# TRANSFER FAIRNESS UNDER DISTRIBUTION SHIFT

**Bang An**[†], **Zora Che**[§], **Mucong Ding**[†], **Furong Huang**[†]
[†]University of Maryland, [§]Boston University

## ABSTRACT

As machine learning systems are increasingly employed in high-stakes tasks, algorithmic fairness has become an essential requirement for deep learning models. In this paper, we study how to transfer fairness under distribution shifts, a crucial issue in real-world applications. We first 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. Experiments on synthetic and real datasets demonstrate that our approach can effectively transfer fairness as well as accuracy under distribution shifts, especially under domain shift which is a more challenging but practical scenario.

## 1 INTRODUCTION

The social impact of machine learning has increased as it is widely used to aid our decision-making in real-world applications, such as hiring, loan approval, facial recognition and criminal justice. To avoid the discrimination on a subset of population (e.g. with respect to race, gender), many efforts on algorithmic fairness have been carried out (Chouldechova, 2016; Friedler et al., 2016; Zafar et al., 2017; Mehrabi et al., 2019; Rajkomar et al., 2018; Corbett-Davies & Goel, 2018; Caton & Haas, 2020). Although existing works have achieved remarkable success in ensuring fairness, most of them assume the distribution of data is identical to that in training stage, which hinders their use in practice since distribution shifts always happen in reality. Recent studies show that the fairness of a model is likely to collapse when encountering a distribution shift. For example, Ding et al. (2021) observe that a fair income predictor trained with data from one state might not be fair when used in other states. Schrouff et al. (2022) try to maintain fairness in healthcare settings, but a model that performs fairly according to the metric in "hospital A" shows unfairness when deployed in "hospital B". Such observations motivate us to investigate how to transfer fairness under distribution shifts. Specifically, when there is a fair model in a source domain, we investigate how to adapt it 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. We follow the taxonomy in Koh et al. (2021) which categorize distribution shifts into two types: *domain shift* where source and target distributions comprise data from related but distinct domains (e.g. deploy a system to a new environment), and *subpopulation shift* where two domains overlap, but relative proportions of subpopulations differ (e.g. the increase of female candidates). We develop a synthetic dataset benchmark to simulate possible distribution shifts and find that domain shift is more challenging than subpopulation shift when transferring fairness. While recent work explores many methods to transfer fairness (Singh et al., 2021; Rezaei et al., 2020; Giguere et al., 2022), the most considered settings fall into subpopulation shifts. This encourages us to focus on domain shifts and hybrid shifts which are more difficult but practical settings.

Recent progress of self-training (Wei et al., 2021; Cai et al., 2021; Zhang et al., 2021b; Berthelot et al., 2021; Sagawa et al., 2021; Sohn et al., 2020) shows that input consistency regularization enables label propagation and has become a powerful approach for transferring accuracy. By taking demography into consideration, we first extend self-training assumption to intra-group expansion (as in Figure 1) and then derive a sufficient condition for transferring fairness under such assumption. The key insight is to have a fair teacher classifier and ensure the model gains the same input consistency on different groups in order to avoid biased label propagation. Guided by our theoretical analysis, we propose a practical algorithm which combines LAFTR (Madras et al., 2018), an adversarial learning method for fairness, and FixMatch (Sohn et al., 2020), a self-training framework. To balance the consistency, we also improves FixMatch with a novel fair consistency regularization.

