Transferring Fairness under Distribution Shifts via Fair Consistency Regularization

Bang An
Department of Computer Science
University of Maryland, College Park
bangan@umd.edu

Zora Che
Department of Computer Science
Boston University
zche@bu.edu

Mucong Ding
Department of Computer Science
University of Maryland, College Park
mcding@umd.edu

Furong Huang
Department of Computer Science
University of Maryland, College Park
furongh@umd.edu

Abstract

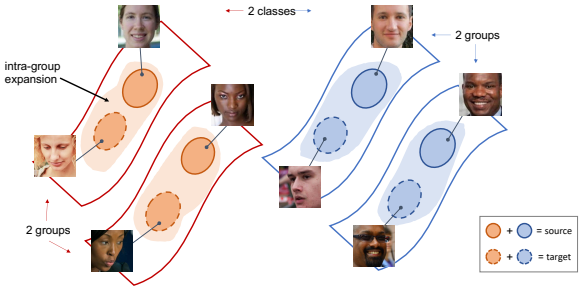
The increasing reliance on ML models in high-stakes tasks has raised a major concern on fairness violations. Although there has been a surge of work that improves algorithmic fairness, most of them are under the assumption of an identical training and test distribution. In many real-world applications, however, such an assumption is often violated as previously trained fair models are often deployed in a different environment, and the fairness of such models has been observed to collapse. In this paper, we study how to transfer model fairness under distribution shifts, a widespread issue in practice. We conduct a fine-grained analysis of how the fair model is affected under different types of distribution shifts, and find that domain shifts are more challenging than subpopulation shifts. Inspired by the success of self-training in transferring accuracy under domain shifts, we derive a sufficient condition for transferring group fairness. Guided by it, we propose a practical algorithm with a fair consistency regularization as the key component. A synthetic dataset benchmark, which covers all types of distribution shifts, is deployed for experimental verification of the theoretical findings. Experiments on synthetic and real datasets including image and tabular data demonstrate that our approach effectively transfers fairness and accuracy under various types of distribution shifts[1].

1 Introduction

Machine learning's social impact has broadened as it is widely used to aid decision-making in real-world applications, such as hiring, loan approval, facial recognition, and criminal justice. To avoid discrimination against a subset of population (e.g., w.r.t race or gender), many efforts on algorithmic fairness have been carried out [11, 20, 64, 43, 45, 14, 7]. Although existing work has achieved remarkable success in ensuring fairness, most of them assume the distribution of data at test time is identical to that in the training set. However, recent studies show that the fairness of a model is likely to collapse when encountering a distribution shift. For example, [18] observes that a fair income predictor trained with data from one state might not be fair when used in other states. [49] tries to maintain fairness in healthcare settings, but a model that performs fairly according to the metric evaluated in "Hospital A" shows unfairness when applied to "Hospital B". Such observations

[1] Code is available at https://github.com/umd-huang-lab/transfer-fairness.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



	Source	Target
Sampled from	UTKFace [70]	FairFace [29]
Class	Gender	
Sensitive attribute	Race	
Group	Gender + Race	
Shared	Underlying data generation process	
Differ	Capturing-bias, e.g., different angles, facial expressions	

Figure 1: **Illustration of intra-group expansion assumption in the input space.** An example of gender classification task with the sensitive attribute being race. Intra-group expansion assumes that different groups are separated but every group is self-connected under certain transformations. If a model has consistent predictions under those transformations, we can propagate labels within each group. Under this assumption, we propose to obtain fairness and accuracy in both domains by a self-training algorithm with fair consistency regularization.

motivate us to find the reason behind the collapse of fairness, and investigate how to transfer fairness under distribution shifts. Specifically, when we have labeled data in source domain and unlabeled data in target domain, we investigate how to adapt the fair source model to a target domain with the goal of achieving both accuracy and fairness in both domains.

Intuitively, the fairness of a model in target domain strongly depends on the nature of distribution shifts. In this paper, we only consider cases where the oracle model is the same in two domains. We characterize distribution shifts by assuming a unified latent variable model for the underlying data generation process. We categorize distribution shifts into three types [31]: 1) *Domain shift* where source and target distributions comprise data from related but distinct domains (e.g., train a model in hospital A but test it in hospital B). 2) *Subpopulation shift* where two domains overlap, but relative proportions of subpopulations differ (e.g., the proportion of female candidates increases at test time). 3) *Hybrid shift* where domain shift and subpopulation shift happen at the same time. Taking a closer look at fairness under them, we find that domain shift is more challenging since the model's performance is unpredictable in unseen domains. Such finding is supported empirically on a synthetic dataset we develop to simulate diverse types of distribution shifts. While recent work explores methods to transfer fairness [53, 47, 22], most considered settings fall into subpopulation shifts. In this paper, we consider more challenging but practical settings: domain shifts and hybrid shifts.

We draw inspirations from recent progress on self-training in transferring accuracy under domain shifts [60, 5, 69, 3, 48, 54]. The success of self-training is due to an expansion assumption and an input consistency regularization algorithm. The expansion assumption assumes two domains share one underlying generative model and the support of the distribution on each class is a connected compact set under data transformations (i.e., has a good continuity). Under the *expansion assumption*, [60] and [5] prove that self-training which enforces consistent predictions for the same input under different transformations can propagate labels from source to target domain. This approach exhibits superior performance in transferring accuracy [69, 48], but does not consider fairness.

Taking demography into consideration, we relax the expansion assumption to a more realistic *intra-group expansion assumption*, as shown in Figure 1, which only requires continuity of the underlying distribution within every group (i.e., data points with the same class and sensitive attribute) rather than the entire class. Based on intra-group expansion assumption, we derive a sufficient condition that guarantees fairness in both source and target domains. This sufficient condition suggests that ensuring the trained model to gain the same input consistency on different groups under a fair teacher classifier guarantees fairness in both domains. However, such a teacher classifier is not available in practice and we need a practical treatment.

Guided by the theoretical algorithm, we propose a practical self-training algorithm which builds upon Laftr [41], an adversarial learning method for fairness, and FixMatch [54], a self-training framework. To encourage similar input consistency on different groups as suggested by our theoretical algorithm, we propose a novel fair consistency regularization. By reweighting the consistency loss of each group dynamically according to model's performance, it encourages the model to pay more attention to high-error group while training. We evaluate our method under different types of distribution shifts with the synthetic and real datasets. Experiments show that our approach achieves high accuracy and

fairness in target domain without sacrificing the performance in source domain. To the best of our knowledge, this is the first work using self-training to transfer fairness under distribution shifts.

Summary of contributions: **(1)** We provide a fine-grained analysis of fairness under distribution shifts and develop a synthetic dataset to study model fairness under different types of distribution shifts. **(2)** Theoretically, we derive a sufficient condition for transferring fairness under distribution shifts. **(3)** Algorithmically, we propose a theory-guided algorithm for transferring fairness with a fair consistency regularization as the key component. **(4)** Experimentally, we evaluate our method on synthetic data, real image data and real tabular data. All results show the effectiveness of our approach in transferring fairness.

2 Preliminaries and Notations

Transfer Fairness. Let X, A, Y and $\mathcal{X}, \mathcal{A}, \mathcal{Y}$ denote random variables and sample space of input features, sensitive attribute and label. For simplicity, we assume binary sensitive attribute and binary classification, while our method can easily extend to a more general setting (see Appendix E). We aim to learn a classifier $g : \mathcal{X} \to \mathcal{Y}$ and are interested in the fairness of it under distribution shifts. Specifically, with S and T denoting source and target domains, we study how to transfer the fairness when $\mathbb{P}_S(X, A, Y) \neq \mathbb{P}_T(X, A, Y)$, with the labeled source data and unlabeled target data. In self-training algorithm, we use g_{tc} to denote a teacher classifier. We use the word "group" to denote the set of data that has the same label and sensitive attribute.

Unfairness Metric. We use *equalized odds* [26] to measure unfairness, $\Delta_{odds} = \frac{1}{2}\sum_{y=0}^{1}\left|\mathbb{P}(\hat{Y} = y|A = 0, Y = y) - \mathbb{P}(\hat{Y} = y|A = 1, Y = y)\right|$ where $\hat{Y} = g(X)$ is the prediction. To avoid trivial fairness where a model with constant output has $\Delta_{odds} = 0$, we also evaluate the variance of group accuracy, $V_{acc} = Var(\{\mathbb{P}(\hat{Y} = y|A = a, Y = y), \forall a, y\})$, in our experiments.

A Unified Framework to Characterize Distribution Shifts. We characterize distribution shifts by assuming a unified latent variable model for the underlying data generation process. We denote the underlying factors as $Y^1, Y^2, ..., Y^K$ (with $\mathcal{Y}^i$ as the sample space of Y^i), and data point as X.

Assumption 1. *[Underlying data generation process] We assume the data is generated from a latent generative model as $\boldsymbol{y}^{1:K} \sim \mathbb{P}(Y^{1:K})$ and $\boldsymbol{x} \sim \mathbb{P}(X|Y^{1:K} = \boldsymbol{y}^{1:K})$. The generative model is fixed $\mathbb{P}_S(X|Y^{1:K} = \boldsymbol{y}^{1:K}) = \mathbb{P}_T(X|Y^{1:K} = \boldsymbol{y}^{1:K})$ but the marginal distribution of factors varies in two domains $\mathbb{P}_S(Y^{1:K}) \neq \mathbb{P}_T(Y^{1:K})$, causing the distribution shift $\mathbb{P}_S(Y^{1:K}, X) \neq \mathbb{P}_T(Y^{1:K}, X)$.*

It assumes that two domains share the same data generation process. It is realistic, for example, the underlying physical process of cell imaging is fixed, but the distribution of underlying factors (e.g. *gender*, *age* or *equipment*) may vary in two hospitals (i.e. two domains), resulting in the distribution shift of the observed tissue images. Two of the factors are label Y^l (i.e. Y) and sensitive attribute Y^a (i.e. A). We call other factors *nuisance factors* since they are irrelevant to the classification task.

3 Fairness under Distribution Shifts

In this section, we provide a fine-grained analysis of fairness under various types of distribution shifts. Based on the unified framework, we consider two major types of distribution shifts, namely *subpopulation shift* and *domain shift*, which are widely considered in many applications [31].

Definition 3.1. *(Subpopulation Shift) We say it is a subpopulation shift, if for any factor Y^i, the sample space of it remains the same in two domains (i.e., $\mathcal{Y}_S^i = \mathcal{Y}_T^i$), but the marginal distribution of at least one factor changes resulting in $\mathbb{P}_S(Y^{1:K}) \neq \mathbb{P}_T(Y^{1:K})$ and $\mathbb{P}_S(Y^{1:K}, X) \neq \mathbb{P}_T(Y^{1:K}, X)$.*

Definition 3.2. *(Domain Shift) We say it is a domain shift, if at least one nuisance factor $Y^i, i \neq l, i \neq a$, has different sample space in two domains, $\exists y^i \in \mathcal{Y}_T^i$, but $y^i \notin \mathcal{Y}_S^i$, resulting in $\mathbb{P}_S(Y^{1:K}) \neq \mathbb{P}_T(Y^{1:K})$ and $\mathbb{P}_S(Y^{1:K}, X) \neq \mathbb{P}_T(Y^{1:K}, X)$.*

Intuitively, under subpopulation shift, the sample space overlaps, and only the marginal distributions of factors vary in two domains. For example, the proportion of young people versus elder people in training and deployment time differs. In contrast, under domain shift, the source model has never seen the data with factors' values that only exist in the target domain. For instance, the source model is unaware of the equipment used for cell imaging at deployment time.

