfer the virtual source semantic information into the real-world target domain and therefore, successfully spare the cost of annotating target urban scene images.

Despite the impressive performances existing CSP models [4], [38], [47] have already achieved, they may not truly fit the demand for outdoor scene understanding. It is because the technical core of normal CSP models depends on a large magnitude of target images to balance the moment matching objective across domains, where the statistic in the training

images should be promised to facilitate the label information transfer from source to target data. Despite saving the labor for the tremendous cost of manual annotation, it is still very expensive to create an urban target domain dataset to support existing strategies in the industry, since their cost are not merely *due to manual annotation efforts*, but also *due to the difficulty to collect available real-world scene images at scale*. In specific, the urban scene frequently changes because of weather, season, the trend of urbanization, etc. It derives lots of target domains, each of which contains very few target instances for training. In terms of the low-resource circumstance, we are encouraged to consider a few-shot adaptation mechanism for the CSP problem.

Recent few-shot learning research [6], [34], [39], [41], [43] promote to classify objects of the categories that never appear in training, provided with only few examples of each new class. Derived from this principle, few-shot domain adaption (FADA) [30] was raised to transfer a source model toclassify target images. However, target images in the FADA scenario are too few to support the cross-domain feature alignment compared with the conventional cases, owing to the image-level and pixel-wise imbalances between the source and the target data. Existing model performances are unstable and easily stuck in the bottleneck as their imbalance techniques heavily rely on the rebalance-sampling or rebalance-weighting manners. Although they may relieve the imbalance issue across domains but there are no evidence they would bridge the domain gap.

In this paper, we recast the pixel-level FSDA problem into Rubin causality [37] to refresh the traditional understanding of the cross-domain imbalance. Distinct from Pearl's structural causal model [32], Neyman-Rubin causal framework [14], [37] develops a system to estimate the treatment effect (i.e., causal effect) under the condition that the causation in Pearl's causal Bayesian network has already been identified. To this, given a premise of Rubin causality applied in our problem, we reconsider the feature as covariates (confounders) and the domain identity (*i.e.*, source or target domain) as a treatment selection, then achieving the pixel-level cross-domain feature alignment is equivalent with minimizing the average treatment effect (ATE) between a control group (the source domain) and a treatment group (the target domain). In this way, the imbalance domain adaptation may refer to solving the covariate balancing [17] issue between the groups that regularly appears in causal effect estimation, in which the imbalance domain gap could be minimized as long as the ATE converges to zero.

There have been many approaches invented to balance the covariates across distinct groups for causal estimation, where *propensity score* [20], [36] and *counterfactual matching* [25] are two categories that have been mostly adopted. Propensity score refers to predicting the preference of a given instance about which group (source or target domain in our circumstance) the instance is inclined to join. Counterfactual matching implies that given a data instance with its outcome in the control group, we are capable of finding another instance in the treatment group mostly resembling its appearance in the covariates, in order to replace its unobservable counterfactual counterpart for building the causal inference. Their spirits inspire Counterfactually Balancing Feature Alignment (CBFA) that we proposed to tackle the few-shot CSP problem in this paper. CBFA consists of an adversarial domain adaptation objective for pixel-level feature alignment, in which the optimal feature alignment implies that their ATE has converged to zero. In case of the cross-domain data imbalance, CBFA employs inverse propensity (score) weighting (IPW) scheme [9], [19] to reconfigure the adversarial learning objective, in which the propensity score of each instance is produced by the domain discriminator updated in the previous iteration. We employ two domain discriminators to promise the covariate balance for pixel-level and image-level training instances, respectively. Besides, CBFA further incorporates counterfactual matching to specify its discriminators, which both receive a source-target pair constructed for the counterfactual matching rule. The training strategy is alternatively executed to update between the coupled discriminator and the scene parser model to enable learning stability.

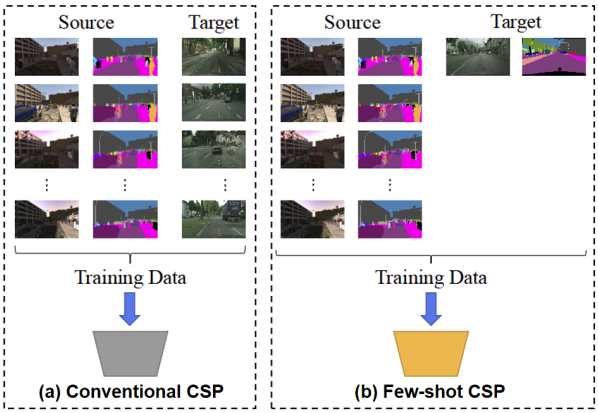


FIGURE 1. The comparison between the problem settings of conventional CSP and the proposed few-shot CSP. (a) is trained with supervised source data and a large amount of unsupervised target data. (b) is trained with supervised source data and very few supervised target data. Although conventional CSP can spare the cost of annotating a large amount of target data, but collecting them is also time-consuming and laborious. In contrast, the few-shot CSP only needs to collect a few target data and the cost of annotating them is completely acceptable.

The contributions are summarized as follows:

- We reconsider few-shot cross-domain semantic segmentation for scene parsing from a Rubin-causal view. It provides a new explanation for few-shot pixel-level domain adaptation to facilitate our solution.
- Based on covariate balance techniques in Rubin causality, we develop counterfactually balancing feature alignment (CBFA) by using propensity score and counterfactual matching to solve the few-shot cross-domain scene parsing problem. It derives a two-stage transfer framework to represent the image-level and pixel-level counterfactual inference in terms of few-shot domain adaptation.
- Extensive experiments and evaluations on two suites

of virtual-to-real scene parsing setups show that our proposed framework achieves superior performance in comparison to the state-of-the-art.

II. RELATED WORK

In this section, we provide some technical backgrounds behind this paper and discuss the relationship and the difference in terms of our work.

A. SEMANTIC SEGMENTATION FOR SCENE PARSING.

Scene parsing is a keen topic in computer vision on the foundation of semantic segmentation. The goal is to classify each pixel of an image into one of several predefined categories. With the development of deep learning, pixel-level prediction tasks like scene parsing and semantic segmentation have achieved great progress. Fully convolutional network (FCN) [22] pioneered to replace fully-connected layers (FC) by convolutional layers, and many successive techniques, dialuted convolution [50], large kernel [33], random-field predictor heads , and others [3] have been invented to further enhance the accuracy and efficiency. There are also several research [4], [47] that use the synthetic datasets based on rendering to handle the data annotation problem, as the labels are usually available directly from computers. Our work also involves synthetic datasets for training. It is more necessary to narrow the domain shift between the synthetic data domain and real-world data domain.

B. DOMAIN ADAPTATION.

Domain adaptation [11] was invented to mitigate the shift between the distributions of the source domain and the target domain. Thanks to the power of deep neural networks (DNNs), the modern emphasis has shifted to aligning features extracted from the networks in an end-to-end manner [8], [23], in which Adversarial domain adaptation [8] [13] is a common way to build the transfer bridge. Chen et al. [5], [53] proposed an elegant adversarial network to transfer multiple source domains with distribution and category shifts. Thereafter, a great deal of variants have been proposed with different loss functions [48] or classifiers [24], and some of them have been developed to solve cross-domain classification on pixels, *i.e.*, domain-adaptive semantic segmentation. For instance, [13] introduces a cycle-GAN backbone as a style transfer model to change the appearance of images to adapt to the target domain; [27] proposes a novel multi-level region expansion strategy to align both the image-level and pixel-level information; [49] incorporates the guidance from self-supervised depth estimation, etc. Compared with the existing works that incorporate massive unlabeled training data drawn from the target domain, our work only employ a few of target-labeled data to guide the transfer from the synthetic datasets, which relieves the burden to collect a large scale of training images in the real world.

C. FEW-SHOT LEARNING.

Few-shot learning aims to recognize novel visual categories from a limited amount of labeled training data. Recent few-shot learning literature has been proposed for image classification [41], [43] and semantic segmentation [6], [34]. They require amounts of labeled data drawn from old classes in the real world as they are unable to decrease the imbalance domain shift in the scenarios. More recently, some research for domain-adaptive few-shot learning [26] has been further proposed to solve few-shot image classification across domains, whereas our work differs from these approaches in two aspects: first, these approaches lean on classifying image-level labels rather than predicting the label for each pixel; second, their evaluations are based upon C-way K-shot meta-learning setup [7], [41], implying that they can not be reconfigured to suit our problem.

The literature mostly relevant to this research are few-shot domain adaption problems for image classification [28], [30], [46] and semantic segmentation [44], which both pair source-target data for each class to alleviate the imbalance. In contrast, our work derives the wisdom from Rubin causality to form a counterfactual inference grounding for our data balancing strategy across domains.

D. CAUSAL INFERENCE.

Causality has been investigated in the past few decades, where the research could be broadly categorized into two branches: *Structural Causal Model (SCM)* (Pearl's Causality) [32], [51] and *Potential Outcome (PO) Framework* (Rubin's Causality) [14], [25], [36], [37]. SCM starts from Bayesian Networks to develop a non-parametric analytical system in mathematics, which rigorously tells us a story about how an event, state, or object (a cause) contributes to the production of another event, state, or object (an effect). The most important function of SCM is causal identification, which judges the existence of causality between two nodes (events or objects) in a causal diagram when the diagram has been provided. In contrast, the PO framework was derived from random experiments, which pay more attention to estimating the difference between a pair of treatment and control groups. There have been massive studies that leverage SCM to solve diverse tasks in vision while seldom observe these problems from a lens of Rubin's causality. Instead, we prefer the PO framework as our analytical toolkit since the techniques of SCM do not solve the covariate balance that few-shot CSP refers to.