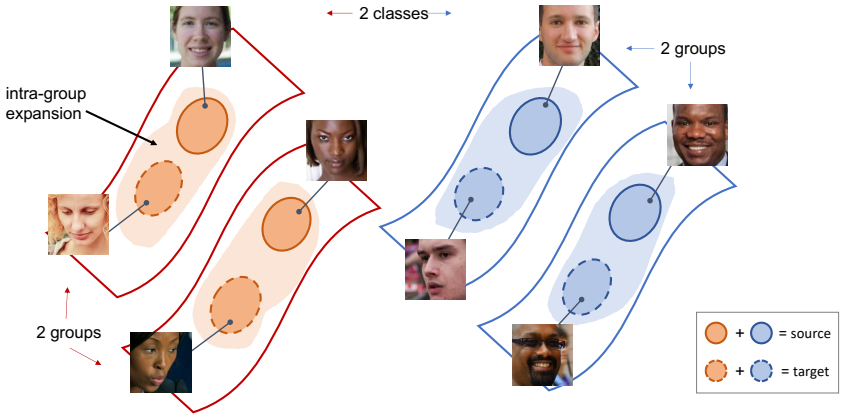



Figure 1: Illustration of intra-group expansion assumption in the input space. Here, we consider a gender classification task where the sensitive attribute is the race. Source images are sampled from UTKFace dataset (Zhang et al., 2017) and target images are sampled from FairFace dataset (Kärkkäinen & Joo, 2019). Images in two domains carry different capture-bias (e.g. different angles, diversity of facial expressions) but lie on the same population distribution. Intra-group expansion assumes that different groups in the same class are separated by the sensitive attribute while every group is self-connected with certain transformations. Under such assumptions, we propose to obtain fairness and accuracy in both domains by training with a fair teacher classifier and a group-balanced input consistency.

We evaluate our method under different types of distribution shifts with a synthetic dataset and also test it on real datasets. Experiments show that our approach performs 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.

## 2 Transfer Fairness via Fair Consistency Regularization

### 2.1 Problem Setting

We consider a classification problem in this paper. Let $X, A, Y$ denote random variables and $\mathcal{X}, \mathcal{A}, \mathcal{Y}$ denote corresponding spaces of input features, sensitive attribute (e.g. male and female) and label. For simplicity, we assume binary sensitive attribute and binary classification, $\mathcal{A} = \{0, 1\}, \mathcal{Y} = \{0, 1\}$. We aim to learn a model $g : \mathcal{X} \to \mathcal{Y}$ (or $g : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$) and are interested in the fairness of it when a distribution shift happens. Specifically, with $S$ and $T$ denoting source and target distributions, we consider the case that $\mathbb{P}_S(X, A, Y) \neq \mathbb{P}_T(X, A, Y)$ but two domains share the same ground truth classifier $g^*$. Suppose we have $m$ labeled examples $\{(\boldsymbol{x}_S^{(i)}, a_S^{(i)}, y_S^{(i)})\}_{i=1}^m$ in source domain and $m'$ unlabeled examples $\{(\boldsymbol{x}_T^{(i)}, a_T^{(i)})\}_{i=1}^{m'}$ in target domain. With labeled data, we can find a fair and accurate model in source domain denoted as $g_{tc}$. We study how to adapt this teacher classifier $g_{tc}$ to target domain with high accuracy and fairness without sacrificing the performance in source domain.

We use *equalized odds* Hardt et al. (2016) as the fairness metric in our theoretical analysis which requires the true positive rate (TPR) and the true negative rate (TNR) to be the same among groups. Thus, the unfairness is defined as

$$\Delta_{odds} = \frac{1}{2}|\mathbb{P}(\hat{Y} = 1|A = 0, Y = 1) - \mathbb{P}(\hat{Y} = 1|A = 1, Y = 1)| + \frac{1}{2}|\mathbb{P}(\hat{Y} = 0|A = 0, Y = 0) - \mathbb{P}(\hat{Y} = 0|A = 1, Y = 0)|.$$

where $\hat{Y} = g(X)$ is the prediction by our model. Since this unfairness metric separates the data into groups based on values for both $A$ and $Y$, we will use "group" to indicate such separation instead of conventional understanding that only considers $A$. In next section, we aim to bound the unfairness and error in both domains.

### 2.2 THEORETICAL ANALYSIS

Under the assumption that two domains share one underlying population distribution which has good continuity within each class or sub-class, Wei et al. (2021) and Cai et al. (2021) prove that self-training with input consistency regularization can propagate labels from source domain to target domain so that to transfer accuracy. Taking the demography and fairness into consideration, we extend the expansion assumption to intra-group expansion (as shown in Figure 1) which is more realistic. Under this assumption, we then upper bound the unfairness and error for a self-training method.

First of all, let's characterize the data distribution as follows. We assume that the source data and target data are from the same underlying population distribution where classes are separated and different groups in the same class are separated as well (i.e. separate by sensitive attribute).