Why do distribution shifts cause unfairness? Suppose the marginal distributions of a binary nuisance factor Y^i differ in two domains with $\mathbb{P}_S(Y^i) \neq \mathbb{P}_T(Y^i)$. The unfairness in two domains are

$$\begin{aligned}\Delta^S_{odds} &= \mathbb{P}_S(Y^i=0) \times \Delta^S_{odds}|_{Y^i=0} + \mathbb{P}_S(Y^i=1) \times \Delta^S_{odds}|_{Y^i=1} \\ \Delta^T_{odds} &= \mathbb{P}_T(Y^i=0) \times \Delta^T_{odds}|_{Y^i=0} + \mathbb{P}_T(Y^i=1) \times \Delta^T_{odds}|_{Y^i=1}.\end{aligned} \tag{1}$$

Since $\mathbb{P}_S(X|Y^i=y^i) = \mathbb{P}_T(X|Y^i=y^i)$, we have $\Delta^S_{odds}|_{Y^i=y^i} = \Delta^T_{odds}|_{Y^i=y^i}, \forall y^i \in \{0,1\}$. Under subpopulation shift, Y^i has the same sample space in two domains but with different proportions (e.g., $\mathbb{P}_S(Y^i=0)=0.9, \mathbb{P}_S(Y^i=1)=0.1, \mathbb{P}_T(Y^i=0)=0.1, \mathbb{P}_T(Y^i=1)=0.9$), while under domain shift the sample space differs (e.g. $\mathbb{P}_S(Y^i=0)=1, \mathbb{P}_T(Y^i=1)=1$). It is easy to see from (1) that if a model is perfectly fair on data with $Y^i=0$ but totally unfair on data with $Y^i=1$, then the model is highly fair in the source domain but highly unfair in the target domain under both cases. Therefore, if the model has inconsistent performance on data generated from different nuisance factors' values, then the shift marginal distribution of those factors may cause fairness collapse.

How to transfer fairness under distribution shifts? Based on the above analysis, one way is to train the model to be fair under any factors' values. It is possible under subpopulation shift as states in the following proposition (see proof and discussion in Appendix B).

Proposition 3.1. *[Transfer of fairness under subpopulation shift] Consider the subpopulation shift that is caused by the shifted marginal distribution of nuisance factor Y^i (i.e., $\mathbb{P}_S(Y^i) \neq \mathbb{P}_T(Y^i)$), while $\mathcal{Y}^i_S = \mathcal{Y}^i_T = \mathcal{Y}^i$. If model f is strictly fair in source domain under any value of factor Y^i satisfying $\mathbb{P}_S(f(X)=y^l|Y^a=0, Y^l=y^l, Y^i=y^i) = \mathbb{P}_S(f(X)=y^l|Y^a=1, Y^l=y^l, Y^i=y^i), \forall y^i \in \mathcal{Y}^i, y^l \in \{0,1\}$, then f is also fair in target domain with $\Delta_{odds}=0$.*

Our empirical results (Figure 3) also support this finding. However, domain shift is more challenging. The source model's performance on the target data is unpredictable due to the distinct sample space. One promising way to tackle domain shift is to enforce model's invariance to nuisance factors, so that source model would have the same behavior on target data. Note that, this solution also works for subpopulation shift since it will lead to the case in Proposition 3.1 directly. Above analysis motive us to transfer accuracy and fairness by doing consistency regularization in the next section.

4 Transfer Fairness via Fair Consistency Regularization

4.1 Theoretical Analysis: A Sufficient Condition for Transferring Fairness

In reality, distribution shifts are usually hybrid and we may not know all the attributes, while the underlying data generation process remains unchanged. In this section, we consider a general distribution shifts case with access to only input X, label Y, and sensitive attribute A. Due to the fixed underlying data generation process, we assume that the data X from two domains are from the same underlying conditional distribution of $X|_{Y,A}$ where groups are separated (i.e. by label and sensitive attribute) which is characterized as follows.

Assumption 2 (Separability of the input)**.** *Let S^y_a and T^y_a denote the sample space of $X|_{A=a,Y=y}$ in source and target domains. The ground truth class and sensitive attribute for $\boldsymbol{x} \in S^y_a \cup T^y_a$ are consistent, which are $y \in \{0,1\}$ and $a \in \{0,1\}$. We assume the sample spaces of X in two domains are $S = \cup_y \cup_a S^y_a$ and $T = \cup_y \cup_a T^y_a$, where groups are separated where 1) $S^y_a \cap S^y_{a'} = T^y_a \cap T^y_{a'} = S^y_a \cap T^y_{a'} = \emptyset, \forall y, a \neq a'$, and 2) $S^y_a \cap S^{y'}_{a'} = T^y_a \cap T^{y'}_{a'} = S^y_a \cap T^{y'}_{a'} = \emptyset, \forall a, a', y \neq y'$.*

This is a realistic assumption as illustrated in Figure 1. We define $U^y_a = \frac{1}{2}(S^y_a + T^y_a)$ as the group distribution, and U as the distribution on the entire data. Next, we characterize the good continuity of group distributions with the definition of *neighbor* and *intra-group expansion* assumption.

Definition 4.1 (Neighbor)**.** *Let $\mathcal{T}$ denote a set of input transformations (e.g. via data augmentation) and define the transformation set of $\boldsymbol{x}$ as $\mathcal{B}(\boldsymbol{x}) \triangleq \{\boldsymbol{x}'|\exists t \in \mathcal{T}, \text{s.t. } \|\boldsymbol{x}' - t(\boldsymbol{x})\| \leq r\}$. For any $\boldsymbol{x} \in S^y_a \cup T^y_a$, we define the neighbor of $\boldsymbol{x}$ as $\mathcal{N}(\boldsymbol{x}) := (S^y_a \cup T^y_a) \cap \{\boldsymbol{x}'|\mathcal{B}(\boldsymbol{x}) \cap \mathcal{B}(\boldsymbol{x}') \neq \emptyset\}$ and define the neighbor of a set $V \in \mathcal{X}$ as $\mathcal{N}(V) := \cup_{\boldsymbol{x} \in V \cap (\cup_y \cup_a S^y_a \cup T^y_a)} \mathcal{N}(x)$.*

Intuitively, two examples are neighbors if they are near each other after applying some transformations. Note that, we only consider neighbors that have the same class and sensitive attribute (i.e. from

the same group). Based on this definition, we characterize the continuity of group distribution with *intra-group expansion* assumption where any small set has a large neighbor in its group.

Assumption 3 (Intra-group expansion)**.** *We say that U_a^y satisfies (α, c)-multiplicative expansion for some constant $\alpha \in (0,1)$ and $c > 1$, if for all $V \subset U_a^y$ with $\mathbb{P}_{U_a^y}(V) \leq \alpha$, the following holds:*

$$\mathbb{P}_{U_a^y}(\mathcal{N}(V)) \geq \min\{c\mathbb{P}_{U_a^y}(V), 1\}.$$

Different from the *expansion* assumption proposed in [60] which considers the class continuity, *intra-group expansion* assumes group continuity. As shown in Figure 1, this is more realistic since groups are separated by both label and sensitive attribute. We can also interpret it as the transformations that change the value of nuisance factors will generate neighbors from the same group.

This assumption allows us to propagate labels within the group from one domain to another by encouraging the consistency under transformations. We use $R_{U_a^y}(g) \triangleq \mathbb{P}_{U_a^y}[\exists \boldsymbol{x}' \in \mathcal{B}(\boldsymbol{x}), \text{s.t. } g(\boldsymbol{x}) \neq g(\boldsymbol{x}')]$ to denote the *input consistency loss* of classifier g on the group distribution U_a^y, which is the fraction of examples where g is not robust to input transformations. Since we only have partial supervision (no labels in the target domain), we use a self-training framework to obtain a model that is accurate and fair in both domains (i.e. on U_a^y). Based on the theory of self-training in [60], we derive a sufficient condition in Theorem 4.1 that bounds the unfairness and error on the population distribution. We use 0-1 loss to evaluate the *error* of g as $\varepsilon_{U_a^y}(g) \triangleq \mathbb{P}_{U_a^y}[g(\boldsymbol{x}) \neq g^*(\boldsymbol{x}')]$, and the *disagreement* between g and a teacher classifier g_{tc} as $L_{U_a^y}(g, g_{tc}) \triangleq \mathbb{P}_{U_a^y}[g(\boldsymbol{x}) \neq g_{tc}(\boldsymbol{x}')]$.

Theorem 4.1 (Guarantee fairness)**.** *Suppose we have a teacher classifier g_{tc} with bounded unfairness such that $|\varepsilon_{U_a^y}(g_{tc}) - \varepsilon_{U_{a'}^{y'}}(g_{tc})| \leq \gamma, \forall a, a' \in \mathcal{A}$ and $y, y' \in \mathcal{Y}$. We assume intra-group expansion where U_a^y satisfies $(\bar{\alpha}, \bar{c})$-multiplicative expansion and $\varepsilon_{U_a^y}(g_{tc}) \leq \bar{\alpha} < 1/3$ and $\bar{c} > 3, \forall a, y$. We define $c \triangleq \min\{1/\bar{\alpha}, \bar{c}\}$, and set $\mu \leq \varepsilon_{U_a^y}(g_{tc}), \forall a, y$. If we train our classifier with the algorithm*

$$\min_{g \in G} \max_{a,y} R_{U_a^y}(g), \qquad \text{s.t.} \quad L_{U_a^y}(g, g_{tc}) \leq \mu \quad \forall a, y$$

then the error and unfairness of the optimal solution $\hat{g}$ on the distribution U are bounded with

$$\varepsilon(\hat{g}) \leq \frac{2}{c-1}\varepsilon_U(g_{tc}) + \frac{2c}{c-1}R_U(\hat{g}), \tag{2}$$

$$\Delta_{odds} \leq \frac{2}{c-1}(\gamma + \mu + c\max_{a,y} R_{U_a^y}(\hat{g})) \tag{3}$$

Remark. This sufficient condition suggests us to fit a teacher classifier which is fair on the population distribution, and minimize the *consistency loss* of each group. The unfairness of the resulting model is bounded by the quality (unfairness and error) of the teacher classifier, and the worst-group consistency loss. Intuitively, we can understand the consistency loss as the model invariance to the nuisance factors. With a group-balanced consistency loss, the model would have similar invariance to the nuisance factors resulting in similar group-performance on the unseen data, so that to transfer accuracy and fairness.

4.2 Practical Algorithm: Fair Consistency Regularization

There are two challenges in realizing the theoretical algorithm in Theorem 4.1. First, we need a high-quality teacher model but the model trained with labeled source data is only fair and accurate in source domain. Second, existing consistency regularization methods do not take fairness into consideration. We tackle the first problem by leveraging the iterative self-training paradigm that updates teacher model with student model while training, and thus makes it fairer and fairer. We tackle the second problem by proposing a novel fair consistency regularization.

Algorithm. Figure 2 shows the overall training diagram. There are three major components:
(1) In every training epoch, we use the student model obtained in last epoch as the teacher model, and automatically fit the teacher model by initializing the student model to be the same as the teacher model. In other words, there is only one model training itself iteratively.
(2) To ensure the accuracy and fairness in the source domain, we adopt Laftr [41], an adversarial learning method consisting of a classification loss L_{cls} and a fairness loss L_{fair}.
(3) To transfer fairness and accuracy, we do consistency training on all the data. Following FixMatch [54], we use the pseudo-labels generated by the teacher model as supervision for consistency training where the model should have consistent predictions under transformations. Different from FixMatch, we propose a fair consistency regularization with a balanced group consistency loss $L_{fconsis}$.