III. PIXEL-LEVEL DOMAIN ADAPTATION IN THE LANGUAGE OF RUBIN CAUSALITY

In this section, we first review pixel-level domain adaptation to formally understand the cross-domain scene parsing problem in many-shot and few-shot scenarios. We further provide an overview of PO framework and how it may reflect pixel-level adaptation from a view of average treatment estimation.

A. PROBLEM DEFINITION

Here we go to consider the task setups of pixel-level domain adaptation. More specifically, we have N_s and N_t training images drawn from the scenarios with regards to the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$, respectively. For each source image $\mathbf{x}_i \in \mathcal{S}$ or target image $\mathbf{x}_i \in \mathcal{T}$, $\{x_j^{(i)}\}_{j=1}^{n_i}$ denotes their pixels within. For a source image i, the pixels are given with their labels $\{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$ therefore someone would expect to train a pixel-level classifier with N_s source images for parsing the scenes in $\mathcal{S}$, however, it fails to segment the images drawn from the target domain due to the distributional shift between $\mathcal{S}$ and $\mathcal{T}$. Unsupervised domain adaptation approaches incorporate a huge amount of training target images in the absence of labels (*i.e.*, each target image i performs as an unlabeled pixel set $\{(x_j^{(i)})\}$) for building a constraint to diminish the domain shift. The constraint changes according to the specific adaptation strategies, *e.g.*, maximum mean discrepancy [24], adversarial learning [8], under the condition of sufficient target images for training.

The unsupervised setup for pixel-level domain adaptation needs massive images collected in each target domain, which could be very difficult to satisfy in many situations. Instead, this work considers the few-shot setup for pixel-level domain adaptation that only requires few-shot, even single-shot target image ($N_s >> N_t$) with labels in pixels, *i.e.*, $\forall \mathbf{x}_i \in \mathcal{T}$ we have $\{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$ for each target scenario from N cities.

B. FEW-SHOT ADAPTATION IN PO FRAMEWORKS

Since the sparsity of target images breaks the statistical law in domain adaptation strategies employed in the unsuperivsed setup, few-shot domain adaptation is very challenging. In this paper, we intend to overcome this difficulty by reconsidering pixel-level domain adaptation in the language of Rubin causality [37]. It reformulates domain adaptation into average treatment estimation (ATE), in which the cross-domain data imbalance issue in the few-shot setup may refer to the covariate balance problem regularly found in the community of treatment effect estimation.

In particular, we first review the background of the Potential Outcome (PO) Framework. Let's consider a treatment T that we are interested in how it causes the outcome y in terms of individual $\mathbf{x}$. For simplicity, we consider T is binary so that $T = 1$ implies the treatment has been taken, $T = 0$ then vice and versa. It is obvious that the causal effect of z towards y can be estimated if given an individual instance $\mathbf{x}$, we are able to simultaneously observe the value of y when $T = 0$ and $T = 1$, respectively. In this way, the average causal effect in the population of interest is presented as follows:

$$\gamma_{\mathrm{ATE}} = \mathbb{E}\gamma = \mathbb{E}_{\mathbf{x}\in\mathcal{G}}\mathbb{E}\Big(y(T=1) - y(T=0)|\mathbf{x}\Big), \quad (1)$$

where $\gamma = \Big(y(T=1) - y(T=0)\Big)$ denotes individual treatment effect (ITE) of an instance, γ_{ATE} denotes the average treatment effect (ATE) in terms of the population $\mathcal{G}$, and $y(T=j)$ represents the potential outcome when the treatment T was taken as j. Note that, ITE inference is intractable since it is impossible to observe all potential outcomes given individual $\mathbf{x}$ in practice unless we have a time machine to go back and choose the treatment distinct from what we have taken at the first time. It is the fundamental problem of causal inference and the potential outcome without observation is well-known as *counterfactual*.

In spite of the impossibility of inferring ITE for each individual, we are able to infer γ_{ATE} for some group of the population under some assumptions. Specifically, we assume that provided a group of the population, each variable of potential outcome is independent with treatments (*Unconfoundness*); there exists the non-zero probability for each individual to receive any of the treatments (*Possibility*) ; and the treatment taken by one does not affect another's decision to choose the treatment (*stable unit-treatment value, SUTVA*)[1]. Under these circumstances, γ_{ATE} can be decomposed as

$$\begin{aligned}\gamma_{\mathrm{ATE}} &= \mathbb{E}_{\mathbf{x}}\mathbb{E}\Big(y(T=1) - y(T=0)|\mathbf{x}\Big) \\ &= \mathbb{E}_{\mathbf{x}}\Big(y(T=1)|\mathbf{x}\Big) - \mathbb{E}_{\mathbf{x}}\Big(y(T=0)|\mathbf{x}\Big) \\ &= \mathbb{E}_{\mathbf{x}}\Big(\mathbb{E}_y \frac{\Pr(y, T=1, \mathbf{x})}{\Pr(T=1|\mathbf{x})} y\Big) - \mathbb{E}_{\mathbf{x}}\Big(\mathbb{E}_y \frac{\Pr(y, T=0, \mathbf{x})}{\Pr(T=0|\mathbf{x})} y\Big),\end{aligned} \quad (2)$$

where $\Pr(y, T=j, \mathbf{x})$ denotes the joint distribution of outcome, treatment, and individual in the population; $\Pr(T = j|\mathbf{x})$ implies that given individual $\mathbf{x}$, its probabilistic preference to join the treatment or the control group. The equations demonstrate that even if the counterfactuals are absent, we are able to measure the causal effect from T to y by separately checking the average factual outcomes of the treatment and the control groups. Since the distributions with regards to the groups are different, the estimator should be balanced by the inverse of $\Pr(T = j|\mathbf{x})$.

Henceforth we reconsider domain adaptation from the lens of the adjustment formula in Eq.2. We reconsider the feature couple of a pixel and the image it belongs to (*i.e.*, $(x, \mathbf{x})$) as the covariate of an individual; and consider the domain identity (*i.e.*, which domain the pixel belongs to) as a treatment. In this sense, γ_{ATE} implies how much the variation of domain identity in the population would alter the model performance to predict y. Interestingly, the optimal distribution alignment drives the source and target features to obtain almost identical performance, hence the causal effect of domain identity converges to zero ($\gamma_{\mathrm{ATE}} \to 0$). It means that minimizing the absolute value of γ_{ATE} results in the optimal feature alignment. To this, the few-shot adaptation becomes the mirror of the data imbalance between the treatment and control groups, which can be resolved by resembling the covariate balancing spirit in existing methods.

IV. METHODOLOGY

In the PO framework, there are two common ways of alleviating the statistical imbalance between the treatment and

[1]Since our work only takes causality to inspire our algorithm, we prefer to skipping the principled interpretation in this paper. The formal definitions for the three assumptions can be found in [14].

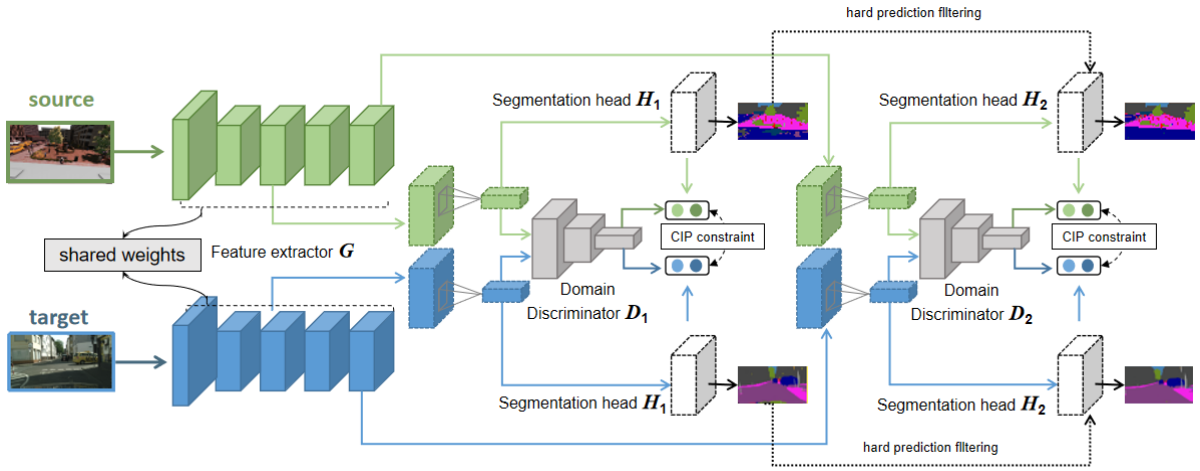


FIGURE 2. Framework overview. In each iteration, our framework concurrently receive source and target images to obtain their output predictions and features. The first-stage (S_1) outputs an auxiliary mask that is used to calculate the first segmentation loss ($\mathcal{L}_{seg1}$) and also can be used to filter the original label. The second-stage (S_2) outputs a semantic mask that is used to calculate the second segmentation loss ($\mathcal{L}_{seg2}$). D_1 and D_2 are used to distinguish which domain the input belongs to, which play the roles of the propensity score simulators also. The adversarial loss is implemented by training two discriminators, then alternatively taking their predictions to construct CIP constraint to update the scene parser.

the control groups. The first one is *inverse propensity score weighting* (IPW) that employs the propensity score to balance the factual outcome between the groups. It could be understood as the inverse of $\Pr(T = j|\mathbf{x})$ in Eq.2, which is implemented by training a treatment classifier to judge which group a given individual may come from. In terms of domain adaptation, the propensity score replays the role of domain discriminator that is trained to classify which domain a pixel-level feature comes from. The second is *counterfactual matching*, which is derived from the spirit of simulating the unobserved counterfactual given each observed individual.