**Assumption 1.** *Let $S_a^i$ and $T_a^i$ denote the support set of $\{X|A=a,Y=i\}$ in source and target domains. We assume* $\text{supp}(S) = \cup_i \cup_a S_a^i$ *and* $\text{supp}(T) = \cup_i \cup_a T_a^i$*, where (1) $S_a^i \cap S_{a'}^i = T_a^i \cap T_{a'}^i = S_a^i \cap T_{a'}^i = \emptyset, \forall i, a \neq a'$, and (2) $S_a^i \cap S_{a'}^j = T_a^i \cap T_{a'}^j = S_a^i \cap T_{a'}^j = \emptyset, \forall a, a', i \neq j$.*

This is a realistic assumption as illustrated in Figure 1. In the following analysis, we abuse the notation to let $S_a^i$, $T_a^i$ also denote the conditional distribution (probability measure) of $S, T$ on the set $S_a^i, T_a^i$ respectively and define $U_a^i = \frac{1}{2}(S_a^i + T_a^i)$ as the group population distribution. We use $U$ to denote the population distribution of the entire data. Next, we define the neighborhood and intra-group expansion based on it.

**Definition 2.1.** *Let $\mathcal{T}$ denote a set of input transformations (e.g. via data augmentations) 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 $\boldsymbol{x} \in S_a^i \cup T_a^i$, we define the neighbor of $\boldsymbol{x}$ as $\mathcal{N}(\boldsymbol{x}) := (S_a^i \cup T_a^i) \cap \{\boldsymbol{x}' | \mathcal{B}(\boldsymbol{x}) \cap \mathcal{B}(\boldsymbol{x}') \neq \emptyset\}$ and define the neighbor of a set $V \in \mathcal{X}$ as $\mathcal{N}(V) := \cup_{\boldsymbol{x} \in V \cap (\cup_i \cup_a S_a^i \cup T_a^i)} \mathcal{N}(x)$.*

**Assumption 2** (Intra-group expansion)**.** *We say that $U_a^i$ satisfies $(\alpha, c)$-multiplicative expansion for some constant $\alpha \in (0,1)$ and $c > 1$, if for all $V \subset U_a^i$ with $\mathbb{P}_{U_a^i}(V) \leq \alpha$, we have $\mathbb{P}_{U_a^i}(\mathcal{N}(A)) \geq \min\{c\mathbb{P}_{U_a^i}(A), 1\}$.*

Because of the Assumption 1 on data distribution, we define the neighbor and expansion assumption within the group. Definition 2.1 can be understood as two examples are neighbors if they are near each other after applying some transformations and the neighbor of a set is the union of neighbors of its elements. As shown in Figure 1, intra-group expansion states that for every sufficiently small set of points in a group, the group conditional probability mass of its neighbor is sufficiently large. We can also interpret it as the manifold of this group has sufficient connectivity. This assumption allows us to propagate labels within the group from one domain to another by encouraging the consistency under transformations. In the following, we investigate how to use this nice property to transfer fairness and accuracy with self-training.

Let's use $R_{U_a^i}(g) \triangleq \mathbb{P}_{U_a^i}[\exists \boldsymbol{x}' \in \mathcal{B}(\boldsymbol{x}), \text{s.t. } g(\boldsymbol{x}) \neq g(\boldsymbol{x}')]$ to denote the consistency loss of classifier $g$ on distribution $U_a^i$ which is the fraction of examples where $g$ is not robust to input transformations. We use 0-1 loss to evaluate the error of $g$ as $\varepsilon_{U_a^i}(g) \triangleq \mathbb{P}_{U_a^i}[g(\boldsymbol{x}) \neq g^*(\boldsymbol{x}')]$, and the disagreement between $g$ and a teacher classifier $g_{tc}$ as $L_{U_a^i}(g, g_{tc}) \triangleq \mathbb{P}_{U_a^i}[g(\boldsymbol{x}) \neq g_{tc}(\boldsymbol{x}')]$. The following theorem states a sufficient condition for transferring fairness as well as accuracy (see proof inAppendix B).

**Theorem 2.1.** *If we have a teacher classifier $g_{tc}$ with bounded unfairness such that $|\varepsilon_{U_a^i}(g_{tc}) - \varepsilon_{U_{a'}^{i'}}(g_{tc})| \leq \gamma, \forall a, a' \in \mathcal{A}$ and $i, i' \in \mathcal{Y}$. We assume that $U_a^i$ satisfies $(\bar{\alpha}, \bar{c})$-multiplicative expansion and $\varepsilon_{U_a^i}(g_{tc}) \leq \bar{\alpha} < 1/3$ and $\bar{c} > 3, \forall a, i$. We define $c \triangleq \min\{1/\bar{\alpha}, \bar{c}\}$. Set $\mu \leq \varepsilon_{U_a^i}(g_{tc}), \forall a, i$. If we train our classifier using the following algorithm*