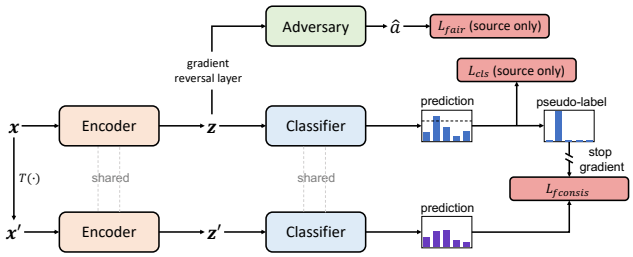


Figure 2: Training diagram.

We train the model with the weighted summation of these three losses as shown in Figure 2. We defer the detailed lose functions of L_{cls} and L_{fair} with a detailed algorithm desciption to Appendix D.

Fair Consistency Regularization. To tighten the upper bound of the unfairness in Theorem 4.1, we need to minimize the worst-group consistency loss. However, the consistency regularization in FixMatch [54] does not distinguish groups which might amplify the bias as observed in [75] and our experiments. Instead, we propose to use a fair consistency regularization that evaluates the consistency loss per group and minimize the balanced consistency loss $L_{fconsis}$ defined as below.

$$L_{fconsis}(g) = \sum_{y=0}^{1}\sum_{a=0}^{1} \lambda_a^y L_a^y(g) \tag{4}$$

$$\text{where} \quad L_a^y(g) = \frac{1}{\sum_{\boldsymbol{x}_a^y}} \sum_{\boldsymbol{x}_a^y} \mathbb{1}(\max(g_{tc}(\boldsymbol{x}_a^y)) \geq \tau) H(\operatorname{argmax}(g_{tc}(\boldsymbol{x}_a^y)), g(t(\boldsymbol{x}_a^y))) \tag{5}$$

where $\boldsymbol{x}_a^y$ denotes an input with sensitive attribute $A = a$ and class $Y = y$. $L_a^y(g)$ is model g's consistency in the group of $\{\boldsymbol{x}_a^y\}$, and λ_a^y is the corresponding weight of the group consistency loss. Here, we abuse the usage of $g(\boldsymbol{x})$ to denote the output logits of model g on input $\boldsymbol{x}$ and thus, $\operatorname{argmax}(g_{tc}(\boldsymbol{x}_a^i))$ is the pseudolabel generated by teacher classifier. $t(\boldsymbol{x}_a^y)$ is the transformed input as defined in Definition 4.1. We use a cross-entropy loss $H(\cdot)$ to encourage the consistency under transformation $t(\cdot)$ and only consider examples that the teacher model has high confidence in with a confidence threshold τ. Note that data is classified into groups according to the true sensitive attribute and pseudolabels. To balance the group consistency loss, we propose to weight each group inversely with the number of confident pseudolabels, and set λ_a^y as

$$\hat{\lambda}_a^y = \frac{1}{\sum_{\boldsymbol{x}_a^y} \mathbb{1}(\max(g_{tc}(\boldsymbol{x}_a^y)) \geq \tau)}, \quad \lambda_a^y = \hat{\lambda}_a^y / \sum_{a,y} \hat{\lambda}_a^y. \tag{6}$$

The weights will dynamically change while training. Heuristically, if the teacher model is only confident in a few examples in a group, the model's consistency in this group is more likely to be low. With the proposed weights, larger penalty will be applied to such groups. Therefore, the proposed fair consistency regularization will enforce the model to pay more attention to higher-error groups.

5 Related Work

This section features related work for transferring of fairness. Other discussion of related work in fair machine learning, domain adaptation and self-training is deferred to Appendix A. Out-of-distribution fairness remains an under-explored area. We categorize prior works into five classes. 1) *Group-wise distribution matching*. [50] derive an upper bound for fairness in the target domain which suggests training a fair model in source domain and matching the distributions of relevant groups from two domains in feature space at the same time. [63] also do group-wise distribution matching but with Wasserstein distance. Such methods are hard to achieve if we do not have supervision in target domain and it also shares the drawback of distribution matching methods. 2) *Reweighting*. When the proportions of groups differ in two domains, reweighting the examples in source domain can approximate the target distribution. [15] use reweighting to deal with fairness problem under covariate shift and [22] use reweighting together with a fairness test to guarantee the fairness under demographic shift. Reweighting methods strongly rely on the support cover assumption which is not satisfied under domain shift. 3) *Distributionally robust optimization (DRO)*. This line of work considers unknown target data that can be any arbitrary weighted combinations of the source dataset, and train a fair

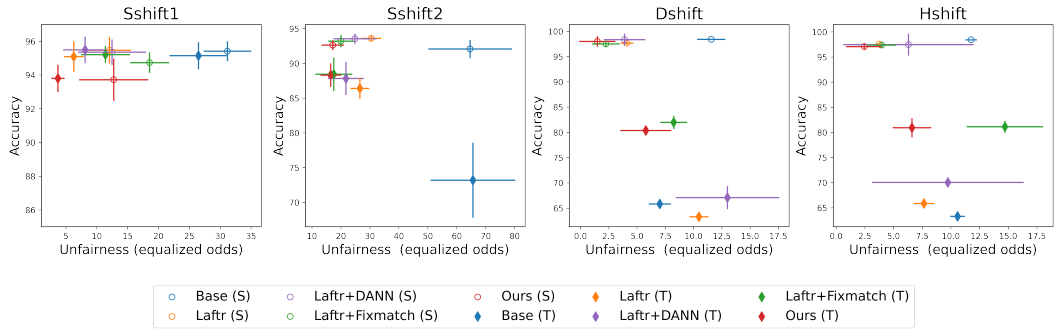


Figure 3: Accuracy and unfairness (error bar denotes the standard deviation) in two domains under subpopulation shifts (Sshift 1, Sshift 2), domain shift (Dshift), and hybrid shift (Hshif). (S) and (T) denotes the evaluation in the source and target domains respectively. Results show that domain shift is more challenging than subpopulation shift, and our method can effectively transfer accuracy and fairness under all the distribution shifts considered.

model that is robust to the worst-case shift [47, 42]. These methods also assume subpopulation shift instead of domain shift. 4) *Causal inference*. [53] do causal domain adaptation and DRO based on a well-characterized causal graph that describes the data construction and distribution shift. Causal methods highly rely on the correct causal graph which is hard to obtain in reality. For example, [49] find that the causal graph in real applications (e.g. predicting the skin condition in dermatology) is far more complicated which violates normal assumptions, and thus making those approaches inapplicable. 5) *Others*. There are also studies aim to maintain fairness under distribution shifts through online learning [68] and loss curvature matching [58]. To the best of our knowledge, this is the first work that uses self-training to transfer fairness. Some work also studies self-supervised learning and fairness, yet they use unlabeled data and self-training to improve the in-distribution fairness [13, 67, 9] which is different from our goal.

6 Experiments

6.1 Evaluation under Different Types of Distribution Shift with a Synthetic Dataset

In order to study the fairness under distribution shifts and verify our theoretical findings, we develop a synthetic dataset to simulate different types of distribution shifts.

Synthetic dataset. The synthetic dataset is adapted from the 3dshapes dataset [30] which contains images of 3D objects generated from six independent latent factors (*shape*, *object hue*, *scale*, *orientation*, *floor hue*, *wall hue*). This dataset satisfies our assumption on the underlying data generation process very well. We simulate different types of distribution shift by shifting the marginal distributions of latent factors. By sampling images according to different marginal distributions of latent factors, we are able to simulate distribution shifts (see Appendix D.1 for details).

Distribution shifts. We set the image as input X, and select three latent factors to be class ($Y =$ *shape*), sensitive attribute ($A =$ *object hue*), and a nuisance factor that might shift ($D =$ *scale*). We consider four widely observed distribution shifts in reality ($\mathbb{P}_S(X, Y, A, D) \neq \mathbb{P}_T(X, Y, A, D)$):
(1) **Sshift 1**: Subpopulation shift where only the nuisance factor shift (i.e. more small objects in source but more large objects in target), $\mathbb{P}_S(Y, A) = \mathbb{P}_T(Y, A)$, $\mathbb{P}_S(D) \neq \mathbb{P}_T(D)$.
(2) **Sshift 2**: Subpopulation shift where A and Y have different correlations in two domains (i.e. most red objects are cubes in source but are capsules in target), $\mathbb{P}_S(Y, A) \neq \mathbb{P}_T(Y, A)$, $\mathbb{P}_S(D) = \mathbb{P}_T(D)$.
(3) **Dshift**: Domain shift where the nuisance factor has different sample spaces (i.e. small objects in source but large objects in target), $\mathbb{P}_S(Y, A) = \mathbb{P}_T(Y, A)$, $\mathbb{P}_S(D) \neq \mathbb{P}_T(D)$, $\mathcal{Y}_S^d \neq \mathcal{Y}_T^d$.
(4) **Hshift**: Hybrid shift of (2) and (3).

Baselines. We do shape classification task with an MLP model and compare our method with four baselines: Base (standard ERM); Laftr; Laftr+DANN (a combination of Laftr and a domain adaptation method [21]); Laftr+FixMatch. In our method, we also use Laftr and FixMatch but with the proposed fair consistency regularization. Since the shifted nuisance factor is *scale*, we use random

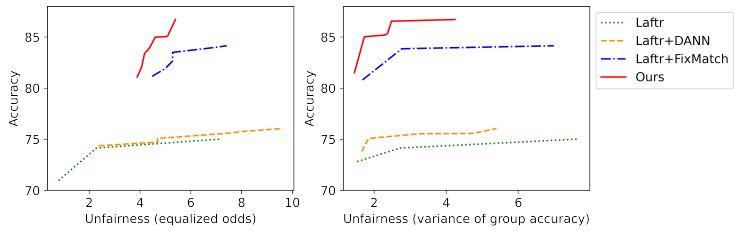
padding and cropping as transformations in our method and Laftr+FixMatch. We train Base and Laftr with labeled source data, and train others with unlabeled target data as well.

Domain shift is more challenging than subpopulation shift. Figure 3 shows that under subpopulation shifts, the source fair model trained with Laftr also has high accuracy and fairness in the target domain although it has not seen any target date. We observe that this is because the sample space is shared (e.g. small and large objects both exists in the source data), and the model has similar performance under all factor values. Thus, the good performance remains even if the proportion of data changes, verified Proposition 3.1. In contrast, under domain shifts and hybrid shifts, the source fair model performs poorly on data from a different sample space, suggesting the difficulty of domain shift.

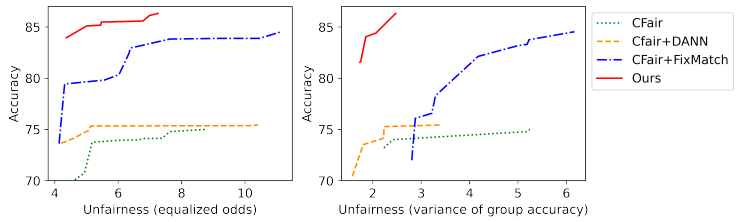
Our method can transfer fairness and accuracy under various types of distribution shift. Under domain shifts, the domain adaptation method DANN does not help in transferring fairness or accuracy. Consistency regularization enforces the model to behave consistently under cropping and flipping, resulting in a model that has similar predictions regardless of the object's scale, and thus to transfer accuracy. However, it may cause bias as shown in the results of Laftr+FixMatch. With the proposed fair consistency regularization, the model gains similar consistency for all groups, resulting in similar performance on all groups in the target domain. Therefore, our method achieves both high accuracy and fairness in two domains under all the considered distribution shifts.