In this section, we propose a domain adaptation framework that incorporates IPW and counterfactual matching, for efficiently tackling the few-shot data imbalance across domains.

A. BASED ADAPTATION FRAMEWORK

In order to model the propensity score by adversarial domain discriminators, we need to propose a based adversarial pixel-level domain adaptation framework for the few-shot CSP problem. It is composed of three subnetworks: a *Scene Feature Extractor* G, a *Propensity Domain Discriminator* D, and a *Semantic Segmentation Head* C. The scene feature extractor G is the key part to learn the domain-invariant representation, which is restricted by the cross-domain data imbalance. The domain discriminator D is adversarially trained to bridge the domain gap with regard to the cross-domain imbalance from a view of causality. The segmentation head C receives the output of G to predict the category on each pixel. The framework is illustrated in Figure.2.

1) Adversarial loss

In terms of the multi-scale granularity behind pixel-level domain adaptation, we promote a two-stage architecture for the feature extractor G to juggle the merits of low-level and high-level feature maps then make pixel-level predictions on images. Accordingly our domain discriminator D including two adversarial learning stages (*i.e.*, D_1 and D_2) with the cascading feature outputs judges which domain a pixel-level feature ($G_1(x;\mathbf{x})$) and its neighbor region ($G_2(x;\mathbf{x})$) comes from, respectively.

Specifically, we employ adversarial loss below to train the domain discriminator. The summation of losses across stages is proposed to train G as follows:

$$\begin{aligned}\mathcal{L}_{adv} = &\sum_{\forall \mathbf{x}\in\mathcal{S}}\sum_{\forall x\in\mathbf{x}}(1-z_{\mathbf{x}})\log D_1(G_1(x;\mathbf{x})) \\ &+\sum_{\forall \mathbf{x}\in\mathcal{T}}\sum_{\forall x\in\mathbf{x}} z_{\mathbf{x}}\log D_1(G_1(x;\mathbf{x})) \\ \lambda_{adv}\Big(&\sum_{\forall \mathbf{x}\in\mathcal{S}}\sum_{\forall x\in\mathbf{x}}(1-z_{\mathbf{x}})\log D_2(G_2(x;\mathbf{x})) \\ &+\sum_{\forall \mathbf{x}\in\mathcal{T}}\sum_{\forall x\in\mathbf{x}} z_{\mathbf{x}}\log D_2(G_2(x;\mathbf{x}))\Big)\end{aligned} \tag{3}$$

where $z_{\mathbf{x}}$ denotes the domain identity of an image $\mathbf{x}$ so source and target domains correspond to 0 and 1, respectively. λ_{adv} denotes the hyper-parameter to balance the adversarial loss at different levels. The minimization of $\mathcal{L}_{adv}$ drives the discriminator to categorize pixels into their domains, in contrast, the maximization turns to fuse the cascading features across domains. We train the discriminators D_1 and D_2 saperately to stablize the adversarial adaptation process.

2) Segmentation loss

Observe that, each training image contains its category mask (each pixel with its label), so it is very important to deriving a based supervised training algorithm for G and C. We propose a two-stage learning objective in terms of the cascading output. In particular, our architecture predicts the segmentation masks from high-level features, constructing the first stage segmentation loss ($\mathcal{L}_{seg1}$) through classifying pixels with

their original labels y:

$$\mathcal{L}_{seg1} = -\sum_{\forall \mathbf{x}\in\mathcal{S}\cup\mathcal{T}}\sum_{\forall x\in\mathbf{x}} y\log H_1\big(G_1(x;\mathbf{x})\big). \tag{4}$$

Then we take the predicted masks with a threshold β to filter out the simple pixel labels y. The prediction confidence of a pixel higher than β implies that the pixel is easy to recognize and thus, our framework prefers to omitting its segmentation loss at the second stage. It encourages the second-stage focus more on the difficult pixels, *e.g.*, those lying around the edges between objects.

$$\mathcal{L}_{seg2} = -\sum_{\forall \mathbf{x}\in\mathcal{S}\cup\mathcal{T}}\sum_{\forall x\in\mathbf{x}} \hat{y}\log H_2\big(G_2(x;\mathbf{x})\big), \quad s.t. \quad \hat{y} = y\odot 1\Big(H_1^{(y)}\big(G_1(x;\mathbf{x})\big)>\beta\Big). \tag{5}$$

We combine $\mathcal{L}_{seg1}$ and $\mathcal{L}_{seg2}$ to derive a supervised learning loss for the semantic segmentation task:

$$\mathcal{L}_{seg} = \mathcal{L}_{seg1} + \lambda_{seg}\mathcal{L}_{seg2}, \tag{6}$$

in which λ_{seg} balances the powers between $\mathcal{L}_{seg1}$ and $\mathcal{L}_{seg2}$.

B. COUNTERFACTUAL BALANCING FEATURE ALIGNMENT

The aforementioned adversarial model trained by balance resampling/reweighting can help to prevent the trained model overfitting to few target data. However, no evidence ensures the domain adaptation since the existing balance resampling and reweighting strategies never promised to bridge the domain gap, so the trained models are still haunted by the negative transfer effect. Aiming to tackle this inconsistency, we introduce new balanced resampling or reweighting methods from the wisdom of covariate balancing the PO framework.

1) Cross-domain Inversed Propensity Constraint

In Eq.2, we may observe that the treatment preference of each individual can be used to obtain the unbiased estimation for ATE, as its minimization refers to the optimal feature alignment across domains. In this case, we rewrite the adjustment formula to derive a novel constraint for supervised learning. Specifically, since we have connected domain adaptation with minimizing γ_{ATE}, in which $\Pr(y, T=j, \mathbf{x})(j\in\{0,1\})$ refers to our segmentation head output to predict the label on pixels, the propensity score $\Pr(T=j|\mathbf{x})$ can be achieved by incorporating the discriminator output on these pixels. To this, given a specific category $y\in\mathcal{Y}$ ($\mathcal{Y}$ denotes the label space), there is a pair of analogies for the pixels in the source and target domains:

$$\begin{aligned}\frac{\Pr(y,T{=}0,\mathbf{x})}{\Pr(T=0|\mathbf{x})} &\rightarrow \frac{H^{(y)}\big(G(x;\mathbf{x}_s)\big)}{1-D(G(x;\mathbf{x}_s))};\\ \frac{\Pr(y,T{=}1,\mathbf{x})}{\Pr(T=1|\mathbf{x})} &\rightarrow \frac{H^{(y)}\big(G(x;\mathbf{x}_t)\big)}{D(G(x;\mathbf{x}_t))}.\end{aligned} \tag{7}$$

By executing the analogies for the two-stage outputs, we may present our *Cross-domain Inversed Propensity Constraint*:

$$\begin{aligned}\mathcal{L}_{\mathrm{CIP}}{=}\mathbb{E}_y\Bigg(\Big\|\Big(\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{S}}\frac{H_1^{(y)}\big(G_1(x;\mathbf{x})\big)}{1-D_1(\bar{G}_1(x;\mathbf{x}))}\\ -\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{T}}\frac{H_1^{(y)}\big(G_1(x;\mathbf{x})\big)}{D_1(\bar{G}_1(x;\mathbf{x}))}\Big)\Big\|+\\ \Big\|\Big(\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{S}}\frac{H_2^{(y)}\big(G_2(x;\mathbf{x})\big)}{1-D_2(\bar{G}_2(x;\mathbf{x}))}-\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{T}}\frac{H_2^{(y)}\big(G_2(x;\mathbf{x})\big)}{D_2(\bar{G}_2(x;\mathbf{x}))}\Big)\Big\|\Bigg).\end{aligned} \tag{8}$$

Note that we only take the discriminators' outcome to balance the adaptation therefore the feature extractors would not be updated through the discriminators in this stage. The CIP constraint is proposed to minimize $||\gamma_{\mathrm{ATE}}||$ in terms of each category group. It is because the imbalance case may be very different across categories. The classwise constraint helps to restrict the imbalance within each class, otherwise, the correct prediction outcome is switched to the wrong category, thus, chasing for the minimization over the whole population.

