FREDOM: Fairness Domain Adaptation Approach to Semantic Scene Understanding

Thanh-Dat Truong[1], Ngan Le[1], Bhiksha Raj[2], Jackson Cothren[3], Khoa Luu[1]
[1]CVIU Lab, University of Arkansas, USA [2]Carnegie Mellon University, USA
[3]Dep. of Geosciences, University of Arkansas, USA
{tt032, thile, jcothre, khoaluu}@uark.edu, bhiksha@cs.cmu.edu

Abstract

Although Domain Adaptation in Semantic Scene Segmentation has shown impressive improvement in recent years, the fairness concerns in the domain adaptation have yet to be well defined and addressed. In addition, fairness is one of the most critical aspects when deploying the segmentation models into human-related real-world applications, e.g., autonomous driving, as any unfair predictions could influence human safety. In this paper, we propose a novel Fairness Domain Adaptation (FREDOM) approach to semantic scene segmentation. In particular, from the proposed formulated fairness objective, a new adaptation framework will be introduced based on the fair treatment of class distributions. Moreover, to generally model the context of structural dependency, a new conditional structural constraint is introduced to impose the consistency of predicted segmentation. Thanks to the proposed Conditional Structure Network, the self-attention mechanism has sufficiently modeled the structural information of segmentation. Through the ablation studies, the proposed method has shown the performance improvement of the segmentation models and promoted fairness in the model predictions. The experimental results on the two standard benchmarks, i.e., SYNTHIA → Cityscapes and GTA5 → Cityscapes, have shown that our method achieved State-of-the-Art (SOTA) performance[1].

1. Introduction

Semantic segmentation has achieved remarkable results in a wide range of practical problems, including scene understanding, autonomous driving, and medical imaging, by using deep learning models, e.g., Convolutional Neural Networks (CNN) [3, 4, 24], Transformers [45]. Despite the phenomenal achievement, these data-driven approaches still need to improve in treating the prediction of each class. In particular, the segmentation models typically treat unfairly between classes in the dataset according to the class distributions. It is known as the fairness problem of semantic segmentation. The unfair predictions of segmentation models can lead to severe problems, e.g., in autonomous driving, unfair predictions may result in wrong decisions in motion planning control and therefore affect human safety. Moreover, the fairness issue of segmentation models is even well observed or exaggerated when the trained models are deployed into new domains. Many prior works alleviate the performance drop on new domains by using unsupervised domain adaptation, but these approaches do not guarantee

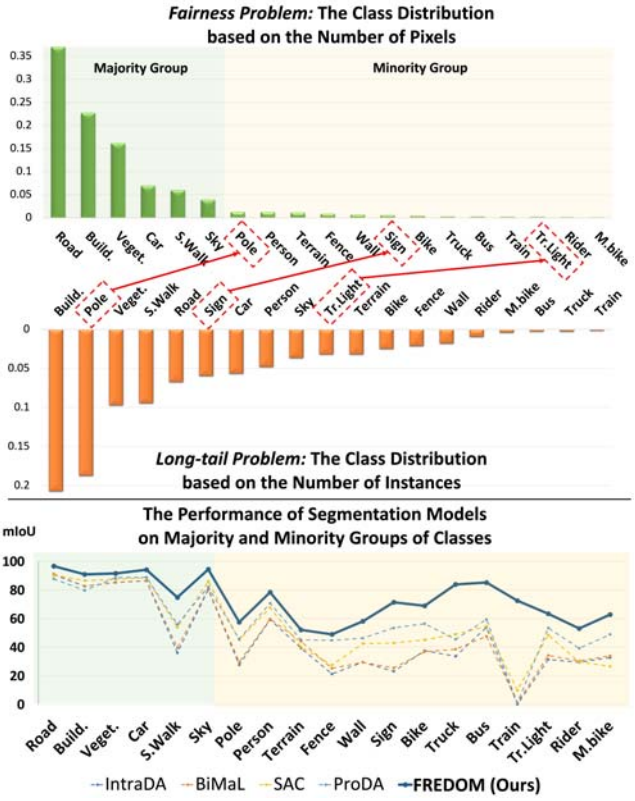

Figure 1. **The class distributions on Cityscapes are defined for Fairness problem and Long-tail problem.** In *long-tail* problem, several head classes frequently exist in the dataset, e.g., Pole, Traffic Light, or Sign. Still, these classes belong to a minority group in the *fairness* problem as their appearance on images does not occupy too many pixels. Our FREDOM has promoted the fairness of models illustrated by an increase of mIoU on the minority group.

[1]The implementation of FREDOM is available at https://github.com/uark-cviu/FREDOM

Figure 2. **Illustration of the Presence of Classes between Major (green boxes) and Minor (red boxes) Groups**. Classes in the minority group typically occupy fewer pixels than the ones in the majority group (Best view in color and 2× zoom).

the fairness property.

There needs to be more attention on addressing the fairness issue in semantic segmentation under the supervised or domain adaptation settings. Besides, the definition of fairness in semantic segmentation needs to be better defined and, therefore, often needs clarification with the long-tail issue in segmentation. In particular, the ***long-tail problem*** in segmentation is typically caused by ***the number of existing instances*** of each class in the dataset [21, 44]. Meanwhile, the ***fairness problem*** in segmentation is considered for ***the number of pixels*** of each class in the dataset. Although there could be a correlation between fairness and long-tail problems, these two issues are distinct. For example, several objects constantly exist in the dataset, but their presence often occupies only tiny regions of the given image (containing a small number of pixels), e.g., the Pole, which is a head class in Cityscapes, accounts for over 20% of instances while the number of pixels does only less than 0.01% of pixels. Hence, upon the fairness definition, it should belong to the minor group of classes as its presence does not occupy many pixels in the image. Another example is Person, which accounts for over 5% of instances, while the number of pixels does only less than 0.01% of pixels. Traffic Lights or Signs also suffer a similar problem. Fig. 2 illustrates the appearance of classes in the majority and minority groups. Therefore, although instances of these classes constantly exist in the dataset, these are still being mistreated by the segmentation model. Fig. 1 illustrates the class distributions defined based on long-tail and fairness, respectively.

Several works reduce the class imbalance effects using weighted (balanced) cross entropy [13, 21, 44], focal loss [1], data augmentation or rare-class sampling techniques [1, 19]. Still, these need to address the fairness problem directly. Indeed, many prior domain adaptation methods [6, 17, 28, 34, 36–39] have been used to improve the overall performance. However, these methods often ignore unfair effects produced by the model caused by the imbalanced class distribution. Besides, in some adaptation approaches using entropy minimization [29, 42], the model's bias caused by the class imbalance between majority and minority groups is even exaggerated [7, 35]. Meanwhile, other approaches using re-weighted or focal loss [1] often assume pixel independence and then penalize the loss contribution of each pixel individually and ignore the structural information of images. Then, pixel independence is relaxed by adopting the Markovian assumption [3, 48] to model segmentation structures based on neighbor pixels. In the *scope of our work*, we are interested in addressing the fairness problem in semantic segmentation between classes under the unsupervised domain adaptation setting. It should be noted that our interested problem is practical. In real-world applications (e.g., autonomous driving), deep learning models are typically deployed into new domains compared to the training dataset. Then, unsupervised domain adaptation plays a role in bridging the gap between the two domains.