	Source			Target		
	Acc	Unfairness		Acc	Unfairness	
Method		V_{acc}	Δ_{odds}		V_{acc}	Δ_{odds}
Base	92.85±0.49	2.30±0.97	4.81±0.69	74.49±0.83	5.79±3.49	9.90±1.27
Laftr	93.24±0.41	1.19±0.46	2.44±0.51	74.35±1.46	6.92±0.72	9.79±1.54
CFair	92.51±0.22	1.76±0.53	4.75±0.85	73.53±0.89	7.51±0.73	7.26±1.95
Laftr+DANN	91.33±0.08	2.12±1.72	2.70±0.67	74.28±1.63	6.25±2.59	8.27±2.11
CFair+DANN	90.89±0.76	2.01±0.70	4.43±1.36	74.62±1.06	6.23±0.90	5.26±2.07
Laftr+FixMatch	96.62±0.06	0.77±0.21	2.23±0.44	83.87±0.48	8.21±0.67	9.32±1.01
CFair+FixMatch	96.13±0.53	1.28±0.53	2.78±0.74	83.11±0.49	7.87±1.86	7.89±0.40
Ours (w/ Laftr)	96.08±0.07	0.96±0.39	2.59±0.35	85.52±0.40	2.82±0.87	5.70±0.52
Ours (w/ CFair)	95.65±0.22	1.56±0.37	3.85±0.97	84.48±0.42	2.88±0.99	5.43±0.65

Table 1: Transfer fairness and accuracy from UTKFace to FairFace



(a) Pareto frontiers of ours (w/ Laftr) and baselines in the target domain.

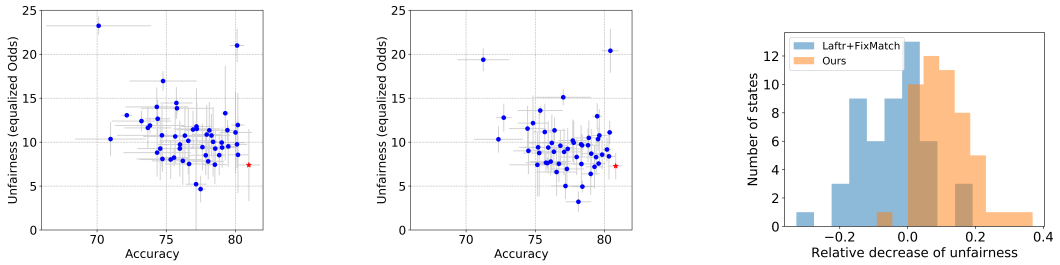


(b) Pareto frontiers of ours (w/ Cfair) and baselines in the target domain.

Figure 4: Comparison of Pareto frontiers. Upper left is preferred. Our method outperforms baseline methods in achieving accuracy and fairness at the same time.

6.2 Evaluation on Real datasets

Evaluation on images. We use UTKFace [70] as the source data and FairFace [29] as the target data. Although both of them are facial images, there is a distribution shift between them due to different



(a) Laftr (b) Ours (c) Decrease of unfairness

Figure 5: Unfairness and accuracy tested on NewAdult. CA as the source domain (red star) and other states as the target domain (blue dots).

image sources. We consider a gender classification task with the race as the sensitive attribute. We use VGG16 [52] as the model and RandAugment [17] (excluding transformations that may change the group) as the transformation function. Additional to previous baselines, we also compare with Laftr+MMD [38] (a domain adaptation method), and use CFair [72] as the method for in-distribution fairness. As shown in Table 1, there is indeed a distribution shift as the source model trained with Laftr or CFair is no longer accurate or fair in the target domain. Two domain adaptation methods are not able to transfer accuracy and may harm the fairness. As expected, self-training (Laftr+Fixmatch and CFair+Fixmatch) significantly improves the accuracy in the target domain but the unfairness is high. With the proposed fair consistency regularization, our method outperforms it remarkably on fairness with a 70% decrease of the group accuracy variance.

Evaluation on tabular data. We further evaluate our method on NewAdult dataset [18] which contains census data from all states of the United States. We do income classification task with an MLP as the model. We set CA as the source domain and all the other states as the target domain. We use random perturbation on tabular data (see details in Appendix D) as the transformations. Results are shown in Figure 5. When applying to other states, the fair model trained on CA becomes unfair, as shown in Figure 5a. Our method improves the fairness in most of the states with a slight improvement of accuracy as well.

6.3 Ablation Study

Fair consistency is essential in transferring fairness. To see whether boosted consistency improves accuracy and whether imbalanced consistency causes unfairness as Theorem 4.1 suggests, we evaluate the accuracy and consistency for every group on target data in UTK-FairFace. The consistency is measured by testing the model's agreement on the outputs under two different random transformations. As shown in Figure 6, groups that obtain higher consistency have higher accuracy, which validates the ability of consistency regularization for transferring accuracy. The training methods that use standard consistency regularization (e.g. Laftr+FixMatch) have been observed to be unfair in the target domain. Figure 6 shows that it is because the model has imbalanced group consistency. With our fair consistency regularization, the model gains similar consistency for all groups, resulting in similar group accuracy. Such finding verifies that fair consistency is essential in transferring fairness.

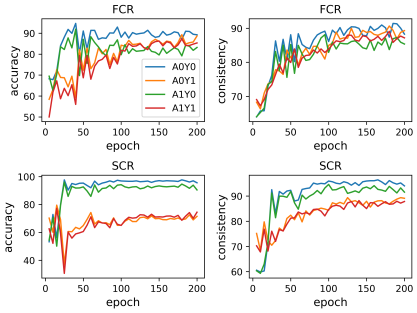


Figure 6: Per-group accuracy and consistency evaluated on the target data. Accuracy is correlated with consistency. Comparing with the standard consistency regularization (SCR), the model trained with fair consistency regularization (FCR) has more balanced consistency and accuracy.

The role of transformation. We design transformation functions based on our domain knowledge of the latent factors. To satisfy *intra-group expansion assumption*, the transformations should be group-preserving (i.e. do not change class and sensitive attribute), so we exclude transformations such as color jittering in RandAugment when *race* is the sensitive attribute. To investigate the importance of transformations, we test a weaker set of transformations (cropping and flipping) and report the performance in Table 2. Comparing with RandAugment (Table 1), consistency to only cropping and flipping is less effective in transferring

Method	Source			Target		
	Acc	Unfairness		Acc	Unfairness	
		V_{acc}	Δ_{odds}		V_{acc}	Δ_{odds}
Laftr+FixMatch	94.08±0.70	1.64±0.46	3.51±1.46	77.05±0.26	12.23±3.83	6.55±1.54
CFair+FixMatch	94.09±0.33	0.97±0.36	2.16±0.97	77.25±0.21	12.93±2.66	9.77±0.95
Ours (w/ Laftr)	94.25±0.22	1.06±0.46	2.09±0.55	77.32±0.21	2.35±1.67	4.27±1.41
Ours (w/ CFair)	94.24±0.26	1.67±0.38	4.43±0.63	77.96±0.38	3.34±1.08	5.70±1.14

Table 2: Transfer fairness and accuracy from UTKFace to FairFace with weak transformations

accuracy since the neighbor generated by them is much smaller. The limited transformations also restrict the performance of our method on tabular data (see Appendix E). Though the ability of transferring accuracy is limited by weak transformations, our method can always make the transferring process fair.

The role of components in fair consistency regularization. Table 3 shows that the consistency in both domains matters. Giving every group the same weight instead of using dynamic weights results in increased unfairness. Fixing the teacher classifier to be the fair source model, we observe a significantly decrease of the accuracy, suggesting the important role of iterative self-training in our algorithm.

Method	Acc	Unfairness	
		V_{acc}	Δ_{odds}
Ours	85.52±0.40	2.82±0.87	5.70±0.52
w/o consistency in target	82.43±1.05	6.80±1.30	5.85±0.40
w/o consistency in source	82.5±1.58	6.63±0.71	8.18±1.27
w/o dynamic weights	84.34±0.19	6.86±0.50	7.68±0.81
w/o updating g_{tc}	79.13±0.52	3.49±0.63	6.65±1.31

Table 3: Ablation study on UTK-FairFace task

7 Conclusion

In this paper, we explore transferring fairness under distribution shifts. We present a theory-guided method of self-training based on an intra-group expansion assumption. To provide fair consistency regularization, we suggest dynamically reweighting consistency loss with the number of confident pseudolabels. We also offer a synthetic dataset simulating different types of distribution shifts to examine this topic. Our technique and other self-training methods rely on a well-defined data transformation set. Future work will relax this limitation for application to more real-world problems.

Acknowledgements

This work is supported by National Science Foundation NSF-IIS-FAI program, DOD-ONR-Office of Naval Research, DOD-DARPA-Defense Advanced Research Projects Agency Guaranteeing AI Robustness against Deception (GARD), and Adobe, Capital One and JP Morgan faculty fellowships.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 2018.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010.

[3] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation, 2021.

[4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations, 2017.

[5] Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1170–1182. PMLR, 18–24 Jul 2021.

[6] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.

[7] Simon Caton and Christian Haas. Fairness in machine learning: A survey, 2020.

[8] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 319–328, New York, NY, USA, 2019. Association for Computing Machinery.

[9] Joymallya Chakraborty, Huy Tu, Suvodeep Majumder, and Tim Menzies. Can we achieve fairness using semi-supervised learning? *arXiv preprint arXiv:2111.02038*, 2021.

[10] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems*, 33:21061–21071, 2020.

[11] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[12] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[13] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning, 2018.

[15] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor, editors, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 91–98. ACM, 2019.

[16] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR, 2019.

[17] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[18] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[19] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2796–2806, 2018.

[20] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness, 2016.

[21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[22] Stephen Giguere, Blossom Metevier, Yuriy Brun, Philip S. Thomas, Scott Niekum, and Bruno Castro da Silva. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*, 2022.

[23] Bruce Glymour and Jonathan Herington. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In danah boyd and Jamie H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 269–278. ACM, 2019.

[24] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021.

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[26] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[27] 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 Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR, 10–15 Jul 2018.

[28] Taotao Jing, Bingrong Xu, and Zhengming Ding. Towards fair knowledge transfer for imbalanced domain adaptation. *IEEE Transactions on Image Processing*, 30:8200–8211, 2021.

[29] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[30] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR, 10–15 Jul 2018.

[31] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[32] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020.

[33] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076, 2017.

[34] Chao Lan and Jun Huan. Discriminatory transfer. *arXiv preprint arXiv:1707.00780*, 2017.

[35] Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Pengfei Xu, Wei Zhou, Yoshua Bengio, and Kurt Keutzer. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*, 2020.

[36] Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[37] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation, 2021.

[38] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[39] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[40] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[41] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3381–3390. PMLR, 2018.

[42] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. *Advances in neural information processing systems*, 33:18445–18456, 2020.

[43] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[44] Luca Oneto and Silvia Chiappa. Fairness in machine learning. *Studies in Computational Intelligence*, page 155–196, 2020.

[45] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[46] Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero-Soriano, Samira Shabanian, and Sina Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[47] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. *CoRR*, abs/2010.05166, 2020.

[48] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori B. Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the wilds benchmark for unsupervised adaptation. *ArXiv*, abs/2112.05090, 2021.

[49] Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine Heller, Silvia Chiappa, and Alexander D'Amour. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications?, 2022.

[50] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. Transfer of machine learning fairness across domains. *CoRR*, abs/1906.09688, 2019.

[51] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 3–13. ACM, 2021.

[54] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[55] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173. PMLR, 2019.

[56] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33:19276–19289, 2020.

[57] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[58] Haotao Wang, Junyuan Hong, Jiayu Zhou, and Zhangyang Wang. Equalized robustness: Towards sustainable fairness under distributional shifts, 2022.

[59] Tongxin Wang, Zhengming Ding, Wei Shao, Haixu Tang, and Kun Huang. Towards fair cross-domain adaptation via generative learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 454–463, 2021.