In practice, the propensity score near 0 and 1 would cause a large variation to the counterfactual inference estimator. To this, we design a truncation function ϕ:

$$\phi_\nu(x) = \max\big(\nu, \min(x, 1-\nu)\big), 0<\nu\ll 1, \tag{9}$$

in which we configure $\nu == 0.05$ (sightly larger than 0) and the truncation function $\phi(x;\nu)$ would pull back the scalar x into the range $[\nu, 1-\nu]$ so as to prevent the estimator from the unacceptable variation. We take ϕ to modify Eq.8:

$$\begin{aligned}\mathcal{L}_{\mathrm{CIP}}^{(\nu)}{=}\mathbb{E}_y\Bigg(\Big\|\Big(\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{S}}\frac{H_1^{(y)}\big(G_1(x;\mathbf{x})\big)}{\phi_\nu\big(1-D_1(\bar{G}_1(x;\mathbf{x}))\big)}\\ -\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{T}}\frac{H_1^{(y)}\big(G_1(x;\mathbf{x})\big)}{\phi_\nu\big(D_1(\bar{G}_1(x;\mathbf{x}))\big)}\Big)\Big\|\\ +\Big\|\Big(\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{S}}\frac{H_2^{(y)}\big(G_2(x;\mathbf{x})\big)}{\phi_\nu\big(1-D_2(\bar{G}_2(x;\mathbf{x}))\big)}\\ -\mathbb{E}_{x\in\mathbf{x},\mathbf{x}\in\mathcal{T}}\frac{H_2^{(y)}\big(G_2(x;\mathbf{x})\big)}{\phi_\nu\big(D_2(\bar{G}_2(x;\mathbf{x}))\big)}\Big)\Big\|\Bigg).\end{aligned} \tag{10}$$

Since the inverse propensity score weighting assumes that the propensity score has been determined, the CIP constraint is only used to balance the training for semantic segmentation. As for the discriminators D_1, D_2, we propose another loss reweighting approach inspired by counterfactual matching to prevent their training from the imbalance issue.

2) Counterfactual Adversarial Contrastive Loss

The goal of training a discriminator is categorizing the pixels of images into the domain they belong to. It is not relevant to the equivalence between ATE and domain adaptation, thus, could not be solved by covariate balancing. The supervised training for the discriminators may heavily suffer from the imbalance problem because there are too few target images to support the binary classification in fair, in which the domain with the majority of samples dominates the training and disable the discriminator. Thanks to the progress of imbalance

learning [16], [21], we are able to propose a new solution to suit our problem.

We reconsider the role of discriminator losses in reducing the distribution shift for domain adaptation:

$$\mathcal{L}_{adv} = \sum_{\forall \mathbf{x}\in\mathcal{S}} \sum_{\forall x\in\mathbf{x}} (1 - z_{\mathbf{x}}) \log D(G(x;\mathbf{x})) + \sum_{\forall \mathbf{x}\in\mathcal{T}} \sum_{\forall x\in\mathbf{x}} z_{\mathbf{x}} \log D(G(x;\mathbf{x})) \tag{11}$$

In the loss construction, each instance only belongs to the source or target, which is extremely different in their magnitude. To this, we introduce contrastive learning to refine the objective above.

$$\mathcal{L}_{con}(D,G) = \mathcal{L}_{adv} + \sum_{\substack{\forall \mathbf{x}\in\mathcal{T},\\ \forall \mathbf{x}'\in\mathcal{S}}} \sum_{\forall x,x^+\in\mathbf{x},} -\log \frac{\exp\left(f_{D,G}^{\top}(x;\mathbf{x}) f_{D,G}(x^+;\mathbf{x})\right)}{\exp\left(f_{D,G}^{\top}(x;\mathbf{x}) f_{D,G}(x^+;\mathbf{x})\right) + \sum_{\forall x^-\in\mathbf{x}'}^{m} \exp\left(f_{D,G}^{\top}(x;\mathbf{x}) f_{D,G}(x^-;\mathbf{x}')\right)} \tag{12}$$

in which $f_{D,G}(x;\mathbf{x})$ denotes the pixel-level feature extracted from the last layer ahead of the discriminator output. To this, given each feature $f_{D,G}(x;\mathbf{x})$ there is a tuple ($f_{D,G}(x;\mathbf{x})$, $f_{D,G}(x^+;\mathbf{x})$, $\{f_{D,G}(x^-;\mathbf{x}')\}_{x^-\in\mathbf{x}'}^{m}$) constructed as a training instance that simultaneously includes source and target data. It relieves the training data imbalance between source and target domains.

We further propose a counterfactual sampling technique to improve the efficiency of tackling the imbalance. As we know in Eq.12, provided each target pixel-level feature $f_{D,G}(x;\mathbf{x})$, there are only m source pixel-level features chosen to construct the loss. Resembling the spirit of counterfactual matching, we prefer to select the m source features mostly similar with $f_{D,G}(x;\mathbf{x})$:

$$C(\mathbf{x}_t,\mathbf{x}_s) = \underset{\substack{A'\subset\mathbf{x}_s,\\ |A'|=m}}{\arg\max} \sum_{x_s\in A'} \exp\left(\frac{1}{|\mathbf{x}_t|} \sum_{x\in\mathbf{x}_t} f_{D,G}^{\top}(x;\mathbf{x}_s) f_{D,G}(x_s;\mathbf{x}_s)\right), \tag{13}$$

where $|\mathbf{x}_t|$ indicates the number of pixels in the target image $\mathbf{x}_t$, and we may refine the Eq.12 into

$$\mathcal{L}_{con}^{(\mathrm{CF})}(D,G) = \mathcal{L}_{adv} + \sum_{\substack{\forall \mathbf{x}\in\mathcal{T},\\ \forall \mathbf{x}'\in\mathcal{S}}} \sum_{\forall x,x^+\in\mathbf{x},} -\log \frac{\exp\left(f_{D,G}^{\top}(x;\mathbf{x}) f_{D,G}(x^+;\mathbf{x})\right)}{\exp\left(f_{D,G}^{\top}(x;\mathbf{x}) f_{D,G}(x^+;\mathbf{x})\right) + \sum_{\substack{\forall x^-\in\\ C(\mathbf{x},\mathbf{x}')}} \exp\left(f_{D,G}^{\top}(x;\mathbf{x}) f_{D,G}(x^-;\mathbf{x}')\right)} \tag{14}$$

This motivation is widely accepted in hard negative sample mining [40]. We tend to interpret it from a causal lens since the technique indeed corresponds to CIP constraint. In this way, the other goal of discriminator is to simulate a propensity score that guides the causality-inspired objective (Eq.10) instead of disambiguating source and target features, which is known as usual. Our empirical study exhibits its superiority compared with normal adversarial DA approaches.

Algorithm 1 Counterfactual Balancing Feature Alignment (CBFA)

Input: N_s images drawn from the source domain $\mathcal{S}$ and N_t images drawn from the target domain $\mathcal{T}$, which are provided with their segmentation masks as pixelwise labels and $N_s >> N_t$ implies that target images are scarce in training; two-stage adaptation frameworks, including pre-trained feature extractors G_1, G_2; label predictors H_1, H_2; and domain discriminators D_1, D_2.

Output: Well-trained feature extractors G_1^*, G_2^* and label predictors H_1^*, H_2^*.

1: **for** episode = 1, E **do**
2: Get a sequence of mini-batches constructed by source images and target images that goes through all source images and repeatedly sampling target images, in which the number of source and target images are equal.
3: **for** each mini-batch **do**
4: Construct a pixelwise loss for Eq.15: given each pixel of a target image, CBFA chooses m source pixels across all source images in the mini-batch that satisfy Eq.13, then construct Eq.14 and Eq.15.
5: Train domain discriminators D_1, D_2 and fine-tune feature extractors G_1, G_2 with a reversed gradient layer [8].
6: Freeze the parameters of the discriminators D_1, D_2, then construct the CIP constraint based on Eq.10.
7: Using Eq.6 to update the parameters of feature extractors G_1, G_2 and predictors H_1, H_2 with the CIP constraint.

C. SUMMARY

The training objective for our model could be summarized as two objectives that we alternatively execute them to learn an ideal two-stage scene parser. We first adopt a pre-training strategy in [56]. Afterwards, we incorporate the adversarial constrastive loss in Eq.14 to learn a two-stage discriminator to approximate the propensity score:

$$\min_{D_1,D_2,G_1,G_2} \mathcal{L}_{con}^{(\mathrm{CF})}(D_1,G_1) + \lambda_{adv}\mathcal{L}_{con}^{(\mathrm{CF})}(D_2,G_2), \tag{15}$$

in which we employ λ_{adv} to balance the training between the first and second discriminators. Given this, we take the trained discriminators to construct the CIP constraint, deriving the few-shot transfer learning for semantic segmentation to scenes:

$$\min_{G_1,G_2} \mathcal{L}_{seg}(G_1,G_2) + \mathcal{L}_{\mathrm{CIP}}^{(\nu)}(G_1,G_2,D_1,D_2). \tag{16}$$

The stochastic training process has been summarized into Algorithm.1.

Complexity discussion. Suppose the model trained with E epoches until it converges and we consider the computational complexity. For each epoch, we consider there are N_s source images and N_t target images for training, then given m_s source images for each training batch, the number of iterations for each epoch is determined by $[\frac{N_s}{m_s}]$+1 due to N_s>N_t. For each iteration, CBFA first adversarially updates the discriminators and feature extractors, then constructs the CIP constraint with our discriminators, eventually fine-tuning the parameters of the feature extractors and predictor. Since most of the operations behind the mini-batch have been opacified and automatically paralleled by the Pytorch platform, we only discuss the complexity outside the mini-batch forward

TABLE 1. Mean IoU (mIoU) results of GTA5-to-CITYSCAPES. We compare our results with the simple baselines FT, JT, and two competitive state-of-the-art methods, *i.e.*, FSDA and PixDA. FT denotes fine-tuning. JT denotes jointly training the scene parser using supervised source data and few supervised target data.