Contributions of This Work: This work presents a novel Unsupervised **Fair**n**ess** **Dom**ain Adaptation (FREDOM) approach to semantic segmentation. To the best of our knowledge, this is one of the first works to address the fairness problem in semantic segmentation under the domain adaptation setting. Our contributions can be summarized as follows. First, the new fairness objective is formulated for semantic scene segmentation. Then, based on the fairness metric, we propose a novel fairness domain adaptation approach based on the fair treatment of class distributions. Second, the novel Conditional Structural Constraint is proposed to model the structural consistency of segmentation maps. Thanks to our introduced Conditional Structure Network, the spatial relationship and structure information are well modeled by the self-attention mechanism. Significantly, our structural constraint relaxes the assumption of pixel independence held by prior approaches and generalizes the Markovian assumption by considering the structural correlations between all pixels. Finally, our ablation studies have shown the effectiveness of different aspects in our approach to the fairness improvement of segmentation models. Through experiments, our FREDOM has promoted the fairness property of segmentation models and achieved state-of-the-art (SOTA) performance on two standard benchmarks of unsupervised domain adaptation, i.e., SYNTHIA → Cityscapes and GTA5 → Cityscapes.

2. Related Work

Unsupervised Domain Adaptation (UDA) in Semantic Segmentation is a vital research topic as its ability to reduce the necessity for massive volumes of labeled data. Adversarial learning [9, 15, 18, 26, 38, 40], and self-supervised training [1, 14, 19, 47] are common approaches to UDA.

Adversarial Learning is a common approach to UDA in semantic segmentation. The model is simultaneously trained on source and target domains in this approach. Hoffman *et al*. [17] introduced the first adversarial approach to UDA in segmentation. Then, Chen *et al*. [10] improved the model by utilizing pseudo labels in parallel with the global and class-wise adaptation learning process. The distillation loss with spatial-aware model [9] proposed by Chen *et al*. has been utilized to improve the spatial structures of seg-

mentation. Other methods have approached the UDA problem by using image translation [16, 27, 49]. SPIGAN [23] embed depth information as its privileged information to improve the UDA model for semantic segmentation. Similarly, Vu *et al*. [43] proposed a depth-aware framework using privileged depth information. Vu *et al*. [42] presented the first adversarial entropy minimization approach to UDA in segmentation. Then, [29, 46] presented a curriculum adaptation training from easy to complex samples ranked by the entropy level. Truong *et al*. [22, 35] improved the performance of segmentation models by introducing a bijective maximum likelihood approach.

Self-supervised Approach has gained a SOTA performance in UDA in semantic segmentation in recent years [1, 14, 19, 47, 50]. In self-training approaches, a new model is trained on unlabeled data using pseudo-labels derived from a trained model. Araslanov *et al*. [1] proposed an augmentation consistency approach to automatically evolve pseudo labels without using further training rounds. Zhang *et al*. [47] introduced a knowledge distillation approach to improving the performance of models while also correcting the soft pseudo labels online. Hoyer *et al*. [19] improved the performance of UDA via a new Transformer-based backbone and training recipe. Then, [19] is further improved by introducing a context-aware high-resolution framework that utilizes the advantages of small high-resolution crops for maintaining precise segmentation and large low-resolution crops for capturing context dependencies [20].

Class Imbalance Approaches: Jiawei *et al*. [30] presented a balanced Softmax loss that helps reduce labels' distribution shift and alleviates the long-tail issue. Wang *et al*. [44] proposed a Seesaw loss that reweights the contributions of gradients produced by positive and negative instances of a class by using two regularizers, i.e., mitigation and compensation. Ziwei *et al*. [25] proposed an algorithm that handles imbalanced classification, few-shot learning, and open-set recognition using dynamic meta-embedding. Chu *et al*. [11] proposed a stochastic training scheme for semantic segmentation, which improves the learning of debiased and disentangled representations. Szabo *et al*. [33] proposed tilted cross-entropy loss to reduce the performance differences, which promotes fairness among the target classes.

3. The Proposed Fairness Domain Adaptation Approach to Semantic Segmentation

Let $\mathbf{x}_s \in \mathcal{X}_s$ and $\hat{\mathbf{y}}_s \in \mathcal{Y}_s$ be an input image and its corresponding segmentation label in the source domain drawn from the source distribution p_s, $\mathbf{x}_t \in \mathcal{X}_t$ and $\hat{\mathbf{y}}_t \in \mathcal{Y}_t$ be the input image and the segmentation label in the target domain drawn from the target distribution p_t. In unsupervised domain adaptation, the ground-truth segmentation $\hat{\mathbf{y}}_t$ of image $\mathbf{x}_t$ is not available. Let $F : \mathcal{X} = \mathcal{X}_s \cup \mathcal{X}_t \rightarrow \mathcal{Y} = \mathcal{Y}_s \cup \mathcal{Y}_t$ be the deep network parameterized by θ that maps the input image $\mathbf{x} \in \mathcal{X}$ into the segmentation $\mathbf{y} \in \mathcal{Y}$, i.e $\mathbf{y}_s = F(\mathbf{x}_s, \theta)$, and $\mathbf{y}_t = F(\mathbf{x}_t, \theta)$. The standard domain adaptation can be mathematically formulated as in Eqn. (1).

$$\theta^* = \arg\min_{\theta} \left[\mathbb{E}_{\mathbf{x}_s, \hat{\mathbf{y}}_s \sim p_s(\mathbf{y}_s, \hat{\mathbf{y}}_s)} \mathcal{L}_s(\mathbf{y}_s, \hat{\mathbf{y}}_s) + \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t)} \mathcal{L}_t(\mathbf{y}_t)\right] \tag{1}$$

where $\mathcal{L}_s$ is the supervised cross-entropy (CE) loss in the source domain. Meanwhile, $\mathcal{L}_t$ is the unsupervised learning loss in the target domain that can be defined as the adversarial loss [29, 38, 39, 42], or the self-supervised loss [1, 19, 47]. In recent studies, the self-supervised loss defined by the cross-entropy loss with pseudo labels has achieved SOTA performance and outperformed other prior methods. Therefore, our proposed approach also defines $\mathcal{L}_t$ as the self-supervised loss [1, 19] with the novel fairness guarantee.

3.1. The Fairness Objective Function

Under the fairness constraint in semantic segmentation, the performance of each class should be equally treated by the deep model. Thus, the goal of fairness in semantic segmentation can be defined as in Eqn. (2).

$$\arg\min_{\theta} \sum_{c_i, c_j} \left| \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sum_k \mathcal{L}(y^k = c_i) - \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sum_k \mathcal{L}(y^k = c_j) \right| \tag{2}$$

where y^k denotes the k^{th} pixel of the segmentation $\mathbf{y}$, c_i and c_j are the class categories, i.e, $c_i, c_j \in [1..C]$ (where C is the number of classes), $\mathcal{L}$ is the loss function measuring the error rates of predictions. Formally, for all pairs of classes in the dataset, Eqn. (2) aims to minimize the difference in the error rates produced by the model between classes. Therefore, it guarantees all classes in the dataset are treated equally. Eqn. (2) can be further derived as in Eqn. (3).