$$\min_{g \in G} \max_{a,i} R_{U_a^i}(g) \tag{1}$$

$$\text{s.t.} \quad L_{U_a^i}(g, g_{tc}) \leq \mu \quad \forall a, i \tag{2}$$

*and denote the optimal solution as $\hat{g}$. Then the error and unfairness of $\hat{g}$ on the population distribution $U$ are bounded as*

$$\varepsilon(\hat{g}) \leq \frac{2}{c-1}\varepsilon(g_{tc}) + \frac{2c}{c-1}R_U(\hat{g}) \tag{3}$$

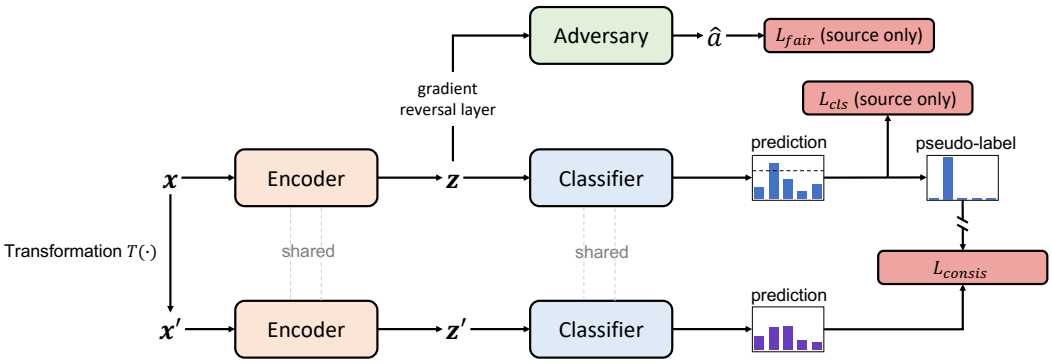



Figure 2: Training diagram.

$$\Delta_{odds} \leq \frac{2}{c-1}(\gamma + \mu + c \max_{a,i} R_{U_a^i}(\hat{g})) \tag{4}$$

Intuitively, this sufficient condition suggests us to fit a fair teacher classifier and minimize a balanced consistency among groups in order to guarantee a low unfairness and error in both domains. One challenge of applying this algorithm in practice is the fair teacher classifier. We can get a fair teacher classifier in source domain with labeled data, but it might not be fair on population distribution. Interestingly, the iterative self-training paradigm is able to update the teacher classifier and thus making it fairer and fairer. Another challenge is how to balance the consistency. Existing self-training approaches do not take fairness into consideration and may have a biased consistency. In the next section, we propose a fair consistency regularization to tackle it.

### 2.3 Practical Algorithm

Guided by theorem 2.1, we propose a practical algorithm (as shown in Figure 2) by combining LAFTR (Madras et al., 2018), an adversarial learning method for fairness, and FixMatch (Sohn et al., 2020), a self-training framework. LAFTR is one of the famous in-processing fairness methods, it enforces the accuracy and fairness with a supervised classification loss $L_{cls}^S$ and an adversarial loss $L_{fair}^S$ that bounds the equalized odds. With labeled data, we can train a fair model in source domain using LAFTR. To transfer fairness and accuracy, we combine it with FixMatch Sohn et al. (2020), a self-training framework. We train a student model to fit the teacher classifier and, at the same time, to minimize the consistency loss $L_{consis}$ with unlabeled data from two domains. Since we need a teacher classifier that is fair on population distribution instead of only source distribution, we iteratively update the teacher classifier by using the student model from the previous epoch as the new teacher classifier. Overall, the loss function in our algorithm is the weighted summation of three terms: $L = L_{cls}^S + \alpha L_{fair}^S + \beta L_{consis}$.