GTA5 → Cityscapes																				
Base	Method	road	sidewalk	building	wall	fence	pole	light	sign	veg	terrain	sky	person	rider	car	trunk	bus	train	motor	bike	mIoU
Resnet-101	Source-only	85.94	40.34	81.4	24.19	16.63	26.58	28.3	15.04	79.75	27.5	83.47	49.81	20.91	71.97	22.11	21.03	0.01	18.41	24.05	38.76
	JT (1-shot)	90.79	56.41	81.57	**33.85**	18.45	30.64	29.31	28.73	83.66	37.27	84.71	49.69	17.52	69.5	24.68	27.1	1.98	16.47	26.15	42.55
	FT (1-shot)	93.41	60.58	82.72	21.19	23.93	30.66	27.04	29.57	84.45	39.13	66.37	52.32	20.44	84.54	35.7	29.65	1.05	12.81	33.76	43.65
Resnet-101	PixDA [44] (1-shot)	93.33	59.19	**84.12**	28.33	28.05	**37.15**	38.44	32.05	**85.85**	**45.12**	**86.18**	63.08	26.79	85.76	35	**50.69**	8.5	12.81	55.02	51.16
	FSDA [56] (1-shot)	94.49	62.67	82.76	25.03	19.23	32.59	28.83	36.9	84.71	39.81	83.55	54.87	25.19	84.87	31.85	29.39	0.0	16.8	48.56	46.43
	Ours (1-shot)	**95.19**	**62.71**	83.96	31.01	**29.23**	35.59	**39.11**	**37.19**	85.76	44.88	85.78	**64.17**	**29.16**	**86.08**	**35.85**	47.93	**10.2**	**18.9**	**56.85**	**51.56**
	JT (2-shot)	92.03	53.0	82.93	32.69	24.66	32.06	30.46	30.75	84.45	41.44	84.59	52.4	13.22	80.98	**42.19**	37.24	1.43	11.26	42.34	45.8
	FT (2-shot)	94.51	62.45	83.62	23.88	28.17	32.05	31.72	30.47	84.84	39.97	78.62	53.14	8.76	85.09	34.58	28.3	16.66	14.19	40.76	45.88
Resnet-101	PixDA [44] (2-shot)	94.66	64.5	84.98	27.48	**30.63**	**38.5**	37.69	38.85	85.95	46.21	85.26	**61.91**	19.36	85.06	37.27	**47.12**	**25.64**	32.45	56.33	52.78
	FSDA [56] (2-shot)	94.07	61.62	84.68	35.9	25.56	34.29	34.51	37.83	86.41	43.65	85.07	55.97	17.91	84.54	39.74	36.03	2.22	21.94	51.37	49.12
	Ours (2-shot)	**95.26**	**64.02**	**85.11**	**37.13**	30.56	36.92	**41.17**	**40.06**	**86.69**	**46.67**	**87.12**	59.97	**25.47**	**85.33**	41.04	46.53	22.11	**34.91**	**57.71**	**53.88**
	JT (3-shot)	92.43	54.1	83.0	33.81	24.46	32.49	34.29	35.76	84.91	39.82	86.13	53.72	25.5	82.27	31.04	30.89	12.38	19.52	45.38	47.47
	FT (3-shot)	94.65	63.45	84.72	26.2	25.69	34.64	34.45	38.06	86.29	41.93	88.55	56.28	24.52	86.04	32.1	7.52	20.7	26.42	44.78	48.26
Resnet-101	PixDA [44] (3-shot)	94.26	62.22	85.68	31.7	26.57	38.55	**43.55**	44.11	**86.9**	**48.88**	**88.85**	64.24	35.6	**87.72**	41.16	47.21	14.96	33.37	**59.78**	54.54
	FSDA [56] (3-shot)	94.34	64.64	85.25	35.63	27.35	36.13	36.93	40.12	86.25	45.17	85.85	58.3	31.09	83.39	31.82	29.28	11.25	29.2	56.42	50.97
	Ours (3-shot)	**95.32**	**65.82**	**86.13**	**38.16**	**31.26**	**39.39**	41.96	**46.03**	86.77	46.17	87.51	**66.03**	**38.68**	85.28	**44.24**	**46.53**	**24.56**	**34.91**	57.62	**55.91**
	JT (4-shot)	92.67	54.58	83.27	29.0	25.25	34.16	32.36	34.87	84.6	41.82	82.01	54.71	26.97	80.7	36.8	28.49	3.16	20.33	48.1	47.04
	FT (4-shot)	94.35	65.17	84.54	28.67	**30.52**	34.82	32.99	39.72	85.78	44.85	77.77	56.45	25.08	86.05	13.17	31.72	0.03	7.76	45.3	46.57
Resnet-101	PixDA [44] (4-shot)	95.24	67.28	**86.05**	30.5	28.87	40.14	40.09	**47.71**	**86.95**	47.95	89.34	64.6	41.32	88.12	47.53	47.23	8.54	30.03	58.19	55.04
	FSDA [56] (4-shot)	94.42	62.99	85.21	38.52	29.34	35.36	33.16	45.07	86.73	45.11	88.83	59.18	32.64	85.6	40.65	29.17	0.0	21.08	53.64	50.88
	Ours (4-shot)	**95.82**	**69.19**	85.91	**40.28**	30.31	**41.11**	**41.16**	45.97	86.77	**48.15**	**89.63**	**66.1**	**41.4**	**88.66**	**48.86**	**49.01**	**21.44**	**28.79**	**58.44**	**56.68**
	JT (5-shot)	92.78	59.9	83.72	31.11	25.92	32.14	34.69	42.54	84.63	40.51	84.36	54.16	26.09	79.25	39.23	43.38	2.71	8.86	49.34	48.18
	FT (5-shot)	95.26	65.53	84.91	22.5	28.87	34.91	33.38	42.07	86.36	42.83	83.86	55.33	27.08	86.43	**47.8**	43.64	3.91	23.08	44.14	50.1
Resnet-101	PixDA [44] (5-shot)	95.03	66.94	**86.33**	26.64	24.3	40.37	**40.84**	**52.52**	86.36	42.83	87.18	64.62	34.78	**88.4**	45.31	49.01	19.91	34.36	**60.39**	55.56
	FSDA [56] (5-shot)	94.61	65.12	85.57	33.68	27.25	37.31	36.75	48.63	86.79	47.94	87.56	60.51	32.12	85.68	41.18	45.27	14.75	32.1	54.6	53.55
	Ours (5-shot)	**95.61**	**68.24**	85.37	**38.48**	**31.42**	**42.29**	40.7	51.66	**87.13**	**48.89**	**88.97**	**66.59**	**42.71**	88.62	46.24	**50.04**	**24.47**	**36.61**	58.87	**57.52**

inference and backward parameter update. To this, the complexity amounting is dominated by Eq.13 and the process of building the CIP constraint. Given m_p pixels at most for each source image, it basically consumes $\mathcal{O}(m_p \log(m_p))$ to achieve the differentiable proxy for ranking [2], which could be parallelly executed across different images within a mini-batch. The CIP constraint is built in terms of the specific categories we consider, and the ATE computation for each class is parallelly executed across pixels and images for each batch. To this end, given $|\mathcal{Y}|$ as the number of categories, it requires $\mathcal{O}\Big(E([\frac{N_s}{m_s}]+1)(\mathcal{O}(m_p \log(m_p))+|\mathcal{Y}|)\Big)=\mathcal{O}\Big(E([N_s \log(m_p) + \frac{N_s}{m_p}|\mathcal{Y}|)\Big)$ as the computation overhead for the training process.

V. EXPERIMENTS

In this section, we evaluate our method on the synthetic-to-real transfer learning tasks SYNTHIA-to-CITYSCAPES and GTA5-to-CITYSCAPES. Our entire training contains two phases: the first is adversarial domain adaptation, *i.e*, training the feature extractor and discriminators jointly, and the second is training the feature extractor independently using target data with CIP constraint. We gradually reduce the domain shift by performing the first training phase and the second training phase alternately in an end-to-end manner.

A. EXPERIMENT SETUP

CITYSCAPES is a real-world image dataset which consists of 2975 images in the training set, and 500 images in the verification set. There are 18 sub-folders in the training set representing 18 different cities. The resolution of the images is 2048×1024. And the pixel-level labels of 19 semantic categories are provided. Our few-shot target domain images are randomly selected from the training set. Finally, we use the verification set to test the trained model.

SYNTHIA is a synthetic dataset of urban scenes, it contains 9400 images compatible with the CITYSCAPES annotated categories. Similar to [47], in all the experiments with SYNTHIA as the source domain, we evaluate on the CITYSCAPES verification set with 13 categories.

GTA5 is a synthetic dataset containing 24966 images with the resolution of 1914×1052. The images are from a video game based on the city of Los Angeles. There are 19 semantic categories compatible with CITYSCAPES dataset. In all the experiments with GTA5 as the source domain, we evaluate all baselines on the CITYSCAPES verification set with 19 categories.

We consider several baselines to illustrate the validity of our CFBA. The first is Source Only, i.e., training the based scene parser only with the source data. Joint Training (JT) is a model trained by concatenating source and target images for

TABLE 2. Mean IoU (mIoU) results of Synthia-to-CITYSCAPES. We compare our results with the simple baselines FT, JT, and two competitive state-of-the-art methods, *i.e.*, FSDA and PixDA. FT denotes fine-tuning. JT denotes jointly training the scene parser using supervised source data and few supervised target data.