$$\begin{aligned}
&\sum_{c_i, c_j} \left| \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sum_k \mathcal{L}(y_s^k = c_i) - \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sum_k \mathcal{L}(y_s^k = c_j) \right| \\
&\leq \sum_{c_i, c_j} \left(\mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sum_k \mathcal{L}(y_s^k = c_i) + \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \sum_k \mathcal{L}(y_s^k = c_j) \right) \\
&= 2C \mathbb{E}_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{y}) = 2C \left[\mathbb{E}_{\mathbf{x}_s \in \mathcal{X}_s} \mathcal{L}_s(\mathbf{y}_s) + \mathbb{E}_{\mathbf{x}_t \in \mathcal{X}_t} \mathcal{L}_t(\mathbf{y}_t) \right] \\
&= 2C \left[\mathbb{E}_{\mathbf{x}_s, \hat{\mathbf{y}}_s \sim p_s(\mathbf{y}_s, \hat{\mathbf{y}}_s)} \mathcal{L}_s(\mathbf{y}_s, \hat{\mathbf{y}}_s) + \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t)} \mathcal{L}_t(\mathbf{y}_t) \right]
\end{aligned} \tag{3}$$

From Eqn. (3), we can observe that the fairness objective in Eqn. (2) is bounded by the standard optimization of domain adaptation in Eqn. (1). Although optimizing the standard domain adaptation as in Eqn. (1) could impose the constraint of fairness under the upper bound in Eqn. (3), the imbalance class distributions of pixels cause the model to behave unfairly between classes when optimizing Eqn. (1). In particular, Eqn. (1) can be rewritten as follows,

$$\begin{aligned}
&\arg\min_{\theta} \left[\int \mathcal{L}_s(\mathbf{y}_s, \hat{\mathbf{y}}_s) p_s(\mathbf{y}_s) p_s(\hat{\mathbf{y}}_s) d\mathbf{y}_s d\hat{\mathbf{y}}_s + \int \mathcal{L}_t(\mathbf{y}_t) p_t(\mathbf{y}_t) d\mathbf{y}_t \right] \\
&= \arg\min_{\theta} \left[\int \sum_{k=1}^{N} \mathcal{L}_s(y_s^k, \hat{y}_s^k) p_s(y_s^k) p_s(\mathbf{y}_s^{\setminus k} | y_s^k) p_s(\hat{\mathbf{y}}_s) d\mathbf{y}_s d\hat{\mathbf{y}}_s \right. \\
&\qquad \left. + \int \sum_{k=1}^{N} \mathcal{L}_t(y_t^k) p_t(y_t^k) p_t(\mathbf{y}_t^{\setminus k} | y_t^k) d\mathbf{y}_t \right]
\end{aligned} \tag{4}$$

where N is the total number of pixels in the image, y_s^k and y_t^k are the k^{th} pixel of predicted segmentations in source and target domains, $\mathbf{y}_s^{\setminus k}$ and $\mathbf{y}_t^{\setminus k}$ are predicted segmentations without the k^{th} pixel in source and target domains, $p_s(y^k)$ and $p_t(y^k)$ are the class distributions of pixels in the source and target domains. The class distributions are computed based on the number of pixels of each class in the dataset. The terms $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ and $p_t(\mathbf{y}_t^{\setminus k}|y_t^k)$ are conditional structure constraints of $\mathbf{y}_s^{\setminus k}$ and $\mathbf{y}_t^{\setminus k}$ on y_s^k and y_t^k.

From imbalance distributions to unfair predictions: In practice, the class distributions of pixels $p_s(y_s^k)$ and $p_t(y_t^k)$ suffer imbalance problems as shown in Fig. 1. When the model is learned by the gradient descent method, the model behaves inequitably between classes. In particular, let us consider the behavior of gradients produced by the gradient descent learning method. Formally, let c_i and c_j be the two classes in the dataset and $p_s(y_s^k = c_i) << p_s(y_s^k = c_j)$. The gradients produced for each class with respect to the predictions can be formed as in Eqn. (5).

$$\left|\left|\frac{\partial \int \sum_{k=1}^{N} \mathcal{L}_s(y_s^k, \hat{y}_s^k) p_s(y_s^k = c_i) p_s(\mathbf{y}_s^{\setminus k}|y_s^k) p_s(\hat{\mathbf{y}}_s) d\mathbf{y}_s d\hat{\mathbf{y}}_s}{\partial \mathbf{y}_s^{(c_i)}}\right|\right| \ll \left|\left|\frac{\partial \int \sum_{k=1}^{N} \mathcal{L}_s(y_s^k, \hat{y}_s^k) p_s(y_s^k = c_j) q_s(\mathbf{y}_s^{\setminus k}|y_s^k) p_s(\hat{\mathbf{y}}_k) d\mathbf{y}_s d\hat{\mathbf{y}}_s}{\partial \mathbf{y}_s^{(c_j)}}\right|\right| \tag{5}$$

where $||.||$ is the magnitude of the vector, $\mathbf{y}_s^{(c_i)}$ and $\mathbf{y}_s^{(c_j)}$ represent the predicted probabilities of label c_i and c_j, respectively. As shown in Eqn. (5), the model inclines to produce significant gradient updates of the classes having a large population in the distributions (*a majority group*); meanwhile, the gradient updates of the class having a small population in the distributions (*a minority group*) are minor and dominated by the gradients of majority groups. Similar behavior can also be observed in the target domain.

3.2. The Proposed Fairness Adaptation Approach

As discussed in the previous section, the fairness problem is typically caused by imbalanced class distributions. Therefore, to address the fairness problem, we first assume that there exists an ideal distribution $p'_s(\mathbf{y}_s)$ and $p'_t(\mathbf{y}_t)$ so that the model trained on the ideal data distributions behave fairly between classes. It should be noted that we assume the ideal data distribution to frame and navigate our proposed approach to the fairness domain adaptation in semantic segmentation. Then, the ideal data distributions will be relaxed later and there is no requirement for the ideal data distribution during the training process. Formally, learning the adaptation framework of Eqn. (1) under the ideal data distribution can be formulated as in Eqn. (6).

$$\arg\min_{\theta} \left[\mathbb{E}_{\mathbf{x}_s \sim p_s(\mathbf{y}_s), \hat{\mathbf{y}}_\mathbf{s} \sim p_s(\hat{\mathbf{y}}_s)} \mathcal{L}_s(\mathbf{y}_s, \hat{\mathbf{y}}_s) \frac{p'_s(\mathbf{y}_s) p'_s(\hat{\mathbf{y}}_s)}{p_s(\mathbf{y}_s) p_s(\hat{\mathbf{y}}_s)} + \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t)} \mathcal{L}_t(\mathbf{y}_t) \frac{p'_t(\mathbf{y}_t)}{p_t(\mathbf{y}_t)}\right] \tag{6}$$

The fraction between ideal and real data distributions, i.e. $\frac{p'_s(\mathbf{y}_s)p'_s(\hat{\mathbf{y}}_s)}{p_s(\mathbf{y}_s)p_s(\hat{\mathbf{y}}_s)}$ and $\frac{p'_t(\mathbf{y}_t)}{p_t(\mathbf{y}_t)}$, can be interpreted as the complement of the model needed to be improved to achieve fairness against the imbalanced data. It should be noted that $p'_s(\hat{\mathbf{y}}_s)$ and $p_s(\hat{\mathbf{y}}_s)$ are constants as they are distributed over segmentation labels, so these could be excluded during training. Then, Eqn. (6) can be further derived as follows,

$$\arg\min_{\theta} \left[\mathbb{E}_{\mathbf{x}_s \sim p_s(\mathbf{y}_s), \hat{\mathbf{y}}_\mathbf{s} \sim p_s(\hat{\mathbf{y}}_s)} \sum_{k=1}^{N} \mathcal{L}_s(y_s^k, \hat{y}_s^k) \frac{p'_s(y_s^k) p'_s(\mathbf{y}_s^{\setminus k}|y_s^k)}{p_s(y_s^k) p_s(\mathbf{y}_s^{\setminus k}|y_s^k)} + \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t)} \sum_{k=1}^{N} \mathcal{L}_t(y_t^k) \frac{p'_t(y_t^k) p'_t(\mathbf{y}_t^{\setminus k}|y_t^k)}{p_t(y_t^k) p_t(\mathbf{y}_t^{\setminus k}|y_t^k)}\right] \tag{7}$$