Guided by the objective in theorem 2.1, we need to minimize and balance the consistency loss evaluated on every group. However, FixMatch (Sohn et al., 2020) does not take fairness into consideration which is shown to have bias in our experiments. Therefore, we propose a fair consistency regularization with the following loss function.

$$L_{consis}(g) = \sum_i \sum_a \lambda_a^i L_a^i(g) \tag{5}$$

$$\text{where} \quad L_a^i(g) = \frac{1}{\sum_{\boldsymbol{x}_a^i} 1} \sum_{\boldsymbol{x}_a^i} \mathbb{1}(\max(g_{tc}(\boldsymbol{x}_a^i)) \geq \tau) H(\operatorname{argmax}(g_{tc}(\boldsymbol{x}_a^i)), g(T(\boldsymbol{x}_a^i))) \tag{6}$$

where $\boldsymbol{x}_a^i$ denotes an input with sensitive attribute $A = a$ and class $Y = i$, $\tau$ is a fixed confidence threshold, $H$ denotes the cross entropy function and $\lambda_a^i$ is the coefficient of the consistency loss of corresponding group. 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. $T(\boldsymbol{x}_a^i)$ is the transformed input as defined in definition 2.1. Minimizing this loss function will lead to a model consistent to certain transformations which is similar to FixMatch. What differs is that

we also need to balance the consistency loss, so we evaluate it per group. The challenge is that we do not know which group one example belongs to since we do not access to labels in target domain. To tackle it, we propose to use pseudolabels for target data. The error is acceptable since only high confident examples are considered by the consistency loss. Then, the simplest way to achieve balanced consistency loss is to set $\lambda_a^i = 1/\#\text{groups}$. However, this setting overlooks the approximation error of consistency loss. By theorem 2.1, our goal is to minimize and balance the consistency loss on population distribution and we approximate it with function 6. Intuitively, if a group has less confident examples, the approximation error of consistency loss is larger. To penalize such cases, we should increase the coefficient of its empirical consistency loss. Therefore, we propose to set $\lambda_a^i$ as

$$\hat{\lambda}_a^i = \frac{1}{\sum_{\boldsymbol{x}_a^i} \mathbb{1}(\max(g_{tc}(\boldsymbol{x}_a^i)) \geq \tau)} \tag{7}$$

and $\lambda_a^i = \hat{\lambda}_a^i / \sum \hat{\lambda}_a^i$. Note that since the teacher classifier evolves while training, this is a dynamic coefficient. With such consistency regularization, the model will dynamically pay more attention to groups it is not confident about which enables us to minimize and balance the consistency at the same time. To allow mini-batch training, we use the number of pseudolabels from the last epoch to calculate the weights of consistency loss with Equitation 7 in every epoch. Although we use LAFTR and FixMatch in this paper, it is also possible to use other in-processing fairness methods such as CFair (Zhao et al., 2019a) and consistency training methods such as UDA (Xie et al., 2019) with the balanced loss. We leave this to future work.

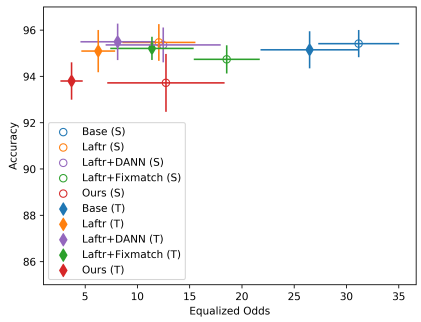
## 3 EXPERIMENTS

We first evaluate our method and benchmark methods under different settings of distribution shift with a synthetic dataset and then conduct experiments on real datasets.

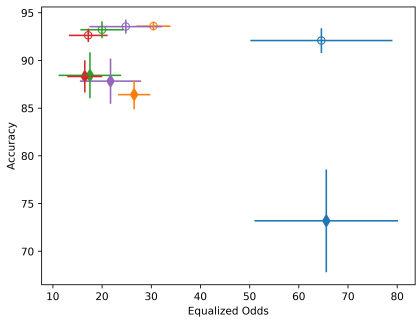
### 3.1 EVALUATE FAIRNESS UNDER DIFFERENT TYPES OF DISTRIBUTION SHIFT

To study the fairness under distribution shifts, we adjust 3D Shapes dataset (Kim & Mnih, 2018), which contains images of 3D objects generated from few independent latent factors, to a synthetic dataset. By sampling according to specific distributions of latent factors, we can simulate different types of distribution shifts (see Appendix C.1 for details). We treat the image as input $X$, the object color as sensitive attribute $A$, the object shape as class $Y$ and the scale of object as domain $D$. Our goal is to train a fair shape classifier under distribution shifts. We simulate four types of distribution shifts and compare accuracy and fairness in source and target domains. We use 2-layer MLP as the encoder and 2-layer MLP as the adversary. We compare our model with four baselines: (1) Base (standard ERM), (2) LAFTR, (3) LAFTR+DANN (a combination of LAFTR and a domain adaptation method (Ganin et al., 2016)), (4) LAFTR+FixMatch. We use random padding and cropping as transformations in consistency loss for (4) and our approach.