Synthia → Cityscapes															
Base	Method	road	sidewalk	building	light	sign	vegetation	sky	person	rider	car	bus	motocycle	bicycle	mIoU
Resnet-101	Source-only	60.12	22.38	66.56	4.63	7.15	75.22	76.58	33.35	10.55	54.53	5.63	1.25	17.65	33.51
	JT (1-shot)	91.24	53.18	79.1	14.81	30.97	82.61	81.8	49.81	16.08	78.67	15.93	5.23	42.47	49.38
	FT (1-shot)	91.53	53.1	80.04	12.84	26.78	83.49	81.9	52.8	9.33	78.36	21.18	3.44	46.82	49.35
Resnet-101	PixDA [44] (1-shot)	94.23	**60.84**	82	22.5	**37.69**	84.05	85.04	59.03	20.63	**84.32**	31.37	8.84	56.46	55.92
	FSDA [56] (1-shot)	92.74	56.54	82.06	17.41	32.95	84.95	84.55	56.12	23.86	82.02	25.59	8.36	52.27	53.8
	Ours (1-shot)	**94.98**	59.67	**82.36**	**23.21**	36.91	**85.55**	**85.99**	**60.21**	**25.39**	83.83	**32.91**	**10.16**	**57.17**	**56.79**
	JT (2-shot)	92.03	54.31	80.69	14.19	27.68	82.52	83.18	50.94	15.98	78.3	16.93	7.54	46.27	50.04
	FT (2-shot)	90.69	51.36	80.43	16.85	25.71	83.55	78.71	54.56	12.09	79.98	15.05	6.4	47.42	49.45
Resnet-101	PixDA [44] (2-shot)	94.24	61.84	83.12	**22.9**	36.94	84.37	84.24	59.43	17.37	**86.23**	36.66	10.16	**57.41**	56.53
	FSDA [56] (2-shot)	92.68	55.44	82.25	24.54	35.43	85.08	83.09	57.01	14.95	79.87	24.47	**17.91**	53.02	54.29
	Ours (2-shot)	**94.98**	**62.42**	**83.41**	21.4	**38.23**	**86.08**	**85.69**	**60.41**	**24.95**	84.97	**37.74**	16.98	57.07	**58.03**
	JT (3-shot)	93.43	58.56	82.33	29.98	35.64	83.75	86.72	52.71	19.3	81.62	7.05	6.36	51.49	53.0
	FT (3-shot)	92.92	57.3	82.95	19.04	30.76	84.47	86.49	57.16	27.9	80.52	8.5	13.39	54.12	53.5
Resnet-101	PixDA [44] (3-shot)	94.36	62.62	84.05	32.39	38.91	**85.44**	88.18	**61.47**	29.33	85.74	23.67	12.27	61.29	58.44
	FSDA [56] (3-shot)	93.96	60.97	83.84	27.95	40.46	85.83	**88.27**	59.44	28.58	84.27	11.6	15.38	58.59	56.86
	Ours (3-shot)	**95.09**	**63.97**	**84.15**	**34.95**	**40.46**	85.01	87.96	60.94	**31.6**	**86.11**	**25.6**	**15.82**	**62.46**	**59.67**
	JT (4-shot)	94.7	64.52	82.88	28.11	38.33	84.5	87.0	55.88	31.91	81.97	6.73	2.7	52.69	54.76
	FT(4-shot)	93.91	61.47	82.69	20.91	32.75	84.44	83.27	56.85	31.61	81.45	13.06	6.06	54.81	54.1
Resnet-101	PixDA [44] (4-shot)	**95.37**	**66.63**	84.52	33.24	45.26	85.58	88.42	62.7	34.67	86.19	25.98	6.91	58.05	59.5
	FSDA [56](4-shot)	94.36	62.68	84.02	29.69	42.01	85.94	87.72	59.16	35.83	83.12	10.58	**20.02**	58.0	57.93
	Ours (4-shot)	95.26	65.88	**84.92**	**36.19**	**47.19**	**86.74**	**88.96**	**64.1**	**38.03**	**86.25**	**28.89**	19.22	**61.71**	**61.8**
	JT (5-shot)	94.17	61.6	82.98	27.27	41.55	84.32	85.65	54.81	26.68	82.57	28.56	15.55	50.5	56.63
	FT (5-shot)	93.48	58.73	83.0	20.64	37.22	85.25	81.84	57.58	30.77	81.8	25.31	22.61	52.22	56.19
Resnet-101	PixDA [44] (5-shot)	**95.5**	66.73	84.63	33.77	50.86	85.99	88.05	63.46	33.31	**87.24**	32.83	21.57	60.3	61.86
	FSDA [56] (5-shot)	94.41	63.31	84.47	30.91	50.87	86.05	88.19	61.34	28.17	86.32	35.5	24.17	58.56	60.94
	Ours (5-shot)	95.29	**68.07**	**84.77**	**35.99**	**54.71**	**86.58**	**88.89**	**64.57**	**38.19**	86.33	**37.75**	**25.51**	**61.92**	**63.74**

4 epochs. Fine-Tuning (FT) is a model fine-tuned for 30,000 iterations with the Source Only model on the target domain. We further compare CBFA with two state-of-the-art models for few-shot cross-domain semantic segmentation: FSDA [56] and PixDA [44], in which we follow the same results reported by the authors. Oughted to be regarded that, NAAE [42] and FDA [54] are the baselines introduced in [44] to compare their performances with PixDA and FSDA whereas have not been shown in our tables. It is because NAAE and FDA were originally invented for cross-domain semantic segmentation with a massive amount of target images for training, which is unfair to compare with the methods specified for few-shot transfer, *i.e.*, FSDA ,and PixDA. The results in [44] elucidate that they both significantly underperform PixDA. It means that if CBFA outperforms PixDA, it would be superior to NAAE and FDA in the few-shot transfer setting.

For a fair comparison, DeepLabV2 with ResNet101 is used across all the baselines. We employ the standard Intersection over Union (IoU) metric to measure the model performance in each category and take their average (mean IoU, mIoU) in summary.

B. IMPLEMENTATION DETAILS

Feature extractor and Segmentation Head G. We adopt the DeepLab-v2 framework with ResNet-101 [12] pre-trained on ImageNet by replacing the final predictor with the Atrous Spatial Pyramid Pooling module (ASPP) [3] as our base architecture. We set the stride of the last two convolutional layers from 2 to 1, so the resolution of the output features can be effectively mapped to 1/8 of the input image size. Finally, we add an up-sampling layer along with the DeepLab-v2 output to match the size of the input images.

Discriminator D. The structures of D_1 and D_2 are mirrored, yet they are two independent discriminators. To preserve the spatial information, we utilize 5 all full-convolutional layers with kernel 4×4, stride of 2, padding of 1 in D_1 and D_2. The channel number of the layers is $\{64, 128, 256, 512, 1\}$. Except for the last layer followed by a *tanh* to limit the output value between -1 and 1, the other layers are followed by a *leaky ReLU* [52]. The final outputs of the discriminators are feature maps. To stabilize the training of the discriminators, we prefer spectral normalization [29] rather than batch-normalization layers [15] after each convolutional layer. It helps to save the memory to promote a larger size to incorporate more images for each training batch.

Optimization. We implement our framework using PyTorch toolbox on four GTX 3090 GPUs with 24GB memory with batch size eight. We use the stochastic gradient descent (SGD) with Nesterov acceleration where the momentum is 0.9 and the weight decay is 10^{-4} to train the scene parser, and the initial learning rate is set as $2.5\times\ 10^{-4}$. To train the

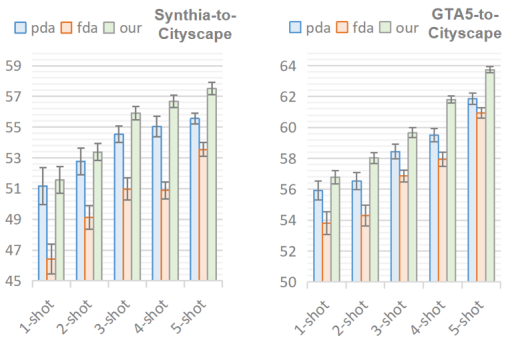


FIGURE 3. Stochastic Results of adapted scene parsing on SYNTHIA-to-CITYSCAPES and GTA5-to-CITYSCAPES. In each scenario, we illustrate the overall performances of PixDA (pda), FSDA (fda), and CBFA (our) across all categories from 1 to 5 shots, respectively.(Best viewed in color)

discriminators, we adopt two Adam optimizers [18] where the initial learning rate is 10^{-4} and the momentum is 0.9 and 0.99. The learning rates of all optimizers are decreased using the polynomial decay with a power of 0.9 as mentioned in [3]. The β is set as 0.95 in Equation (1). Finally, we set the λ_{seg} to 0.1 in Equation (2) and set the λ_{adv1} to 0.0002, λ_{adv2} to 0.001 in Equation (3).

C. RESULTS

GTA5 → CITYSCAPES. Table.1 shows the comparison between our approach and other baselines. Due to the different experimental settings, the results of source-only model reported by us do not match the results reported in [47]. Our method achieved a mIoU of 51.56 in the 1-shot empirical study, reaping a massive gain compared with source-only model and more importantly, achieving significant improvements over the joint training (JT, +9.01%) and fine-tuning (FT, +7.91%). It also outperforms FSDA, whose pipeline is similar with the based adaptation framework of CBFA, while slidely better than PixDA (+0.4%). As the number of target shot increases, the increased margin between CBFA and PixDA grow more significantly, *e.g.*, by +0.4% (1-shot), +1.10% (2-shot), +1.37% (3-shot), +1.64% (4-shot) and +1.96% (5-shot) in the average boost. On the other hand, since FSDA is similar with based adaptation framework, comparing their results may illustrate the effect of our motivation about counterfactual balancing for few-shot cross-domain semantic segmentation. As we have observed, CBFA exceeds FSDA by by +5.13% (1-shot), +4.76% (2-shot), +4.94% (3-shot), +5.80% (4-shot) and +3.97% (5-shot) in the mIoU.