As shown in Eqn. (7), if the conditional structure fractions $\frac{p'_s(\mathbf{y}_s^{\setminus k}|y_s^k)}{p_s(\mathbf{y}_s^{\setminus k}|y_s^k)}$ and $\frac{p'_t(\mathbf{y}_t^{\setminus k}|y_t^k)}{p_t(\mathbf{y}_t^{\setminus k}|y_t^k)}$ are ignored, Eqn. (7) becomes a special case of the weighted class balanced loss [13, 44]. However, conditional structure plays a vital role in semantic segmentation as it provides the constraints and correlation of structures among objects in images. The ignorance of conditional structure fractions could lower the performance of segmentation models. In addition, although the input images of the source and target domains can vary significantly in appearance due to the distribution shift, their segmentation maps between two domains share similar class distributions and structural information [35, 38, 39]. Hence, the distribution of segmentation in the target domain $p_t(\cdot)$ can be practically approximated by distribution in the source domain, i.e., $\frac{p'_t(\mathbf{y}_t)}{p_t(\mathbf{y}_t)} = \frac{p'_s(\mathbf{y}_t)}{p_s(\mathbf{y}_t)}$. In summary, by taking the log of Eqn. (7), the learning process can be formed as follows (the derivation of Eqn. (8) is detailed in the supplementary):

$$\begin{aligned}\theta^* \simeq \arg\min_{\theta} \Big[&\mathbb{E}_{\mathbf{x}_s \sim p_s(\mathbf{x}_s), \hat{\mathbf{y}}_\mathbf{s} \sim p_s(\hat{\mathbf{y}}_s)} \mathcal{L}_s(\mathbf{y}_s, \hat{\mathbf{y}}_s) + \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t)} \mathcal{L}_t(\mathbf{y}_t) \\ &+ \frac{1}{N} \sum_{k=1}^{N} \left(\mathbb{E}_{\mathbf{x}_s \sim p_s(\mathbf{x}_s)} \log\left(\frac{p'_s(y_s^k)}{p_s(y_s^k)}\right) + \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t)} \log\left(\frac{p'_s(y_t^k)}{p_s(y_t^k)}\right)\right. \\ &\left. + \mathbb{E}_{\mathbf{x}_s \sim p_s(\mathbf{x}_s)} \log\left(\frac{p'_s(\mathbf{y}_s^{\setminus k}|y_s^k)}{p_s(\mathbf{y}_s^{\setminus k}|y_s^k)}\right) + \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t)} \log\left(\frac{p'_s(\mathbf{y}_t^{\setminus k}|y_t^k)}{p_s(\mathbf{y}_t^{\setminus k}|y_t^k)}\right)\right)\Big]\end{aligned} \tag{8}$$

In summary, there are three terms in the learning objective of our FREDOM approach. Hence, several properties are brought into the learning process that can be observed.

Domain Adaptation Objective The first two terms stand for the objective of domain adaptation. While $\mathcal{L}_s$ learns to a segment on the source domain in the supervised fashion, $\mathcal{L}_t$ aims to unsupervised adapt knowledge to the target domain.

Fairness Treatment from Class Distributions The next two terms, i.e, $\log\left(\frac{p'_s(y_t^k)}{p_s(y_t^k)}\right)$ and $\log\left(\frac{p'_s(y_t^k)}{p_s(y_t^k)}\right)$, denoted as the $\mathcal{L}_{Class}$, impose the behavior of the model with respect to the class distribution. In particular, these constraints aim to regularize the predictions of classes so that the model should behave fairly between classes with respect to the class distribution. Under the ideal data distribution assumption, the model is expected to equally treat predictions of all

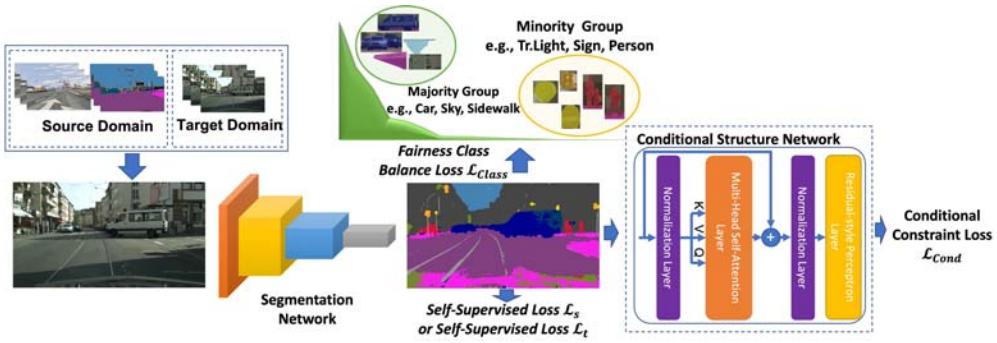

Figure 3. **The Proposed Fairness Framework.** The predictions of the inputs sampled from the source or target domains are penalized by the supervised loss $\mathcal{L}_s$ or the self-supervised loss $\mathcal{L}_t$, respectively. Then, the predictions are imposed by the fairness class balance loss $\mathcal{L}_{Class}$ followed by the Conditional Constraint Loss $\mathcal{L}_{Cond}$ computed via a Conditional Structure Network (Best view in color).

classes. Thus, to achieve the desired goal, the distributions of pixel classes should be uniformly distributed. Therefore, we adopt the uniform distribution of the class distribution $p'_s(y_s^k)$, i.e., $p'_s(y_s^k) = \frac{1}{C}$ where C is the number of classes.

Conditional Structure Constraint The last two terms, i.e., $\log\left(\frac{p'_s(\mathbf{y}_s^{\setminus k}|y_s^k)}{p_s(\mathbf{y}_s^{\setminus k}|y_s^k)}\right)$ and $\log\left(\frac{p'_s(\mathbf{y}_t^{\setminus k}|y_t^k)}{p_s(\mathbf{y}_t^{\setminus k}|y_t^k)}\right)$, denoted as $\mathcal{L}_{Cond}$, impose the conditional structure of the predicted semantic segmentation. This condition plays a role as a metric to measure the structural consistency of predicted segmentation maps with respect to the one under the ideal distributions where the model behaves fairly. Modeling the conditional structure, i.e., $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$, is a challenging problem. Several prior works modeled structural constraints by adopting the Markovian assumption [3, 48] where the models only consider the correlation between the current pixel with its neighbor pixels. However, the smoothness of predicted segmentation maps is highly dependent on the window size used in Markovian approaches (the number of neighbor pixels being selected). In our work, to sufficiently capture the conditional structural constraint, instead of modeling only neighborhood dependencies as Markovian approaches, we generalize it by modeling $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ via a conditional structure network (detailed in Sec. 4) to consider the correlation between all pixels in the segmentation.