[60] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[61] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.

[62] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pages 6872–6881. PMLR, 2019.

[63] Tae-Ho Yoon, Jaewook Lee, and Woojin Lee. Joint transfer of model knowledge and fairness over domains using wasserstein distance. *IEEE Access*, 8:123783–123798, 2020.

[64] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 20–22 Apr 2017.

[65] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[66] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[67] Tao Zhang, tianqing zhu, Jing Li, Mengde Han, Wanlei Zhou, and Philip Yu. Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020.

[68] Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C Weiss, and Wolfgang Nejdl. Farf: A fair and adaptive random forests classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 245–256. Springer, 2021.

[69] Yabin Zhang, Haojian Zhang, Bin Deng, Shuai Li, Kui Jia, and Lei Zhang. Semi-supervised models are strong unsupervised domain adaptation learners, 2021.

[70] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[71] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2019.

[72] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[73] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.

[74] Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. *Advances in Neural Information Processing Systems*, 34, 2021.

[75] Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section F
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Section D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] see Section D
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material

A More on Related Work

Fair machine learning. Generally, fair machine learning methods fall into three categories: pre-processing, in-processing, and post-processing [43, 7]. In this paper, we focus on in-processing methods that modify learning algorithms to remove discrimination during the training process. As for fair classification, several approaches have been proposed including fair representation learning [65, 40, 4, 66, 41, 55, 16, 72], fairness-constrained optimization [19, 1], causal methods [33, 23, 44], and many other approaches with different techniques [8, 12, 24]. All of those works are for in-distribution fairness and we investigate out-of-distribution fairness in this paper. We use LAFTR [41], an adversarial learning method that shows advanced performance on fairness [46], to learn a fair model in the source domain and adapt it to the target domain. We also test CFair[71] in our experiments. Many metrics of fairness have been proposed [14] including demographic parity [6], equalized opportunity, and equalized odds [26] which are most widely adopted. In this paper, we use equalized odds to measure unfairness in both domains.

Distribution shifts. In many real-world applications, distribution shifts are unavoidable. The goal of existing work addressing distribution shifts is simply to transfer accuracy. [31] propose a benchmark of in-the-wild datasets to study the real distribution shifts. We follow their category of distribution shifts, including subpopulation shifts and domain shifts. Their empirical results on many state-of-the-art methods show that self-training outperforms others on image datasets significantly while have limited performance due to the limited data augmentation on non-image modalities [48]. This finding aligns with our experimental results. [61] conduct a fine-grained analysis of various distribution shifts based on an underlying data generation assumption which is similar to ours. They also use 3dshapes dataset to simulate different types of distribution shifts. Additional to accuracy, we aim at transferring fairness at the same time in this paper.

Domain adaptation and self-training. Inspired by the theoretical work [2], numerous distribution matching approaches have emerged for domain adaption over the past decade. Domain-adversarial training [21] and many of its variants [57, 39, 27, 56] that aim at matching the distribution of two domains in the feature space have shown encouraging results in many applications. However, recent studies [62, 73, 35] show that such methods may fail in many cases since they only optimize part of the theoretical bound. We test DANN [21] and MMD [38], two distribution matching methods in our experiments, and also find them less effective in transferring accuracy and fairness. Recently, another line of work that uses self-training draws increasing attention [69, 3]. Those methods enjoy guarantees [60, 5] and demonstrate superior empirical results with desirable properties such as robustness to spurious features [32, 10, 37] and robustness to dataset imbalance [36]. However, all of those work on domain adaptation only aims at transferring accuracy. Although there is work that study fairness issues in current domain adaptation methods [34] and propose to alleviate it by balancing the data [28, 59, 75], fair domain adaptation is still under-explored. Based on the findings that model's consistency to input transformations is important to generalization [74] and is a core component of self-training [51, 54, 25], we improve the consistency regularization in [54] to achieve fair transferring.

B Proof and More Discussion of Fairness under Distribution Shifts

Lemma B.1. *Under Assumption 1, for a subpopulation shift that caused by the shift of marginal distribution of factor Y^i, we have $\mathbb{P}_S(X|Y^i = y^i) = \mathbb{P}_T(X|Y^i = y^i), \forall y^i \in \mathcal{Y}^i$.*

Proof. Under Assumption 1, $\mathbb{P}_S(X|Y^{1:K} = \boldsymbol{y}^{1:K}) = \mathbb{P}_T(X|Y^{1:K} = \boldsymbol{y}^{1:K})$. Since the shift only happens on the factor Y^i, the marginal distribution of other factors remains the same in two domains,

$\mathbb{P}_S(Y^{\{1:K\}\setminus i}) = \mathbb{P}_T(Y^{\{1:K\}\setminus i})$ where we use $\{1:K\}\setminus i$ to denote $1,..,i-1,i+1,...,K$. Then

$$\begin{aligned}
\mathbb{P}_S(X|Y^i=y^i) &= \sum_{\boldsymbol{y}^{\{1:K\}\setminus i}} \mathbb{P}_S(X, Y^{\{1:K\}\setminus i}=\boldsymbol{y}^{\{1:K\}\setminus i}|Y^i=y^i) \\
&= \sum_{\boldsymbol{y}^{\{1:K\}\setminus i}} \mathbb{P}_S(Y^{\{1:K\}\setminus i}=\boldsymbol{y}^{\{1:K\}\setminus i})\mathbb{P}_S(X|Y^i=y^i, Y^{\{1:K\}\setminus i}=\boldsymbol{y}^{\{1:K\}\setminus i}) \\
&= \sum_{\boldsymbol{y}^{\{1:K\}\setminus i}} \mathbb{P}_T(Y^{\{1:K\}\setminus i}=\boldsymbol{y}^{\{1:K\}\setminus i})\mathbb{P}_T(X|Y^i=y^i, Y^{\{1:K\}\setminus i}=\boldsymbol{y}^{\{1:K\}\setminus i}) \\
&= \mathbb{P}_T(X|Y^i=y^i)
\end{aligned}$$

where the second line holds because of the independence of the latent factors $Y^1,...,Y^K$. □

Now, we restate Proposition 3.1 and provide the proof.

Proposition B.1 (Transfer of fairness under subpopulation shift)**.** *Consider the subpopulation shift that is caused by the shifted marginal distribution of a nuisance factor Y^i (i.e., $\mathbb{P}_S(Y^i) \neq \mathbb{P}_T(Y^i)$), while $\mathcal{Y}^i_S = \mathcal{Y}^i_T = \mathcal{Y}^i$. If model f is strictly fair in source domain under any value of factor Y^i satisfying $\mathbb{P}_S(f(X)=y^l|Y^a=0,Y^l=y^l,Y^i=y^i) = \mathbb{P}_S(f(X)=y^l|Y^a=1,Y^l=y^l,Y^i=y^i), \forall y^i\in\mathcal{Y}^i, y^l\in\{0,1\}$, then f is also fair in target domain with $\Delta_{odds}=0$.*

Proof. In the target domain, the equalized odds (unfairness) is defined as

$$\Delta_{odds} = \frac{1}{2}\sum_{y^l=0}^{1}\left|\mathbb{P}_T(f(X)=y^l|Y^a=0,Y^l=y^l) - \mathbb{P}_T(f(X)=y^l|Y^a=1,Y^l=y^l)\right|.$$

Since all latent factors are independent, we have

$$\mathbb{P}_T\left(f(X)=y^l|Y^a=0,Y^l=y^l\right) = \sum_{y^i\in\mathcal{Y}^i}\mathbb{P}_T(Y^i=y^i)\mathbb{P}_T\left(f(X)=y^l|Y^a=0,Y^l=y^l,Y^i=y^i\right)$$

and

$$\mathbb{P}_T\left(f(X)=y^l|Y^a=1,Y^l=y^l\right) = \sum_{y^i\in\mathcal{Y}^i}\mathbb{P}_T(Y^i=y^i)\mathbb{P}_T\left(f(X)=y^l|Y^a=1,Y^l=y^l,Y^i=y^i\right).$$

Therefore, the Δ_{odds} in the target domain can be decomposed into

$$\begin{aligned}
\Delta_{odds} = \frac{1}{2}\sum_{y^l=0}^{1}\Big|\sum_{y^i\in\mathcal{Y}^i}\mathbb{P}_T(Y^i=y^i)\big(&\mathbb{P}_T\left(f(X)=y^l|Y^a=0,Y^l=y^l,Y^i=y^i\right) \\
&- \mathbb{P}_T\left(f(X)=y^l|Y^a=1,Y^l=y^l,Y^i=y^i\right)\big)\Big|.
\end{aligned}$$

Since two domains share the same underlying data generative model, and the distribution shift is caused by the shift of the marginal distribution of factor Y^i, from Lemma B.1, we have

$$\mathbb{P}_T\left(X|Y^a=0,Y^l=y^l,Y^i=y^i\right) = \mathbb{P}_S\left(X|Y^a=0,Y^l=y^l,Y^i=y^i\right)$$

and

$$\mathbb{P}_T\left(X|Y^a=1,Y^l=y^l,Y^i=y^i\right) = \mathbb{P}_S\left(X|Y^a=1,Y^l=y^l,Y^i=y^i\right).$$

Thus the conditional distribution of the model's prediction also remains, as

$$\mathbb{P}_T\left(f(X)|Y^a=0,Y^l=y^l,Y^i=y^i\right) = \mathbb{P}_S\left(f(X)|Y^a=0,Y^l=y^l,Y^i=y^i\right)$$

and

$$\mathbb{P}_T\left(f(X)|Y^a=1,Y^l=y^l,Y^i=y^i\right) = \mathbb{P}_S\left(f(X)|Y^a=1,Y^l=y^l,Y^i=y^i\right).$$

In this case, if the source model is strictly fair that $\forall y^i\in\mathcal{Y}^i, y^l\in\{0,1\}$ the following holds

$$\mathbb{P}_S(f(X)=y^l|Y^a=0,Y^l=y^l,Y^i=y^i) = \mathbb{P}_S(f(X)=y^l|Y^a=1,Y^l=y^l,Y^i=y^i),$$

then it is also fair in the target domain with

$$\begin{aligned}
\Delta_{odds} = \frac{1}{2}\sum_{y^l=0}^{1}\Big|\sum_{y^i\in\mathcal{Y}^i}\mathbb{P}_T(Y^i=y^i)\big(&\mathbb{P}_S\left(f(X)=y^l|Y^a=0,Y^l=y^l,Y^i=y^i\right) \\
&- \mathbb{P}_S\left(f(X)=y^l|Y^a=1,Y^l=y^l,Y^i=y^i\right)\big)\Big| = 0.
\end{aligned}$$

□

This proposition explains why the fair source model is also fair in the target domain under Sshift 1 in our experiments (see Section 6.1). In addition to shifts of nuisance factors, the subpopulation shifts can also be caused by the marginal distribution shift of the label and sensitive attribute. The following proposition argues that the fair model is also in the target domain under such distribution shifts.