Synthia → CITYSCAPES. Table.2 reported the performances of baselines in mIoU across all categories and their average for Synthia → CITYSCAPES. The mIOUs of our method are 56.79, 58.03, 59.67, 61.8 and 63.74 in 1-5 shot scenarios, respectively, exceeding the previous state-of-the-art PixDA by +0.87% (1-shot), +1.50% (2-shot), +1.23% (3-shot), +2.3% (4-shot) and +1.88% (5-shot). In terms of the performances of their source-only models, our CBFA has reaped considerable trasfer benefits by +23.28% (1-shot), +24.52% (2-shot), +26.16% (3-shot), +28.29% (4-shot) and +30.23% (5-shot). Obviously, CBFA behaves more competitively as the number of target images increases.

As we further observe, CBFA not only obtain the most impressive results in the mIoU but also perform very stably across different categories. In GTA5→CITYSCAPES, CBFA achieved more than 60% state of the art in the IoUs across 19 categories; In Synthia→CITYSCAPES, it outperforms more than 75% in 13 categories. Especially, in some under-represented classes that rarely appear in training data or contain few pixels within an image, *e.g.*, *motorcycle*, *rider*, *sign*, *bus* in GTA5→CITYSCAPES and Synthia→CITYSCAPES, CBFA presents a larger gain margin in their IoUs than the other baselines. In Fig.3, the stochastic results of PixDA, FSDA and our CBFA are performed, which are trained by randomly drawing different few-shot examples in the target domain then evaluated in the same manner (we randomly take 5 target training seeds to train the models, then observe their average and deviations). Their average performances are very close to the report in Table.1 and Table.2, whereas CBFA shows the most stable results. It implies that CBFA is probably better at the corner cases that inevitably happen in urban scene parsing.

Figure 4 illustrates the comparison of visualization results, indicating that our CBFA has significantly improved the pixel recognition, especially in street lights, signal lights, persons and bicycles, etc.

D. ANALYSIS

TABLE 3. Ablation study shows the contributions of different components on the mIoU in terms of SYNTHIA-to-CITYSCAPES and GTA5-to-CITYSCAPES.

Ablated baselines	SYNTHIA→ CITYSCAPES	GTA5→ CITYSCAPES
1-shot:		
Source-only(batch size = 8)	35.22	39.06
Based adaptation(batch size = 8)	48.30	42.76
CIP w/o	53.42	44.24
Contrastive learning w/o	53.28	47.55
Counterfactual sampling w/o	**56.80**	50.82
CBFA	56.79	**51.56**
5-shot:		
Source-only(batch size = 8)	35.22	39.06
Based adaptation(batch size = 8)	56.12	52.67
CIP w/o	59.28	51.96
Contrastive learning w/o	60.88	54.18
Counterfactual sampling w/o	62.75	56.89
CBFA	**63.74**	**57.52**

Ablation. We show how much each component contributes to the final performance behind CBFA. We consider 5 models ablated from CBFA: (a) Source only: we reported the performance when only using source data to train a scene parser in terms of our architecture (Note that the source only in Table.1,2 follows the report in [56], which is trained with 2 images for each batch, so their performances are different.). (b) Based adaptation: the CIP constraint and counterfactual paired contrastive learning were discarded. (c) w/o CIP: we discarded the inverse propensity score weighting scheme to balance the adversarial training while remain the other tech-

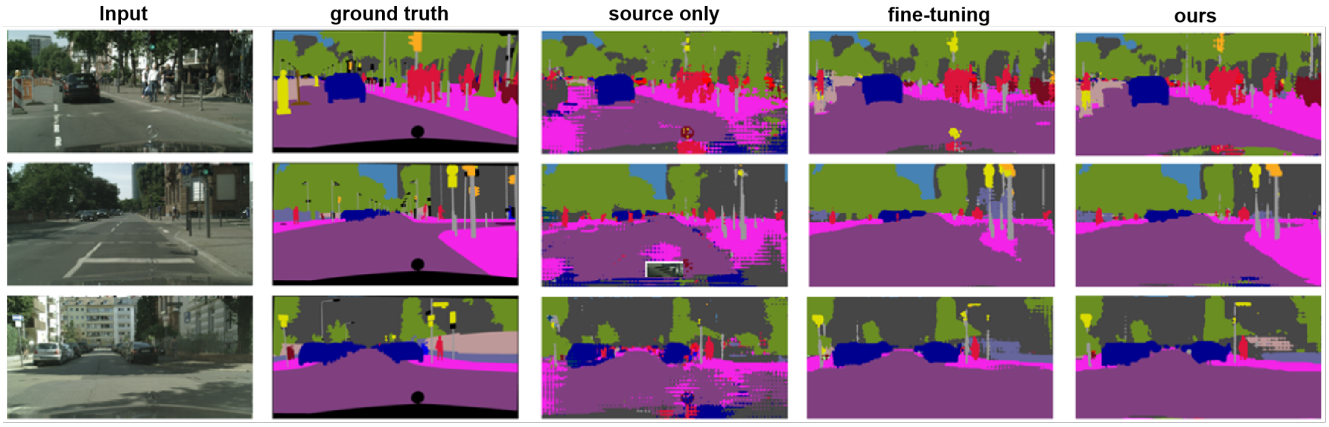


FIGURE 4. Example results of adapted scene parsing on SYNTHIA-to-CITYSCAPES. For each test image, we show the source-only (before adaptation), fine-tuning and our adapted results in the output space.

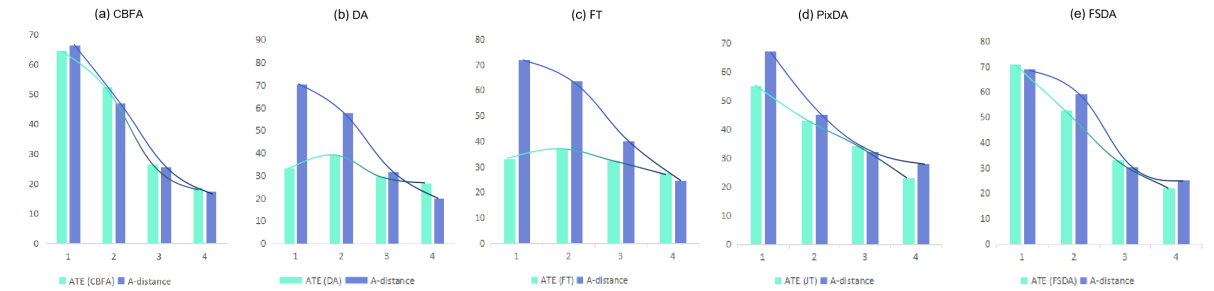


FIGURE 5. Relation between ATE and the approximation of $\mathcal{A}$-distance, *i.e.*, $\hat{d}_{\mathcal{A}}$ (Best viewed in color). Since the value of $\mathcal{A}$-distance is bounded in [0,2], we take $40\hat{d}_{\mathcal{A}}$ for better illustrating the connection between ATE and $\mathcal{A}$-distance.

niques to train the discriminator. (d) w/o contrastive learning: we ignored the contrastive learning part in Eq.14 to train the discriminator. (e) w/o counterfactual sampling: we took the random sampling instead of the sampling strategy in Eq.13. The mIoU results have been reported in Table.3. It can be observed that although (c-e) have played important roles in CBFA, the CIP constraint is the pivotal part to boost the performance of CBFA since it directly determines how to learn the domain-invariant features under the source-target imbalance situation.

Connection between ATE and domain adaptation. It is a primitive motivation of this work that "Covariate balancing in Rubin causal framework may relieve the cross-domain imbalance while reducing the domain gap". In order to validate our motivation, we consider how the relationship between ATE and domain discrepancy changes as the training in process and thus, compare CBFA with the based adaptation model and other baselines[2]. In particular, we take the based ATE formula in Eq.2 to calculate the γ_{ATE} for CBFA and the other baseline models in the range of all source and target images for training and test, in terms of each category. Notice that we abandoned the propensity score simulator to infer γ_{ATE}, in order to fairly compare CBFA with the other baselines (massive data in both source and target domains reduce the statistical importance of propensity score in Eq.2). We extract the pixelwise features in the final layer before segmentation and collect them to calculate γ_{ATE} for each class, then further take their average to obtain ATE. In terms of domain discrepancy, we take $\mathcal{A}$-distance [1] to represent the statistical gap between source and target domains. In terms of the similar training routine Since the exact $\mathcal{A}$-distance is intractable for computation, we follow [23] to calculate its approximation. It is also achieved by measuring the disparity on pixelwise features.

Fig.5 illustrates the relationship between ATE and the magnified $\mathcal{A}$-distance across the 300^{th}, 3000^{th}, 10000^{th}, 30000^{th} iterative processes. We have found that as the CBFA training in progress, the approximated $\mathcal{A}$-distance decreases along with the reduction in ATE. However, this monotonicity would disappear if we switch the baseline into a based adaptation model, which remains a significant ATE value given a well-trained model. It implies that better controlling the value of ATE may refer to reducing the domain discrepancy. It verifies the motivation of this work.