Relaxation of Ideal Data Distribution One of the key challenging problems in optimizing Eqn. (8) is that the conditional ideal data distributions $p'_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ and $p'_s(\mathbf{y}_t^{\setminus k}|y_t^k)$ are not available. Therefore, instead of directly optimizing these terms, let us consider the tight bound as in Eqn. (9).

$$\mathbb{E}_{\mathbf{x}_s\sim p_s(\mathbf{x}_s)}\log\left(\frac{p'_s(\mathbf{y}_s^{\setminus k}|y_s^k)}{p_s(\mathbf{y}_s^{\setminus k}|y_s^k)}\right) + \mathbb{E}_{\mathbf{x}_t\sim p_t(\mathbf{x}_t)}\log\left(\frac{p'_s(\mathbf{y}_t^{\setminus k}|y_t^k)}{p_s(\mathbf{y}_t^{\setminus k}|y_t^k)}\right)$$
$$\leq -\left[\mathbb{E}_{\mathbf{x}_s\sim p_s(\mathbf{x}_s)}\log p_s(\mathbf{y}_s^{\setminus k}|y_s^k) + \mathbb{E}_{\mathbf{x}_t\sim p_t(\mathbf{x}_t)}\log p_s(\mathbf{y}_t^{\setminus k}|y_t^k)\right] \tag{9}$$

With any form of ideal distribution $p'_s(\cdot)$, Eqn. (9) always hold due to $\log p'_s(\cdot) \leq 0$. Hence, optimizing Eqn. (9) also ensure the conditional structural constraint in Eqn. (8) imposed due to the upper bound of Eqn. (9). Therefore, the demand for ideal data distribution is relaxed. Fig. 3 illustrates our proposed fairness domain adaptation framework.

4. The Conditional Structure Network

The conditional structural constraint $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ can be learned on the source dataset due to the availability of the ground-truth segmentation in the source domain. Formally, let $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ be modeled by the conditional structure network G with parameters Θ. Then the conditional structure network can be auto-regressively formed as follows:

$$\begin{aligned}&\arg\min_{\Theta}\mathbb{E}_{\mathbf{y}_s\in\mathcal{Y}_s} - \log p_s(\mathbf{y}_s^{\setminus k}|y_s^k,\Theta)\\ &= \arg\min_{\Theta}\mathbb{E}_{\mathbf{y}_s\in\mathcal{Y}_s}\sum_{i=1}^{N-1} - \log p_s(y^{\sigma_i^k}|y^{\sigma_{i-1}^k},...,y^{\sigma_1^k},y_s^k,\Theta)\end{aligned} \tag{10}$$

where σ^k is the permutation of $\{1...N\}\setminus\{k\}$. Eqn. (10) could be modeled by Recurrent Neural Networks [41]. However, directly adopting recurrent approaches remains some potential limitations. Particularly, as the recurrent approaches use a pre-defined permutation of regressive orders, it requires different conditional structure models for different initial pixel conditions, e.g., $p_s(\mathbf{y}_s^{\setminus k_1}|y_s^{k_1})$ and $p_s(\mathbf{y}_s^{\setminus k_2}|y_s^{k_2})$ should be modeled two different models. This problem could be alleviated by considering the permutation of regressive order as an network's input. However, learning a single network to model conditional structural constraints of different permutations is a heavy task and ineffective.

Instead of regressively forming $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$, we propose to model $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ in the parallel fashion. Particularly, let $\mathbf{m}$ be the binary masked matrix of $\mathbf{y}_s$, where the values of one and zero indicate a given pixel (unmasked pixel) and an unknown pixel (masked pixel), respectively. Then, the conditional structure $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ can be rewritten as $p_s(\mathbf{y}_s\odot(\mathbf{1}-\mathbf{m})|\mathbf{y}_s\odot\mathbf{m})$, where $\odot$ is the element-wise product and the mask $\mathbf{m}$ contains only one unmasked pixel, i.e., the given k^{th} pixel ($m^k = 1$). Learning the conditional structure constraint via binary mask $\mathbf{m}$ can be formed as:

$$\arg\min_{\Theta}\mathbb{E}_{\mathbf{y}_s\in\mathcal{Y}_s,\mathbf{m}\in\mathcal{M}} - \log p_s(\mathbf{y}_s\odot(\mathbf{1}-\mathbf{m})|\mathbf{y}_s\odot\mathbf{m}) \tag{11}$$

where $\mathcal{M}$ is the set of possible binary masks. Through Eqn. (11), modeling the conditional structural constraint $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$ can be equivalently interpreted as learning the condition of *masked pixels* on the given *unmask pixel*. To increase the modeling capability of the conditional structure network, three different strategies of the binary mask are adopted during training. First, the binary mask only contains one unmasked pixel to model the condition structural constraint $p_s(\mathbf{y}_s^{\setminus k}|y_s^k)$. Second, the binary mask does not contain any unmasked pixels (a zero mask). In this case, the model is going to learn the likelihood of the segmentation map $p_s(\mathbf{y}_s)$. Third, the binary mask contains more than one unmasked pixel that aims to increase the generalizability of the conditional structure network in modeling segmentation structures conditioned on the unmasked pixels.

To model conditional structure network G in a parallel fashion, the network G is designed as a Transformer. In particular, considering each pixel as a token, the network G is formed as the Transformer with L self-attention blocks where each block is designed in a residual style and the layer norms are applied to both the multi-head self-attention and multi-perceptron layers. By this design, the spatial relationship and structural dependencies can be modeled by the self-attention mechanism. To effectively optimize the network G, we adopt the learning tactic of Image-GPT [5].

5. Experiments

In this section, we present our experimental results on two standard benchmarks, i.e., SYNTHIA $\rightarrow$ Cityscapes and GTA5 $\rightarrow$ Cityscapes. First, we review datasets and our implementation, followed by analyzing the effectiveness of our approach to fairness improvement in ablation studies. Finally, we compare our experimental results with prior SOTA domain adaptation approaches. The performance of segmentation models is evaluated using the mean Intersection over Union (mIoU) and the IoU's standard deviation.

5.1. Datasets and Implementation

Cityscapes [12], a real-world dataset collected in European, consists of $3,975$ urban images with high-quality, dense annotations of 30 categories. The license of Cityscapes is available for academic and non-commercial purposes.

SYNTHIA [32] is a synthetic dataset for the semantic segmentation task generated from a virtual world. There are $9,400$ pixel-level labeled RGB images in SYNTHIA with 16 standard classes overlapping with Cityscapes. The license of SYNTHIA was registered under Creative Commons Attribution-NonCommercial-ShareAlike 3.0.

GTA5 [31], a synthetic dataset generated from the game engine, contains $24,966$ high-resolution, densely labeled images created for the semantic segmentation task. There are 19 standard classes between GTA5 and Cityscapes. The GTA5 dataset is protected under the MIT License.

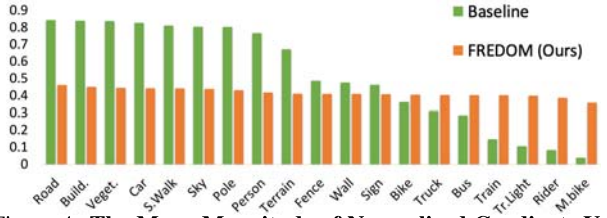

Figure 4. **The Mean Magnitude of Normalized Gradients Updated for Each Class.** Configuration (A) is used as the baseline.