Proposition B.2 (Transfer of fairness under subpopulation shift of sensitive attribute)**.** *Consider the subpopulation shift that is caused by the shifted marginal distribution of sensitive attribute Y^a (i.e., $\mathbb{P}_S(Y^a) \neq \mathbb{P}_T(Y^a)$), while $\mathcal{Y}_S^a = \mathcal{Y}_T^a = \mathcal{Y}^a = \{0,1\}$. If model f is fair in the source domain with $\Delta_{odds}^S = 0$, then it is also fair in the target domain with $\Delta_{odds}^T = 0$.*

Proof. The proof is similar to the previous one. Since

$$\Delta_{odds}^S = \frac{1}{2}\sum_{y^l=0}^{1} \left|\mathbb{P}_S(f(X)=y^l|Y^a=0,Y^l=y^l) - \mathbb{P}_S(f(X)=y^l|Y^a=1,Y^l=y^l)\right|,$$

and from Lemma B.1 we know that

$$\mathbb{P}_S(X|Y^a=0,Y^l=y^l) = \mathbb{P}_T(X|Y^a=0,Y^l=y^l)$$
$$\mathbb{P}_S(X|Y^a=1,Y^l=y^l) = \mathbb{P}_T(X|Y^a=1,Y^l=y^l),$$

thus,

$$\begin{aligned}\Delta_{odds}^T &= \frac{1}{2}\sum_{y^l=0}^{1} \left|\mathbb{P}_T(f(X)=y^l|Y^a=0,Y^l=y^l) - \mathbb{P}_T(f(X)=y^l|Y^a=1,Y^l=y^l)\right| \\ &= \frac{1}{2}\sum_{y^l=0}^{1} \left|\mathbb{P}_S(f(X)=y^l|Y^a=0,Y^l=y^l) - \mathbb{P}_S(f(X)=y^l|Y^a=1,Y^l=y^l)\right| \\ &= \Delta_{odds}^S = 0.\end{aligned}$$

□

Such result also holds for subpopulation shifts that caused by the shift of label Y^l. This proposition explains why the fair source model is also fair in the target domain under Sshift 2 in our experiments (see Section 6.1). It suggests that encouraging fairness is able to alleviate spurious correlation. We leave more studies on this interesting finding to future work.

Remark. All the above analyses are based on the population distribution where $\mathbb{P}_S(f(X)=y^l|Y^a=0,Y^l=y^l) = \mathbb{E}_{\mathbb{P}_S(X,Y^{1:K})}(f(X)=y^l|Y^a=0,Y^l=y^l)$. In practice, it is estimated by finite samples. Insufficient samples would cause estimation error of fairness and bring another challenge for transferring fairness. In this paper, we only consider the fairness measured by population distribution. Future work will investigate the impact of estimation error on transferring fairness, and the way to resolve it.

C Proof of the Sufficient Condition for Transferring Fairness

Our proof is based on the theory in [60].

Theorem C.1. *(Restatement of Lemma A.8 in [60]) We assume that U_a^y satisfies $(\bar{\alpha},\bar{c})$-multiplicative expansion for $\varepsilon_{U_a^y}(g_{tc}) \leq \bar{\alpha} < 1/3$ and $\bar{c} > 3$. We define $c \triangleq \min\{1/\bar{\alpha},\bar{c}\}$. Then for any classifier $g:\mathcal{X}\to\mathcal{Y}$, the error of it on the group U_a^y is upper bounded as:*

$$\varepsilon_{U_a^y}(g) \leq \frac{c+1}{c-1}L_{U_a^y}(g,g_{tc}) + \frac{2c}{c-1}R_{U_a^y}(g) - \varepsilon_{U_a^y}(g_{tc})$$

Theorem C.2. *(A restricted version of the above theorem) We assume that U_a^y satisfies $(\bar{\alpha},\bar{c})$-multiplicative expansion for $\varepsilon_{U_a^y}(g_{tc}) \leq \bar{\alpha} < 1/3$ and $\bar{c} > 3$. We define $c \triangleq \min\{1/\bar{\alpha},\bar{c}\}$. Then for any classifier $g:\mathcal{X}\to\mathcal{Y}$ satisfies $L_{U_a^y}(g,g_{tc}) \leq \varepsilon_{U_a^y}(g_{tc})$, the error of it on the group U_a^y is upper bounded as:*

$$\varepsilon_{U_a^y}(g) \leq \frac{2}{c-1}\varepsilon_{U_a^y}(g_{tc}) + \frac{2c}{c-1}R_{U_a^y}(g)$$

Proof.

$$\varepsilon_{U_a^y}(g) \leq \frac{c+1}{c-1} L_{U_a^y}(g, g_{tc}) + \frac{2c}{c-1} R_{U_a^y}(g) - \varepsilon_{U_a^y}(g_{tc})$$
$$\varepsilon_{U_a^y}(g) \leq \frac{2}{c-1} \varepsilon_{U_a^y}(g_{tc}) + \frac{2c}{c-1} R_{U_a^y}(g) \qquad \text{(because } L_{U_a^y}(g, g_{tc}) \leq \varepsilon_{U_a^y}(g_{tc}))$$

□

Theorem C.3. *If $L_{U_a^y}(g, g_{tc}) \leq \varepsilon_{U_a^y}(g_{tc})$, we have*

$$\varepsilon_{U_a^y}(g) \geq \varepsilon_{U_a^y}(g_{tc}) - L_{U_a^y}(g, g_{tc})$$

Proof. By triangle inequality. □

Now, we restate Theorem 4.1 and provide the proof.

Theorem C.4. *Suppose we have a teacher classifier g_{tc} with bounded unfairness such that $|\varepsilon_{U_a^y}(g_{tc}) - \varepsilon_{U_{a'}^{y'}}(g_{tc})| \leq \gamma, \forall a, a' \in \mathcal{A}$ and $y, y' \in \mathcal{Y}$. We assume intra-group expansion where U_a^y satisfies $(\bar{\alpha}, \bar{c})$-multiplicative expansion and $\varepsilon_{U_a^y}(g_{tc}) \leq \bar{\alpha} < 1/3$ and $\bar{c} > 3, \forall a, y$. We define $c \triangleq \min\{1/\bar{\alpha}, \bar{c}\}$, and set $\mu \leq \varepsilon_{U_a^y}(g_{tc}), \forall a, y$. If we train our classifier with the algorithm*

$$\min_{g \in G} \max_{a,y} R_{U_a^y}(g) \tag{7}$$
$$\text{s.t.} \quad L_{U_a^y}(g, g_{tc}) \leq \mu \quad \forall a, y$$

then the error and unfairness of the optimal solution $\hat{g}$ on the distribution U are bounded with

$$\varepsilon(\hat{g}) \leq \frac{2}{c-1} \varepsilon_U(g_{tc}) + \frac{2c}{c-1} R_U(\hat{g}),$$
$$\Delta_{odds} \leq \frac{2}{c-1} (\gamma + \mu + c \max_{a,y} R_{U_a^y}(\hat{g})).$$

Proof. The upper bound of error is derived from Theorem C.2. For the unfairness, by definition

$$\Delta_{odds} = \frac{1}{2} \left(\left| \varepsilon_{U_0^0}(\hat{g}) - \varepsilon_{U_1^0}(\hat{g}) \right| + \left| \varepsilon_{U_0^1}(\hat{g}) - \varepsilon_{U_1^1}(\hat{g}) \right| \right).$$

Based on the upper bound of group error from Theorem C.2, and the lower bound of it from Theorem C.3, we have

$$\begin{aligned}
&\left| \varepsilon_{U_0^0}(\hat{g}) - \varepsilon_{U_1^0}(\hat{g}) \right| \\
&\leq \max \Big\{ \frac{2}{c-1}\gamma + \frac{2}{c-1} L_{U_1^0}(\hat{g}, g_{tc}) + \frac{2c}{c-1} R_{U_0^0}(\hat{g}), \\
&\qquad \frac{2}{c-1}\gamma + \frac{2}{c-1} L_{U_0^0}(\hat{g}, g_{tc}) + \frac{2c}{c-1} R_{U_1^0}(\hat{g}) \Big\} \qquad \text{(because } c > 3) \\
&= \frac{2}{c-1}\gamma + \frac{2}{c-1} \max \left\{ L_{U_1^0}(\hat{g}, g_{tc}) + c R_{U_0^0}(\hat{g}), L_{U_0^0}(\hat{g}, g_{tc}) + c R_{U_1^0}(\hat{g}) \right\} \\
&\leq \frac{2}{c-1} (\gamma + \mu + c \max_a R_{U_a^0}(\hat{g})).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Delta_{odds} &\leq \frac{2}{c-1} \left(\gamma + \mu + \frac{c}{2} \left(\max_a R_{U_a^0}(\hat{g}) + \max_a R_{U_a^1}(\hat{g}) \right) \right) \\
&\leq \frac{2}{c-1} (\gamma + \mu + c \max_{a,y} R_{U_a^y}(\hat{g})).
\end{aligned}$$

□

Upper bound of V_{acc}. From Theorem C.2 we know that the group accuracy is upper bounded by $\varepsilon_{U_a^y}(g) \leq \frac{2}{c-1} \varepsilon_{U_a^y}(g_{tc}) + \frac{2c}{c-1} R_{U_a^y}(g)$. The variance of group accuracy is defined as

$$\begin{aligned}
V_{acc} &= Var(\{\mathbb{P}(\hat{Y} = y | A = a, Y = y), \forall a, y\}) \\
&= Var(\{\varepsilon_{U_a^y}(g), \forall a, y\})
\end{aligned}$$

If we assume the same estimation error for all the groups when we use the upper bound to estimate the group accuracy with $\varepsilon_{U_a^y}(g) = \frac{2}{c-1}\varepsilon_{U_a^y}(g_{tc}) + \frac{2c}{c-1}R_{U_a^y}(g)$, then

$$V_{acc} = Var\left(\left\{\frac{2}{c-1}\varepsilon_{U_a^y}(g_{tc}) + \frac{2c}{c-1}R_{U_a^y}(g), \forall a, y\right\}\right)$$

When the teacher classifier has bounded unfairness with $|\varepsilon_{U_a^y}(g_{tc}) - \varepsilon_{U_{a'}^{y'}}(g_{tc})| \leq \gamma, \forall a, a', y, y'$, the variance of group accuracy would be mainly affected by the variance of group consistency loss $Var(\{R_{U_a^y}(g), \forall a, y\})$. Therefore, this upper bound also suggests us to balance the consistency loss while minimizing it.

Multi-sensitive attribute and multi-class cases. It is obvious that the Theorem 4.1 still holds for the binary-sensitive attribute and multi-class case where $\mathcal{Y} = \{1, 2, .., M\}$. For the multi-sensitive attribute case, the key problem is how to define the unfairness. If we define the equalized odds in general cases to be the following one, then it is easy to see that Theorem 4.1 still holds.

$$\Delta_{odds} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\max_{a,a'\in\mathcal{A}}\left|\varepsilon_{U_a^y}(g) - \varepsilon_{U_{a'}^y}(g)\right|$$

D Details of Experiments

D.1 Synthetic Dataset

The 3dshapes dataset [2] [30] contains 480000 RGB images (the size is $64 \times 64 \times 3$) of 3D objects. Every image is generated by six latent factors (*shape*, *object hue*, *scale*, *orientation*, *floor hue*, *wall hue*) which are annotated along with images. The sample spaces of these factors are shown in Table 4.

Factor	Sample space
shape	4 values in [0, 1, 2, 3]
object hue	10 values linearly spaced in [0, 1]
scale	8 values linearly spaced in [0, 1]
orientation	15 values linearly spaced in [-30, 30]
floor hue	10 values linearly spaced in [0, 1]
wall hue	10 values linearly spaced in [0, 1]

Table 4: Latent factors in 3dshapes dataset.