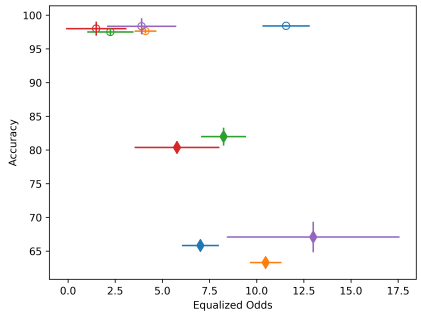
**Domain shift is more challenging than subpopulation shift.** Figure 3 shows that under subpopulation shifts, LAFTR can already achieve high accuracy and fairness in target domain although it is trained to be fair in source domain. We suspect that it is because source and target domains share the same support set, so the change of proportions of subpopulations (defined by scale in Figure 3(a), and by shape and color in Figure 3(b)) will not affect fairness much if the model is already accurate on every group. However, under domain shifts, the model trained to be accurate and fair in source domain (with small object scales) with LAFTR performs poorly in target domain (with large object scales) in both aspects of accuracy and fairness. This observation suggest that domain shift is a more challenging problem. By using unlabeled target data with domain adaptation methods, LAFTR+DANN performs poorly in transferring fairness or accuracy and LAFTR+FixMatch can transfer accuracy but might be unfair in target domain which is consistent with the finding in Zhu et al. (2022). Comparing with them, our method with fair consistency regularization achieves high accuracy and fairness in target domain and also improves the performance in source domain.



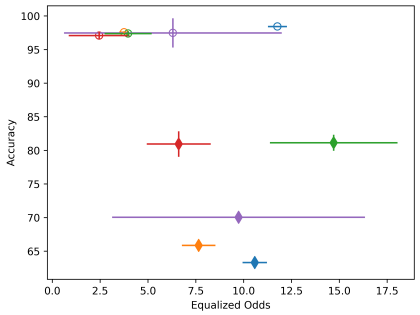

(a) Subpopulation shift, $\mathbb{P}_S(Y, A) = \mathbb{P}_T(Y, A), \mathbb{P}_S(D) \neq \mathbb{P}_T(D)$




(b) Subpopulation shift, $\mathbb{P}_S(Y, A) \neq \mathbb{P}_T(Y, A), \mathbb{P}_S(D) = \mathbb{P}_T(D)$




(c) Domain shift, $\mathbb{P}_S(Y, A) = \mathbb{P}_T(Y, A), \mathbb{P}_S(D) \neq \mathbb{P}_T(D)$




(d) Domain shift, $\mathbb{P}_S(Y, A) \neq \mathbb{P}_T(Y, A), \mathbb{P}_S(D) \neq \mathbb{P}_T(D)$

Figure 3: Accuracy and unfairness in two domains under different types of distribution shift. (S) denotes the source domain and (T) denotes the target domain, we run experiments under every setting for 5 times and visualize the standard deviation as the bar.

### 3.2 Evaluation on real dataset

In this section, we evaluate our method on two face datasets. We use UTKFace dataset (Zhang et al., 2017) as source domain and FairFace dataset (Kärkkäinen & Joo, 2019) as target domain. There is a distribution shift between two domains since the images in two datasets are collected from different sources. We consider a gender classification task where the sensitive attribute is race (white and non-white). We use VGG16 (Simonyan & Zisserman, 2014) as the model and compare with four baselines. We use horizontal flipping and random cropping as transformations in consistency loss. Besides equalized odds, we evaluate the unfairness with two additional metrics. The accuracy disparity $\Delta_{acc} = |\mathbb{P}(\hat{Y} = Y|A = 0) - \mathbb{P}(\hat{Y} = Y|A = 1)|$, and the variance of group accuracy $V_{acc} = Var(\{\mathbb{P}(\hat{Y} = i|A = a, Y = i), \forall a, i\})$. Results are shown in Table 1. Comparing with LAFTR and LAFTR+DANN, LAFTR+FixMatch effectively improves the accuracy in target domain while suffers from high unfairness. As expected, our method can reduce the unfairness via balanced consistency regularization.