Implementation Two different segmentation architectures are used in our experiments, i.e., (1) DeepLab-V2 [3] with the Resnet-101 backbone and (2) Transformer with the MiT-B3 backbone [45]. The Transformer design of [5] has been adapted to our conditional network structure G. Our framework is implemented in PyTorch and trained on four 48GB-VRAM NVIDIA Quadro P8000 GPUs. The model is optimized by the SGD optimizer [2] with learning rate 2.5×10^{-4}, momentum 0.9, weight decay 10^{-4}, and batch size of 4 per GPU. The image size is set to 1280×720 pixels. In the proposed FREDOM framework, the learning strategies and sampling techniques of [1, 19] are adopted for the self-supervised loss $\mathcal{L}_t$ to train our model. Our implementation is further detailed in the supplementary.

5.2. Ablation Study

Our ablation studies evaluate DeepLab-V2 models on two benchmarks under two settings, i.e., With and Without Adaptation. Each setting has three configs, i.e., (A) Model without $\mathcal{L}_{Class}$ and $\mathcal{L}_{Cond}$, (B) ***Fairness model*** with only $\mathcal{L}_{Class}$, and (C) ***Fairness model*** with $\mathcal{L}_{Class}$ and $\mathcal{L}_{Cond}$.

Does Adaptation Improve the Fairness? We evaluate the impact of Domain Adaptation in improving the fairness of classes in the minor group. As shown in Tab. 1, domain adaptation significantly improves fairness. In particular, without adaptation, the segmentation models trained only on the source data retain low performance in classes in the minor group, i.e., Traffic Light, Sign, and Fence. However, with our fairness domain adaptation approach, the overall accuracy and individual IoU of classes in the minor group are significantly boosted. In particular, the mIoU accuracy of segmentation models has been improved by $+22.4\%$ and $+21.6\%$ on SYNTHIA $\rightarrow$ Cityscapes and GTA5 $\rightarrow$ Cityscapes benchmarks. The model's fairness has been improved. Meanwhile, the IoU's STD of classes has been reduced by 1.4% and 4.5% on two benchmarks, respectively.

Does Class Distributions Matter to Fairness Improvement? As shown in Table 1, the fairness treatment from the class distribution loss $\mathcal{L}_{Class}$ contributes a significant improvement to both the overall performance and accuracy of classes in the minority group. In particular, the IoU accuracy of each class in configuration (B) is improved compared to the one in configuration (A) in both with and without adaptation settings. Specifically, in the adaptation setting on benchmark SYNTHIA $\rightarrow$ Cityscapes, the class dis-

Table 1. Effectiveness of our FREDOM (DeepLab-V2) approach to fairness improvement. There are three configurations: (A) Model without $\mathcal{L}_{Class}$ and $\mathcal{L}_{Cond}$. (B) **Fairness Model** with $\mathcal{L}_{Class}$ only. (C) **Fairness Model** with $\mathcal{L}_{Class}$ and $\mathcal{L}_{Cond}$.

Configuration		Majority Group						Minority Group													mIoU	STD
		Road	Build.	Veget.	Car	S.Walk	Sky	Pole	Person	Terrain	Fence	Wall	Sign	Bike	Truck	Bus	Train	Tr.Light	Rider	M.bike		
SYNTHIA → Cityscapes																						
Without Adaptation	(A)	64.9	71.5	73.1	62.9	26.1	71.0	21.7	48.4	–	0.2	3.0	0.2	35.6	–	27.9	–	0.1	20.7	12.0	33.7	27.8
	(B)	65.0	72.1	64.9	65.8	31.9	66.6	23.2	49.6	–	0.2	5.0	2.5	31.7	–	26.8	–	2.4	21.3	18.7	34.4	26.1
	(C)	65.2	73.3	65.4	69.0	32.2	67.7	34.5	50.0	–	0.3	17.5	3.5	39.9	–	27.0	–	3.9	21.9	18.5	36.7	25.4
With Adaptation	(A)	84.9	85.7	86.4	86.8	44.9	88.6	45.8	69.3	–	2.5	31.0	40.5	57.1	–	45.9	–	48.9	31.4	47.4	56.1	25.3
	(B)	84.8	85.8	86.4	86.8	45.2	88.9	47.6	70.1	–	2.6	31.3	43.0	58.5	–	46.0	–	51.9	34.1	49.2	57.0	24.9
	(C)	**86.0**	**87.0**	**87.1**	**87.1**	**46.3**	**89.1**	**48.7**	**71.2**	–	**5.3**	**33.3**	**46.8**	**59.9**	–	**54.6**	–	**53.4**	**38.1**	**51.3**	**59.1**	**24.0**
GTA5 → Cityscapes																						
Without Adaptation	(A)	75.8	77.2	81.3	49.9	16.8	70.3	25.5	53.8	24.6	21.0	12.5	20.1	36.0	17.2	25.9	6.5	30.1	26.4	25.3	36.6	24.0
	(B)	76.2	77.7	83.0	51.2	17.5	71.5	26.0	52.5	28.5	21.7	13.7	22.6	37.7	18.4	26.5	7.1	40.7	27.1	26.3	38.2	23.6
	(C)	77.1	79.4	84.7	52.9	18.5	72.3	28.6	54.4	33.8	22.5	15.6	23.7	38.9	19.7	27.1	7.9	41.6	28.6	28.0	39.7	23.6
With Adaptation	(A)	90.3	87.2	88.1	88.6	53.5	87.3	44.4	67.3	42.2	28.5	41.1	50.1	54.4	52.5	56.9	33.7	48.9	33.1	42.6	57.4	20.9
	(B)	90.6	87.3	88.1	88.8	53.7	87.4	44.9	67.7	42.3	28.6	41.9	52.9	57.6	55.2	57.5	47.6	50.8	36.9	44.9	59.2	19.8
	(C)	**90.9**	**87.8**	**88.6**	**89.7**	**54.1**	**89.5**	**45.2**	**68.8**	**42.6**	**32.6**	**44.1**	**57.1**	**58.1**	**58.4**	**62.6**	**55.3**	**51.4**	**40.0**	**47.7**	**61.3**	**19.1**

tribution loss $\mathcal{L}_{Class}$ has boosted the performance of classes in the minority group, e.g., Traffic Light (from 48.9% to 51.9%), Sign (from 40.5 to 43.0%), Pole (from 45.8% to 48.6%). Without adaptation, improvement is also observed. Moreover, the standard deviation of IoU over classes has been reduced. It shows that the model's fairness has been promoted. Similarly, the performance of models on benchmark GTA5 → Cityscapes is also consistently improved.

Does the Conditional Structure Constraint Contribute to Fairness Improvement? Configuration (C) in Table 1 reports experimental results of our model using conditional structure constraint loss $\mathcal{L}_{Cond}$. Results in Table 1 have shown the de facto role of the conditional structure constraint in performance improvement. Indeed, it enhances the IoU accuracy of each class in the minority group. For example, the average IoU accuracy of Fences, Pole, Traffic Light, and Sign has been improved by 2.3%. Overall, the performance of segmentation models has been improved by a notable margin, i.e., $+2.1\%$ and 2.7% on SYNTHIA → Cityscapes and GTA5 → Cityscapes, respectively. The difference in performance between classes is reduced, illustrated by the decrease of the IoU's standard deviation, which means the model's fairness is improved notably.