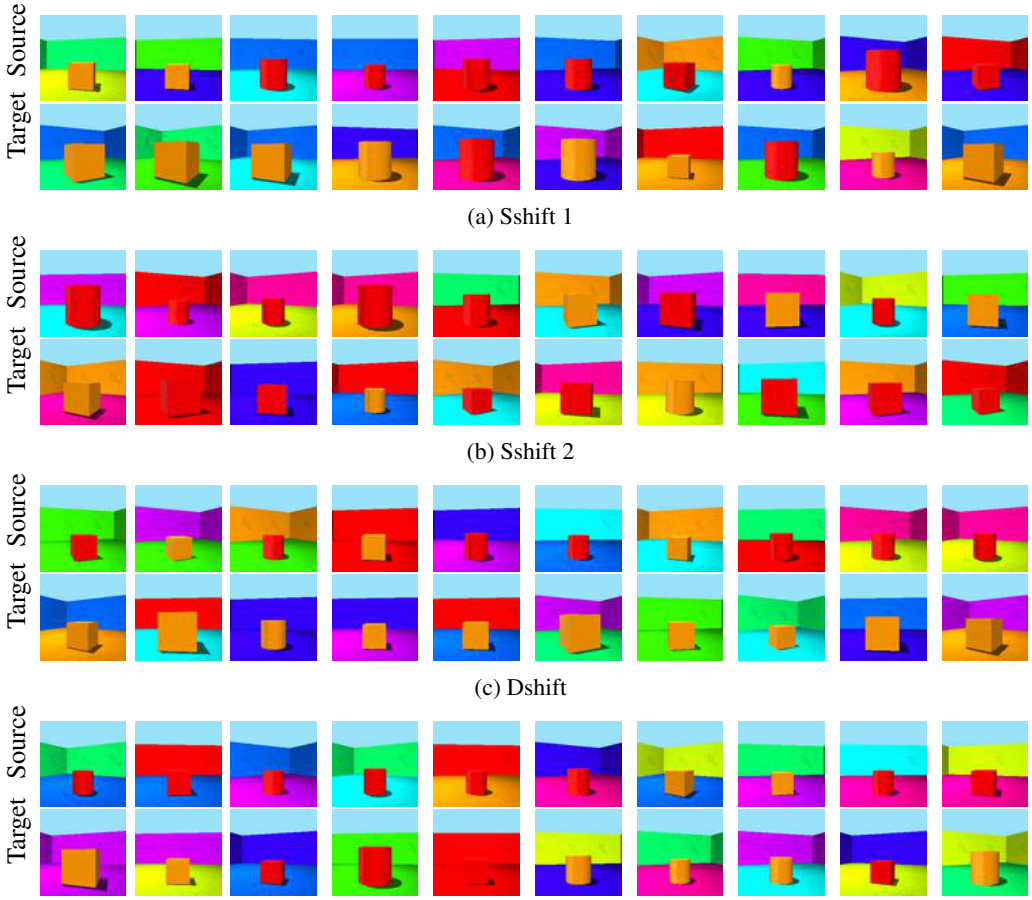
How to simulate different types of distribution shift? By varying the marginal distribution of latent factors and then sample images according to the distribution of latent factors, we can simulate different distribution shifts. In this paper, we set the image as input X, and set class $Y =$ *shape*, sensitive attribute $A =$ *object hue*, and a nuisance factor that might shift to be $D =$ *scale*. We consider a binary case, and restrict the *shape* to be in $\{0, 1\}$ (i.e. $\{\text{cube}, \text{cylinder}\}$) and *object hue* to be in $\{0, 1\}$ (i.e. $\{\text{red}, \text{yellow}\}$). In our experiments, we simulate four types of distribution shift. Their specific settings are shown in Table 5. We show examples from two domains under different shifts in Figure 7.

D.2 Real Datasets

UTKFace[3] [70] is a face dataset with images annotated with age, gender, and race. The data is collected from MORPH, CACD and Web. In our experiments, we use the aligned and cropped face images with ages larger than 10. We do gender classification which is a binary classification task, and set sensitive attribute to be the race. We consider binary-sensitive attribute case in our experiment by setting race to be white or non-white. The statistics of this dataset are shown in Table 6.

[2] https://github.com/deepmind/3d-shapes

[3] https://susanqq.github.io/UTKFace/



(a) Sshift 1

(b) Sshift 2

(c) Dshift

(d) Hshift

Figure 7: Randomly sampled examples from two domains under different distribution shifts.

	Factor	Source	Target
Sshift 1	$\mathbb{P}(Y, A)$	$[0.1, 0.4, 0.4, 0.1]$	same
	$\mathbb{P}(D)$	$[\frac{4}{16}, \frac{4}{16}, \frac{3}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}]$	$[\frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{3}{16}, \frac{4}{16}, \frac{4}{16}]$
Sshift 2	$\mathbb{P}(Y, A)$	$[0.1, 0.4, 0.4, 0.1]$	$[0.4, 0.1, 0.1, 0.4]$
	$\mathbb{P}(D)$	$[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}]$	same
Dshift	$\mathbb{P}(Y, A)$	$[0.1, 0.4, 0.4, 0.1]$	same
	$\mathbb{P}(D)$	$[\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0, 0]$	$[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}]$
Hshift	$\mathbb{P}(Y, A)$	$[0.1, 0.4, 0.4, 0.1]$	$[0.4, 0.1, 0.1, 0.4]$
	$\mathbb{P}(D)$	$[\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0, 0]$	$[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}]$

Table 5: Simulate different distribution shifts. $\mathbb{P}(Y, A)$ is represented by the proportions of four groups as $[\mathbb{P}(Y=0, A=0), \mathbb{P}(Y=0, A=1), \mathbb{P}(Y=1, A=0), \mathbb{P}(Y=1, A=1)]$. $\mathbb{P}(D)$ is represented by the proportions of eight possible values of *scale*. Other factors have uniform marginal distributions. Images in two domains are sampled according to the marginal distributions of six latent factors.

FairFace[4] [29] is another large-scale face dataset with images annotated with age, gender, and race as well. Different from UTKFace, the data in FairFace is collected from Flickr, Twitter, Newspapers

[4]https://github.com/joojs/fairface

(Y, A)		(Male, White)	(Male, Black)	(Female, White)	(Female, Black)
UTK (S)	train	3127	1508	2480	1450
	test	1377	651	1027	617
FairFace (T)	train	7796	4650	6946	5160
	test	984	620	839	635

Table 6: Statistics of UTK and FairFace datasets.

and Web. We also use images with ages larger than 10 in our experiments. We set the label to be the gender and the sensitive attribute to be the race. See statistics in Table 6. All face images in UTK and FairFace are resized to $96 \times 96 \times 3$ in our experiments.

NewAdult[5] [18] is a suite of datasets derived from US Census surveys. The data spans multiple years and all states of the United States which is a good fit for studying distribution shifts. In our experiments, we use 2018 data that span all states, and do income classification with a threshold of 50,000 dollars. We set the gender to be sensitive attribute. We consider a problem that we train a fair classifier in California (source domain) and deploy it in other states (target domain). The statistics are shown in Table 7. The input contains 10 features (see Appendix B.1 in [18]) which are preprocessed to one-hot embeddings in our experiments.

(Y, A)		(High, Male)	(High, Female)	(Low, Male)	(Low, Female)
CA (S)	train	33258	22314	39224	42169
	test	14839	9924	15990	17947
Other states (T)	train	232162	140876	296826	351970
	test	101934	57798	127654	150544

Table 7: Statistics of NewAdult dataset.

D.3 Experimental Settings

D.3.1 Experiments on 3dshapes

Model. We use a two-layer MLP with 512 hidden units as the encoder and one linear layer as the classifier. The adversaries used in Laftr, CFair and DANN are also two-layer MLP with 512 hidden units. ReLU is used as the activation function.

Transformations. We use random center cropping and padding as the transformation functions in consistency regularization. Such transformations can perfectly change the *scale* of the objects, so that to propogate labels from the source domain to the target domain.

Setup. We use SGD as the optimizer, and set the learning rate to be 0.01. We train every model with 200 epochs and select the best model according to the model's performance on the validation set. For Base and Laftr that only have access to the source data, the model selection is based on the source validation set, and for other methods that can access to the unlabeld target data, the model selection is based on the labeled target validation set. Since *accuracy* and *fairness* are both important metrics, we select the best model according to the value of accuracy minus unfairness which is $Acc - \Delta_{odds}$. The coefficients of the fairness loss and consistency loss are both set to be 1. We run every method for five times and report the mean and the standard deviation.

[5] https://github.com/zykls/folktables

D.3.2 Experiments on UTK-FairFace

Model. We use VGG16 and ResNet18 as the model in our experiments. The last linear layer is the classifier and all the previous layers construct the encoder. When we use VGG16 as the model, we set every adversary used in Laftr, CFair and DANN to be a two-layer MLP with 1024 hidden neurons. When ResNet18 is the model, the adversary has 512 hidden neurons.

Transformations. We use RandAugment [17] as the transformation function which contains data augmentations that are the best for CIFAR-10 dataset. To restrict the transformations to be group-preserving, we exclude augmentations that may change the color (so to change the race). The transformation function used in our experiments contains AutoContrast, Brightness, Equalize, Identity, Posterize, Rotate, Sharpness, ShearX, ShearY, TranslateX, TranslateY. In this experiment, we use a weak augmented (with random cropping and flipping) image as the original input $\boldsymbol{x}$, and a strong augmented (with RandAugment) image as the transformed input $t(\boldsymbol{x})$.

Setup. We use SGD as the optimizer. The learning rate for Base, Laftr, and CFair is 0.001 and is 0.001 for other methods. We train every model with 200 epochs and use the way introduced in the 3dshapes experiment to select the best model. The coefficients of the fairness loss and consistency loss are both set to be 1. We run every method for five times and report the mean and the standard deviation.

D.3.3 Experiments on NewAdult

Model. We use a 3-layer MLP with hidden sizes of (256, 512, 256) as the encoder, and a 2-layer MLP with the hidden size of 128 as the classifier. Every adversary is a two-layer MLP with 128 hidden neurons.

Transformations. Studies on data augmentations for tabular data is very limited. In this paper, we use random corruptions on the input features as the transformation function. There are ten features in the input, and every time we only corrupt half of them. Additionally, for important factors that highly correlated with the label, including the OCCP (occupation), COW (class of worker), we do not do any corruptions. For factor SEX (gender), we also do not do any corruptions to preserve the group. For continues factors including AGEP (age), SCHL (educational attainment) and WKHP (work hours), we do perturbations within a range. For other factors, we do uniformly sampling from their value spaces as corruptions. We do such transformations based on our assumption that they do not change the label. For example, two individuals that have a 5 years of age gap but have the same other features should have similar income, and two individuals that only differ in the place of birth should have similar income. We admit that such transformations may not be the best ones. We need better domain knowledge on the income prediction to design more powerful transformations. We leave the improvement of transformations for tabular data to future work.

Setup. We use SGD as the optimizer. The learning rate is set to be 0.01. We train every model with 200 epochs and use the metric introduced in the 3dshapes experiment to select the best model. The coefficients of the fairness loss and consistency loss are both set to be 1. We run every method for five times and report the mean and the standard deviation.

D.4 Baselines

Laftr is an adversarial learning method for algorithmic fairness. The adversary aims at accurately predicting the sensitive attribute based on the representation, while the encoder aims at making it hard. By adversarial learning, the representation will not contain information of sensitive attribute, so that the prediction based on it will be fair. We denote the data in each group to be $\mathcal{D}_a^y = \{\boldsymbol{x} \in \mathcal{D} | A = a, Y = y\}$. The fairness loss is designed to be

$$L_{fair} = \sum_{(a,y)\in\{0,1\}^2} \frac{1}{|\mathcal{D}_a^y|} \sum_{\boldsymbol{x}\in\mathcal{D}_a^y} |h(f(\boldsymbol{x})) - a|.$$

where f is the encoder, h is the adversary. [41] prove that this loss is an upper bound of the equalized odds. The adversary minimizes this loss while the encoder maximizes this loss with a gradient reversal layer.