Convergence. We finally compare the convergence based on our CBFA (1-shot) with its ablated variants in the experiment Synthia→CITYSCAPE. Fig.6 (Right) illustrates that CBFA converges faster than the other models: it converges after 10000 iterations of training, and the loss fluctuation is relatively stable, in contrast, the based adaptation model still diverges after 30000 iterations of training.

[2]The evaluation includes FT, JT and FSDA except for PixDA in terms of its different training manner

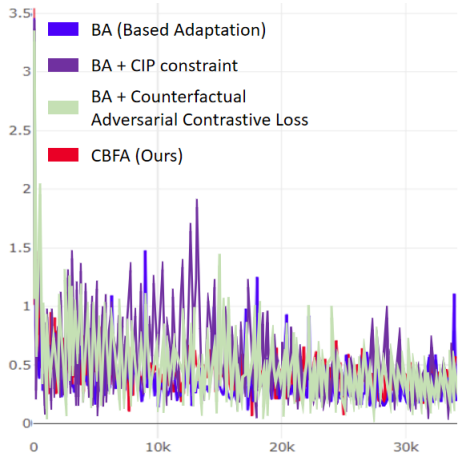


FIGURE 6. The convergence behavior between the based model and the CBFA model in their training losses (1-shot in Synthia→CITYSCAPES). The horizontal axis represents the number of iterations (Best viewed in color).

VI. CONCLUSION

In this paper, we attempt to solve the problem of few-shot cross-domain semantic segmentation for virtual-to-real scene parsing, where only a few of target images with spatial category masks are provided for pixel-level transfer learning. Distinct from existing methods, we recast the problem from a causal len. We treat data drawn from source and target domains as the treat and control groups for estimating the treatment effects, hence, the cross-domain data imbalance may refer to covariate balancing in PO framework. We resemble the spirit of some basic solution for covariate balancing, thus, inverse propensity score weighting and counterfactual matching, which leads to our CBFA approach to solve few-shot cross-domain semantic segmentation. The experimental results on GTA5-to-CITYSCAPES and SYNTHIA-to-CITYSCAPES have demonstrated the effectiveness of our method, and verify some connection between few-shot cross-domain semantic segmentation and Rubin causality.

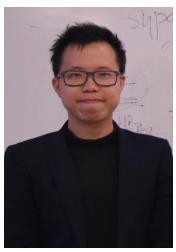
The technical merit of our methodology inspires further investigation to the imbalance learning, especially, for the problems that imbalanced data are distributed across different domains and categories. However, if categories across domains are not consistent, it would still be a great challenge to incorporate the domain-inconsistent classes into CBFA. To this end, the future of our research aims to solve the computer vision problems corresponding to the PO framework where two groups are substantially heterogeneous.

REFERENCES

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, vol.79(1):pp.151–175, 2010.
[2] M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga. Fast differentiable sorting and ranking. In *2020 International Conference on Machine Learning*, pp.950–959.
[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.40(4):pp.834–848, 2018.
[4] Y. Chen, W. Li, and L. Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.7892–7901.
[5] Z. Chen, P. Wei, J. Zhuang, G. Li, and L. Lin. Deep cocktail networks. *International Journal of Computer Vision*, vol.129(8):pp.2328–2351, 2021.
[6] N. Dong and E. P. Xing. Few-shot semantic segmentation with prototype learning. *2018 British Machine Vision Conference*, vol.3:pp.4–10.
[7] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *2017 International Conference on Machine Learning*, pp.1126–1135.
[8] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv:1409.7495*, 2014.
[9] A. N. Glynn and K. M. Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, vol.18(1):pp.36–56, 2010.
[10] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network. In *2018 Proceedings of the European Conference on Computer Vision*, pp.770–785.
[11] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 IEEE International Conference on Computer Vision*, pp.999–1006.
[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.
[13] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *2018 International Conference on Machine Learning*, pp.1989–1998.
[14] G. W. Imbens and D. B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.
[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.
[16] J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, vol.6(1):pp.1–54, 2019.
[17] N. Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *2020 International Conference on Machine Learning*, pp.5067–5077.
[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
[19] K. Kuang, P. Cui, B. Li, M. Jiang, and S. Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *2017 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.265–274*,.
[20] F. Li, L. E. Thomas, and F. Li. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, vol.188(1):pp.250–257, 2019.
[21] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *2019 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2537–2546.
[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3431–3440.
[23] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *2015 International Conference on Machine Learning*, pp.97–105.
[24] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *2016 Advances in Neural Information Processing Systems*, pp.136–144.
[25] W. Luo and Y. Zhu. Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, vol.38(4):pp.888–900, 2020.
[26] X. Luo, J. Xu, and Z. Xu. Channel importance matters in few-shot image classification. In *2022 International Conference on Machine Learning*, pp.14542–14559.
[27] F. Lv, T. Liang, X. Chen, and G. Lin. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In *2020 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4334–4343.
[28] Y. Mansour, M. Mohri, J. Ro, A. T. Suresh, and K. Wu. A theory of multiple-source adaptation with limited target labeled data. In *2021 International Conference on Artificial Intelligence and Statistics*, pp.2332–2340.
[29] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv:1802.05957*, 2018.
[30] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. In *2017 Advances in Neural Information Processing Systems*, pp.6670–6680.
[31] X. Pan, Y. You, Z. Wang, and C. Lu. Virtual to real reinforcement learning

for autonomous driving. *arXiv:1704.03952*, 2017.
[32] J. Pearl. Causality. Cambridge university press, 2009.
[33] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *2017 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4353–4361.
[34] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine. Few-shot segmentation propagation with guided networks. *arXiv:1806.07373*, 2018.
[35] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3234–3243.
[36] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.
[37] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, vol.66(5):pp.688, 1974.
[38] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. *arXiv:1711.06969*, 2017.
[39] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots. One-shot learning for semantic segmentation. *arXiv:1709.03410*, 2017.
[40] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.761–769, 2016.
[41] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *2017 Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp.4080–4090, 2017.
[42] R. Sun, X. Zhu, C. Wu, C. Huang, J. Shi, and L. Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *2019 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4360–4369.
[43] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1199–1208, 2018.
[44] A. Tavera, F. Cermelli, C. Masone, and B. Caputo. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation. In *2022 Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp.1626–1635, 2022.
[45] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium*, pp.1013–1020.
[46] T. Teshima, I. Sato, and M. Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *2020 International Conference on Machine Learning*, pp.9458–9469.
[47] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. *arXiv:1802.10349*, 2018.
[48] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.7167–7176.
[49] Q. Wang, D. Dai, L. Hoyer, L. Van Gool, and O. Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *2021 Proceedings of the IEEE International Conference on Computer Vision*, pp.8515–8525.
[50] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,pp.7268–7277.
[51] K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *2021 Advances in Neural Information Processing Systems*, vol.34, pp.10823–10836.
[52] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv:1505.00853*, 2015.
[53] R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3964–3973.
[54] Y. Yang and S. Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *2020 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4085–4095.
[55] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015.
[56] J. Zhang, Z. Chen, J. Huang, L. Lin, and D. Zhang. Few-shot structured domain adaptation for virtual-to-real scene parsing. In *2019 Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp.0–0.
[57] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan. Perspective-adaptive convolutions for scene parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42(4):pp.909–924, 2019.
[58] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *2017 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.2881–2890.

CODY MINGZHAN SUN received his B.S and M.S. degrees in mathematics and statistics from Sun Yat-sen University, respectively, advised by Professor Yonglong Dai. He founded Plan Technology group and has been committed to the development of FinTech. He was employed as a professional Master tutor of the Department of Finance and the Department of Statistics of Sun Yat-sen University. During his studies, his main research interests include Markov process and stochastic finance. His current research interests include inference and interventions in causal models, and machine learning.



ZILIANG CHEN has received his a Ph.D. degree in computer science and technology at Sun Yat-sen University, advised by Professor IET and IAPR Fellow Liang Lin. He also earned the B.S. degree in mathematics and applied mathematics from Sun Yat-sen University, Guangzhou, China. His research interests include computer vision and machine learning. He published more than 10 papers in ICML, CVPR, ICCV, AAAI, IJCV, TKDE, etc; and served as PCs (Program Commitees) of many leading conferences, *e.g.*,ICML, NeurIPS, AISTATS, AAAI, IJCAI, CVPR, ICCV, ECCV, a senior PC of IJCAI-2021. Now he is an assistant professor of Jinan University in Guangzhou.

LIANG LIN is a Full Professor of computer science at Sun Yat-sen University. He served as the Executive Director and Distinguished Scientist of SenseTime Group from 2016 to 2018, leading the R&D teams for cutting-edge technology transferring. He has authored or co-authored more than 200 papers in leading academic journals and conferences, and his papers have been cited by more than 16,000 times. He is an associate editor of IEEE Trans. Neural Networks and Learning Systems and IEEE Trans. Human-Machine Systems, and served as Area Chairs for numerous conferences such as CVPR, ICCV, SIGKDD and AAAI. He is the recipient of numerous awards and honors including Wu Wen-Jun Artificial Intelligence Award, the First Prize of China Society of Image and Graphics, ICCV Best Paper Nomination in 2019, Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, Best Paper Dimond Award in IEEE ICME 2017, Google Faculty Award in 2012. His supervised PhD students received ACM China Doctoral Dissertation Award, CCF Best Doctoral Dissertation and CAAI Best Doctoral Dissertation. He is Fellows of IET and IAPR.