Does the Network Design Improve the Fairness? Table 2 illustrates the results of our approach using DeepLab-V2 and Transformer networks. As in our results, the performance of segmentation models using a more powerful backbone, i.e., Transformer, outperforms the models using DeepLab-V2. The performance of classes in the minority group has been improved notably, e.g., the performance of classes Fence, Traffic Light, Sign, and Pole has been improved to 9.3%, 65.1%, 60.1%, and 57.3% on the SYNTHIA → Cityscapes benchmark. The major improvements in the performance of overall and individual classes are also perceived in the GTA5 → Cityscapes benchmark. Also, the standard deviation of IoU over classes has been majorly reduced by 3.3%, illustrating that fairness has been promoted.

Does the Model Fairly Treat all Class During Training? Fig. 4 visualizes the gradients produced w.r.t each class in the domain adaptation setting. In particular, we take a subset in Cityscapes and compute the normalized gradients updated for each class. The model with our proposed approach tends to update gradients for each class fairly. Meanwhile, without using our fairness method, the gradients of classes in the minority group are dominated by the ones in the majority group, which could result in models' unfair behaviors.

5.3. Comparison with SOTA Approaches

SYNTHIA → Cityscapes Table 2 presents our experimental results using DeepLab-V2 and Transformer compared to prior SOTA approaches. Our proposed approach achieves SOTA performance and outperforms prior methods using the same network backbone. Specifically, the mIoU accuracy of our approach using Transformer is 67.0% and higher than DAFormer [19] by $+6.1\%$. Although the results of several individual classes are slightly lower than prior methods, overall, the mIoU accuracy and performance of individual classes in the minor group have been significantly promoted. Analyzing the mIoU accuracy of classes in the minor group, our results have been significantly improved compared to the prior SOTA method (i.e., DAFormer [19]). In particular, the performance of Rider, Fence, Pole, Traffic Light, and Sign classes has been improved by 4.1%, $+2.8\%$, $+7.3\%$, $+10.1\%$, and $+5.5\%$, respectively. In addition, the IoU accuracy of classes in the major group is also slightly enhanced. For example, the IoU accuracy of Building, Car, Sidewalk, and Sky has been improved to 87.8%, 89.7%, 54.1%, and 89.5%, respectively. It is vital to highlight that, to enhance the performance of classes in the minority group, the model does not sacrifice its ability to identify classes in the majority group. Instead, to promote the model's fairness, our approach enhances its ability to segment classes in the minor group to reduce the difference in performance between classes in minor and major groups.

GTA5 → Cityscapes As shown in Table 2, on the same network backbone, our FREDOM approach performs better

Table 2. Comparison of Semantic Segmentation Performance with UDA Methods Using DeepLab-V2 (**DL-V2**) and Transformer (**Trans.**).

Approach	Network	Majority Group						Minority Group													mIoU	STD
		Road	Build.	Veget.	Car	S.Walk	Sky	Pole	Person	Terrain	Fence	Wall	Sign	Bike	Truck	Bus	Train	Tr.Light	Rider	M.bike		
SYNTHIA → Cityscapes																						
IntraDA [29]	DL-V2	84.3	79.5	80.0	78.0	37.7	84.1	24.9	57.2	–	0.4	5.3	8.4	36.5	–	38.1	–	9.2	23.0	20.3	41.7	31.0
BiMaL [35]	DL-V2	**92.8**	81.5	82.4	85.7	51.5	84.6	30.4	55.9	–	1.0	10.2	15.9	38.8	–	44.5	–	17.6	22.3	24.6	46.2	30.9
SAC [1]	DL-V2	89.3	85.6	87.1	87.0	**47.3**	89.1	43.1	63.7	–	1.3	26.6	32.0	52.8	–	35.6	–	45.6	25.3	30.3	52.6	27.9
ProDA [47]	DL-V2	87.8	**84.6**	**88.1**	88.2	45.7	84.4	44.0	**74.2**	–	0.6	**37.1**	37.0	45.6	–	51.1	–	**54.6**	24.3	40.5	55.5	26.4
FREDOM	DL-V2	86.0	**87.0**	87.1	87.1	46.3	**89.1**	**48.7**	71.2	–	**5.3**	33.3	**46.8**	**59.9**	–	**54.6**	–	53.4	**38.1**	**51.3**	**59.1**	**24.0**
TransDA [8]	Trans.	**90.4**	86.4	**90.3**	**92.3**	**54.8**	93.0	53.8	71.2	–	1.7	31.1	37.1	49.8	–	66.0	–	61.1	25.3	44.4	59.3	27.3
ProCST [14]	Trans.	84.3	87.7	86.1	87.6	41.1	87.9	50.7	74.7	–	6.1	42.6	54.2	62.5	–	61.4	–	55.5	47.2	53.3	61.4	22.6
DAFormer [19]	Trans.	84.5	88.4	86.0	87.2	40.7	89.8	50.0	73.2	–	6.5	41.5	54.6	61.7	–	53.2	–	55.0	48.2	53.9	60.9	22.8
FREDOM	Trans.	89.4	**89.3**	89.9	90.5	50.8	**93.7**	**57.3**	**79.4**	–	**9.3**	**48.8**	**60.1**	**68.1**	–	**66.0**	–	**65.1**	**51.6**	**62.3**	**67.0**	**22.0**
GTA5 → Cityscapes																						
IntraDA [29]	DL-V2	90.6	82.6	85.2	86.4	36.1	80.2	27.6	59.3	39.3	21.3	29.5	23.1	37.6	33.6	53.9	0.0	31.4	29.4	32.7	46.3	26.7
BiMaL [35]	DL-V2	**91.2**	82.7	85.4	86.6	39.6	80.8	29.6	59.7	44.0	25.2	29.4	25.5	36.8	38.5	47.6	1.2	34.3	30.4	34.0	47.3	25.9
SAC [1]	DL-V2	90.3	86.6	87.5	88.5	53.9	86.0	45.1	67.6	40.2	27.4	42.5	42.9	45.1	49.0	54.6	9.8	48.6	29.7	26.6	53.8	24.2
ProDA [47]	DL-V2	87.8	79.7	88.6	88.8	**56.0**	82.1	**45.6**	**70.7**	**45.2**	**44.8**	**46.3**	**53.5**	56.4	45.5	59.4	1.0	53.5	39.2	48.9	57.5	21.7
FREDOM	DL-V2	90.9	**87.8**	**88.6**	**89.7**	54.1	**89.5**	45.2	68.8	42.6	32.6	44.1	**57.1**	**58.1**	**58.4**	**62.6**	**55.3**	51.4	**40.0**	**47.7**	**61.3**	**19.1**
TransDA [8]	Trans.	94.7	89.2	90.4	92.5	64.2	93.7	50.1	76.7	50.2	45.8	48.1	40.8	55.4	56.8	60.1	47.6	60.2	47.6	49.6	63.9	19.1
ProCST [14]	Trans.	95.8	89.8	90.2	92.3	69.6	93.0	49.8	72.2	50.3	45.0	55.8	63.3	63.1	72.2	78.8	65.1	56.8	44.9	56.4	68.7	17.1
DAFormer [19]	Trans.	95.7	89.4	89.9	92.3	70.2	92.5	49.6	72.2	47.9	48.1	53.5	59.4	61.8	74.5	78.2	65.1	55.8	44.7	55.9	68.3	17.3
FREDOM	Trans.	**96.7**	**90.9**	**91.6**	**94.1**	**74.8**	**94.4**	**57.5**	**78.4**	**52.1**	**49.0**	**58.1**	**71.4**	**68.9**	**83.9**	**85.2**	**72.5**	**63.4**	**53.1**	**62.8**	**73.6**	**15.8**

than previous SOTA methods. In particular, our approach using Transformer achieves the mIoU accuracy of 73.6%, which is the SOTA result; meanwhile, the result of the prior method [19] is 68.3%. Noticeably, the performance results have been significantly enhanced in the classes of the minority group, e.g., in comparison with DAFormer [19], the IoU accuracy of Rider, Motorbike, Pole, Traffic Light, and Sign has been increased by +8.4, +6.9%, 7.9%, +7.6%, and +12.0%. The performance accuracy has also improved in the majority group classes. For example, the accuracy of Building, Car, Sidewalk, and Sky is brought up to 90.9%, 94.1%, 74.8%, and 94.4%. Our FREDOM approach has strengthened the model's ability to segment classes in the minor group to lessen the performance gap between minor and major groups. In addition, the IoU's standard deviation over classes has been decreased compared to prior methods, which means that fairness has been promoted.