CFair is similar to Laftr but uses two adversaries h' and h'' for two classes with a balanced error rate (BER) defined as follows. We denote the data from one class to be $\mathcal{D}^y = \{\boldsymbol{x} \in \mathcal{D} | Y = y\}$.

$$L_{fair} = \text{BER}_{\mathcal{D}^0}(h'(f(\boldsymbol{x})) \parallel A) + \text{BER}_{\mathcal{D}^1}(h''(f(\boldsymbol{x})) \parallel A)$$

where $\text{BER}_{\mathcal{D}^0}(h'(f(\boldsymbol{x})) \parallel A) = \frac{1}{2}\mathbb{P}_{\mathcal{D}^0}(h'(f(\boldsymbol{x})) \neq A | A = 0) + \frac{1}{2}\mathbb{P}_{\mathcal{D}^0}(h'(f(\boldsymbol{x})) \neq A | A = 1)$. In practice, the balanced error rate is estimated by the following cost-sensitive cross-entropy loss.

$$\mathbb{P}_{\mathcal{D}^0}(h'(f(\boldsymbol{x})) \neq A | A = 0) \leq \frac{\text{CE}_{\mathcal{D}_0^0}(h'(f(\boldsymbol{x})) \parallel A)}{\mathbb{P}_{\mathcal{D}^0}(A = 0)}$$

Laftr+FixMatch uses the same framework as our method but with a standard consistency regularization that does not care about group performance. The consistency loss is defined as

$$L_{consis}(g) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \mathbb{1}(\max(g_{tc}(\boldsymbol{x})) \geq \tau) H(\text{argmax}(g_{tc}(\boldsymbol{x})), g(t(\boldsymbol{x})))$$

where $\mathcal{D}$ denotes the entire dataset.

D.5 Time and Space Complexity

Comparing with Base, Laftr and CFair that only uses labeled source data, our method need more training time and memory since we use unlabeled target data as well. Comparing with other baselines that also use target data, such as Laftr+DANN, the time complexity of our method is comparable to theirs. Nevertheless, our method needs much less parameters than Laftr+DANN since it requires an adversary to do domain classification while we do not need it. Actually, our method has the same model parameters as that in Laftr but with an additional consistency loss.

E More Experimental Results

E.1 Additional Results on UTKFace-FairFace with a Different Data Setting

To evaluate our method in extreme circumstances, we conduct the UTK-FairFace experiment with less labeled source data and more unlabeled target data (see Table 8). We also consider the race to be "white" and "non-white". Are shown in Table 9, we get consistent results that our method outperforms all baselines, and can effectively transfer accuracy as well as fairness.

(Y, A)		(Male, White)	(Male, Non-white)	(Female, White)	(Female, Non-white)
UTK (S)	train	1373	750	1650	1227
	test	565	285	614	370
FairFace (T)	train	11429	16574	8024	16838
	test	1712	2453	1176	2518

Table 8: Statistics of UTK and FairFace datasets used in Table 9.

E.2 Additional Results on UTKFace-FairFace with Different Transformations

To investigate the effect of using different transformations in our method, we evaluate 14 transformations in RandAugment and report the results in Table 9. All the transformations can improve the accuracy in both domains. The effect on fairness varies. We find that *Solarize*, *Color* and *TranslateX* increase the unfairness in the source domain the most, and *Contrast*, *Color* and *Solarize* have the highest unfairness in the target domain. Note that, it does not mean that these augmentations always lead to unfairness but that they are not suitable for our method. Recall that our theory and algorithm are built upon the intra-group expansion assumption. Transformations like *Contrast*, *Color* and *Solarize* may change the sensitive attribute "race" and break this assumption. Thus, in our experiments (Table 1) we use all the transformations excluding *Contrast*, *Color* and *Solarize*.

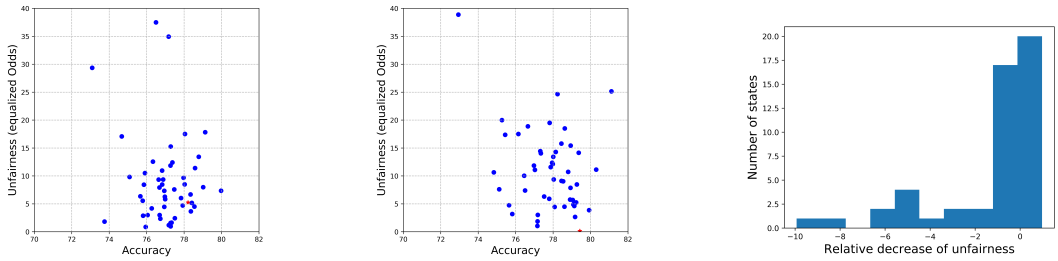
Method	Source			Target		
	Acc	Unfairness		Acc	Unfairness	
		V_{acc}	Δ_{odds}		V_{acc}	Δ_{odds}
Base	89.93±0.43	2.79±0.74	4.65±0.44	73.48±0.56	7.49±3.50	6.09±1.07
Laftr	90.61±0.33	1.28±0.43	3.62±1.17	73.29±0.70	5.42±1.33	7.78±1.77
CFair	90.68±0.35	1.20±0.59	3.61±0.93	73.82±0.81	5.71±1.54	7.37±1.40
Laftr+DANN	90.53±0.98	1.59±0.97	4.62±1.24	74.44±1.38	6.94±1.53	10.26±1.85
CFair+DANN	90.23±0.88	1.82±0.97	4.96±1.15	74.53±1.46	9.27±2.16	9.96±1.49
Laftr+FixMatch	95.01±0.10	1.37±0.44	4.65±1.00	83.77±0.45	11.58±1.16	6.56±1.74
CFair+FixMatch	95.37±0.24	1.13±0.21	3.58±0.90	83.62±0.51	11.96±1.05	5.29±1.76
Ours (w/ Laftr)	94.77±0.33	1.35±0.70	3.28±0.79	84.65±1.13	2.92±0.72	6.99±0.41
Ours (w/ CFair)	94.92±0.43	1.09±0.30	3.00±1.09	84.71±1.10	3.57±0.60	7.34±0.91

Table 9: Transfer fairness and accuracy from UTKFace to FairFace with less source data.

Transformation	Source			Target		
	Acc	Unfairness		Acc	Unfairness	
		V_{acc}	Δ_{odds}		V_{acc}	Δ_{odds}
None	90.61	1.28	3.62	73.29	5.42	7.78
AutoContrast	94.49	0.93	3.36	80.43	4.01	5.77
Brightness	94.6	1.25	4.34	81.45	1.93	4.9
Color	93.84	4.04	5.6	80.05	2.2	7.21
Contrast	94.55	1.3	4.57	81.7	4.86	9.71
Equalize	93.62	1.57	5.87	80.28	1.77	6.33
Identity	94.2	1.37	4.39	81.03	1.26	4.65
Posterize	94.66	1.39	3.7	81.91	3.65	4.97
Rotate	93.78	1.24	3.76	81.8	1.85	6.76
Sharpness	94.44	0.81	2.32	80.54	1.87	4.98
ShearX	94.66	1.23	3.61	80.57	3.1	2.17
ShearY	94.38	1.04	3.72	82.28	1.43	5.22
Solarize	93.89	1.64	6.32	81.7	2.13	7.83
TranslateX	94.11	2.01	5.48	81.56	2.85	5.34
TranslateY	94.87	1.77	2.96	82.71	3.51	4.29

Table 10: Results by using different transformations in our method. Average results of three trials.

E.3 Additional Results on NewAdult with Different Transformations



(a) Laftr (b) Laftr+Mixup (c) Decrease of unfairness

Figure 9: A failure case of using Mixup as the transformation function.

Pre-defined transformation functions require good domain knowledge on the task which is hard to obtain for the NewAdult dataset. To alleviate this issue, we try to use Mixup[6] as the transformation function on the tabular data. Mixup does extrapolation between training examples as the data augmentation which is applicable to any type of data without requiring any prior knowledge. We apply Mixup on NewAdult dataset. Since Mixup requires the access to the label, we only use labeled source data to train the model in this experiment. We use CA as the source domain, and evaluate on all the other states. We set race (white and non-white) to be the sensitive attribute. Figure 9 shows that Mixup improves the in-distribution accuracy and fairness but harms the out-of-distribution fairness with neglectable improvement in accuracy. This preliminary experiment suggests that Mixup that only applies to the source data can only benefit the accuracy and fairness in the source domain.

[6]https://github.com/facebookresearch/mixup-cifar10

We suspect that it is because the augmented data generated by Mixup is hard to expand to other domains, and they are not group-preserving. One interesting future work is to do intra-group Mixup and apply it to the unlabeled target data as well by using confident pseudolabels. All the experiments on NewAdult dataset encourage the community to design more powerful transformations for tabular data, so that to transfer accuracy. After that, the proposed fair consistency regularization is able to help making this transferring process fair.

E.4 A Byproduct: Alleviate the Disparate Impact of Semi-supervised Learning

[75] find that semi-supervised learning methods may have disparate impact. The classes that have high accuracy on labeled data is likely to benefit more from semi-supervised learning on unlabeled data. We test this argument on CIFAR-10 with FixMatch as the semi-supervised learning method. We use ResNet18 as the model. We randomly sample 500 images to be the labeled data and treat others as unlabeled data. We use the benefit ration proposed in [75] as the metric for the benefit of semi-supervised learning, defined as

$$BR(\mathcal{D}) = \frac{a_{semi}(\mathcal{D}) - a_{baseline}(\mathcal{D})}{a_{ideal}(\mathcal{D}) - a_{baseline}(\mathcal{D})}. \tag{8}$$

where $\mathcal{D}$ denotes the data from one class. $a_{semi}(\mathcal{D})$ is the model's test accuracy after semi-supervised learning, $a_{baseline}(\mathcal{D})$ is the test accuracy of the base model that is trained on labeled data, and $a_{ideal}(\mathcal{D})$ is the test accuracy of the ideal model where all data are labeled. We evaluate the benefit ratio of FixMatch on ten classes. As the blue line in Figure 10 shows, the rich gets richer and the poor gets poorer after semi-supervised learning.

Our method (without using Laftr) can directly apply to this task. By using the proposed fair consistency regularization (red line in Figure 10), we can significantly improve the benefit ratio of the poor classes. Therefore, the fair consistency regularization is a byproduct of this paper which is able to alleviate the disparate impact of semi-supervised learning.

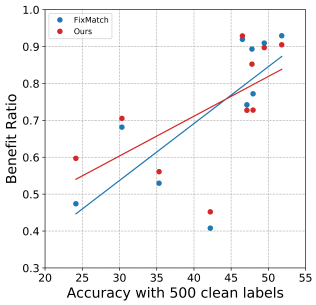


Figure 10: With fair consistency regularization, our method alleviates the disparate impact of FixMatch.

F Impact and Limitations

Fairness of machine learning is a critical problem we are facing in todays real-world applications. When distribution shifts happen, the collapse of fair systems would cause unexpected discrimination, resulting in severe negative social impact. The fairness that robust to distribution shifts is essential but is less explored. In this paper, the theoretical analysis of how the fairness changes under different distribution shifts sheds light on the deep reasons of the collapse of fairness. The theory-guided self-training algorithm proposed in this paper explores a promising way to tackle distribution shifts. We hope our work would inspire more algorithms for this important and practical task.

The major limitation of our method is that it strongly relies on the pre-defined transformations as all the other self-training methods. The transformations are designed to be group-preserving based on our prior knowledge. Our experiments show that self-training with less powerful transformations have limited ability in propagating labels from source to target (i.e. transfer accuracy). Valid transformation functions on image data are thoroughly studied in exiting work, while transformations on non-image

data such as tabular data are much less explored. Our method with more powerful transformations on tabular data is expected to have significant improvement. Future work is encouraged to relax this limitation, such as by using a generative model as the transformation function.