Qualitative Results Fig. 5 illustrates our results of the SYNTHIA → Cityscapces experiment. Our approach produces better quality results than prior UDA methods. Particularly, a significant improvement can be observed from the predictions of classes in the minority group, e.g., the predicted segmentation of signs, persons, and poles is sharper. The model can well segment the classes in the minor group cogently and minimize the region of classes being erroneously classified. The borders between classes are accurately identified and predicted segmentation continuity has improved compared to prior works. Although our predictions contain some noise, the boundaries are still clear and correspond to the labels. More comparisons of quantitative and qualitative results are available in the supplementary.

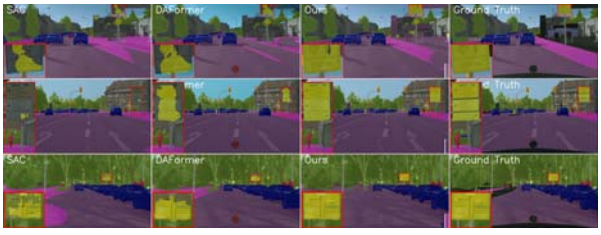

Figure 5. **Qualitative Results on SYNTHIA → Cityscapes** Columns 1-4 are the results of SAC [1], and DAFormer [19], our FREDOM, and ground truths (Best view in 2× zoom and color).

6. Conclusions and Limitations

This paper has presented the new fairness domain adaptation to semantic scene segmentation by analyzing the fairness treatment from class distributions. In particular, the conditional structural constraints have imposed the consistency of the predicted segmentation and modeled the structural information to improve the accuracy of segmentation models. Our ablation studies have analyzed different aspects affecting the fairness of segmentation models. It has also shown the effectiveness of our approach in terms of fairness improvement. Our FREDOM approach has achieved SOTA performance compared to prior methods.

Limitations: One of the potential limitations in our approach is the computational cost of the conditional structural constraint $\mathcal{L}_{Cond}$. As the constraint is computed by conditional structure network G, it requires more computational resources and time during training. Also, our work only utilized specific self-supervised loss, network backbones, and hyper-parameters to support our hypothesis. However, different aspects of learning have yet to be fully exploited, e.g., learning hyper-parameters, additional unsupervised loss $\mathcal{L}_t$ (adversarial loss, self-supervised loss). These could be further exploited in our future work.

Acknowledgment This work is supported by NSF Data Science, Data Analytics that are Robust and Trusted (DART), NSF WVAR-CRESH, and Googler Initiated Research Grant. We also acknowledge the Arkansas High Performance Computing Center for providing GPUs.

References

[1] Nikita Araslanov, , and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 8

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *in COMPSTAT*, 2010. 6

[3] 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. *TPAMI*, 2018. 1, 2, 5, 6

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[5] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. *ICML*, 2020. 6

[6] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, 2019. 2

[7] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, 2019. 2

[8] Runfa Chen, Yu Rong, Shangmin Guo, Jiaqi Han, Fuchun Sun, Tingyang Xu, and Wenbing Huang. Smoothing matters: Momentum transformer for domain adaptive semantic segmentation. *CoRR*, 2022. 8

[9] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *CVPR*, 2018. 2

[10] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017. 2

[11] Sanghyeok Chu, Dongwan Kim, and Bohyung Han. Learning debiased and disentangled representations for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:8355–8366, 2021. 3

[12] 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 *CVPR*, 2016. 6

[13] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 4

[14] Shahaf Ettedgui, Shady Abu-Hussein, and Raja Giryes. Procst: Boosting semantic segmentation using progressive cyclic style-transfer, 2022. 2, 3, 8

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 3

[17] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv:1612.02649*, 2016. 2

[18] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *CVPR*, 2018. 2

[19] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 2, 3, 6, 7, 8

[20] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[21] Ting-I Hsieh, Esther Robb, Hwann-Tzong Chen, and Jia-Bin Huang. Droploss for long-tail instance segmentation. In *Proceedings of the Workshop on Artificial Intelligence Safety 2021 co-located with the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021. 2

[22] Ibsa Jalata, Naga Venkata Sai Raviteja Chappa, Thanh-Dat Truong, Pierce Helton, Chase Rainwater, and Khoa Luu. Eqadap: Equipollent domain adaptation approach to image deblurring. *IEEE Access*, 10:93203–93211, 2022. 3

[23] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPIGAN: Privileged adversarial learning from simulation. In *ICLR*, 2019. 3

[24] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 1

[25] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world, 2019. 3

[26] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[27] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *CVPR*, 2018. 3

[28] Pha Nguyen, Thanh-Dat Truong, Miaoqing Huang, Yi Liang, Ngan Le, and Khoa Luu. Self-supervised domain adaptation in crowd counting. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2786–2790, 2022. 2

[29] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*, 2020. 2, 3, 8

[30] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition, 2020. 3

[31] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 6

[32] 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 *CVPR*, 2016. 6

[33] Attila Szabó, Hadi Jamali-Rad, and Siva-Datta Mannava. Tilted cross-entropy (tce): Promoting fairness in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2305–2310, 2021. 3

[34] Thanh-Dat Truong, Ravi Teja Nvs Chappa, Xuan-Bac Nguyen, Ngan Le, Ashley P.G. Dowling, and Khoa Luu. Otadapt: Optimal transport-based approach for unsupervised domain adaptation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2850–2856, 2022. 2

[35] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, and Khoa Luu. Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *ICCV*, 2021. 2, 3, 4, 8

[36] Thanh-Dat Truong, Chi Nhan Duong, Khoa Luu, Minh-Triet Tran, and Ngan Le. Domain generalization via universal non-volume preserving approach. In *CRV*, 2020. 2

[37] Thanh-Dat Truong, Pierce Helton, Ahmed Moustafa, Jackson David Cothren, and Khoa Luu. Conda: Continual unsupervised domain adaptation learning in visual perception for self-driving cars, 2022. 2

[38] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 2, 3, 4

[39] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative representations. *arXiv:1901.05427*, 2019. 2, 3, 4

[40] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[41] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks, 2016. 5

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 2, 3

[43] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019. 3

[44] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 4

[45] 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 *NeurIPS*, 2021. 1, 6

[46] Zizheng Yan, Xianggang Yu, Yipeng Qin, Yushuang Wu, Xiaoguang Han, and Shuguang Cui. *Pixel-Level Intra-Domain Adaptation for Semantic Segmentation*. Association for Computing Machinery, 2021. 3

[47] 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. *arXiv preprint arXiv:2101.10979*, 2021. 2, 3, 8

[48] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2, 5

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3

[50] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 3