Table 1: Transfer fairness and accuracy from UTKFace to FairFace with VGG16

| Method | Source | | | | Target | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc | $V_{acc}$ | $\Delta_{acc}$ | $\Delta_{odds}$ | Acc | $V_{acc}$ | $\Delta_{acc}$ | $\Delta_{odds}$ |
| Base | 90.79 | 1.73 | 2.53 | 4.88 | 74.22 | 7.01 | 4.57 | 6.62 |
| LAFTR | 91.00 | 1.46 | 1.68 | 3.45 | 73.74 | 4.1 | 4.73 | 8.10 |
| LAFTR+DANN | 90.57 | 1.63 | 2.43 | 4.53 | 73.37 | 6.46 | 4.44 | 9.91 |
| LAFTR+FixMatch | 92.31 | 0.98 | 1.58 | 3.20 | 76.36 | 18.71 | 5.79 | 5.22 |
| Ours | 92.80 | 1.51 | 2.58 | 5.21 | 77.07 | 4.71 | 2.42 | 6.36 |

## 4 CONCLUSION

In this paper, we study the problem of transferring fairness under distribution shift. We derive a sufficient condition for it based on a intra-group expansion assumption and propose a theory guided algorithm with self-training. To further ensure a fair consistency regularization, we propose a simple but effective reweighted consistency loss with dynamic weights calculated from pseudolables. We also introduce a synthetic dataset to study this problem with a easy realization of different types of distribution shift. One limitation of our approach and other self-training approaches is that the performance strongly depends on a well defined transformation set. In future work, we will explore the impact of transformations on our algorithm and apply our method to other real-world applications.

## REFERENCES

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 (eds.), *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*, pp. 60–69. PMLR, 2018. URL http://proceedings.mlr.press/v80/agarwal18a.html.

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. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

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

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

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

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18, 2009. doi: 10.1109/ICDMW.2009.83.

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

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, pp. 319–328, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287586. URL https://doi.org/10.1145/3287560.3287586.

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

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.

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

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. URL https://openreview.net/forum?id=DNl5s5BXeBn.

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.

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

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 (eds.), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pp. 91–98. ACM, 2019. doi: 10.1145/3306618.3314236. URL https://doi.org/10.1145/3306618.3314236.

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 (eds.), *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*, pp. 1436–1445. PMLR, 2019. URL http://proceedings.mlr.press/v97/creager19a.html.

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 (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=bYi_2708mKK.

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 (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2796–2806, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/83cdcec08fbf90370fcf53bdd56604ff-Abstract.html.

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

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.

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. URL https://openreview.net/forum?id=wbPObLm6ueA.

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 (eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pp. 269–278. ACM, 2019. doi: 10.1145/3287560.3287573. URL https://doi.org/10.1145/3287560.3287573.

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. URL https://openreview.net/forum?id=9YlaeLfuhJF.

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.

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

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 (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1989–1998. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hoffman18a.html.

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

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

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

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*, pp. 5637–5664. PMLR, 2021.

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

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 (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4066–4076, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

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

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.

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

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

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.00830.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In Jennifer G. Dy and Andreas Krause (eds.), *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*, pp. 3381–3390. PMLR, 2018. URL http://proceedings.mlr.press/v80/madras18a.html.

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.

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. URL http://arxiv.org/abs/1908.09635.

Luca Oneto and Silvia Chiappa. Fairness in machine learning. *Studies in Computational Intelligence*, pp. 155–196, 2020. ISSN 1860-9503. doi: 10.1007/978-3-030-43883-8_7. URL http://dx.doi.org/10.1007/978-3-030-43883-8_7.

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.

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.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. *CoRR*, abs/2010.05166, 2020. URL https://arxiv.org/abs/2010.05166.

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.

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.

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. URL http://arxiv.org/abs/1906.09688.

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

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

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 (eds.), *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pp. 3–13. ACM, 2021. doi: 10.1145/3442188.3445865. URL https://doi.org/10.1145/3442188.3445865.

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.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *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*, pp. 2164–2173. PMLR, 2019. URL http://proceedings.mlr.press/v89/song19a.html.

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.

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*, pp. 7167–7176, 2017.

Haotao Wang, Junyuan Hong, Jiayu Zhou, and Zhangyang Wang. Equalized robustness: Towards sustainable fairness under distributional shifts, 2022. URL https://openreview.net/forum?id=-dzXGe2FyW6.

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*, pp. 454–463, 2021.

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.

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

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

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. doi: 10.1109/ACCESS.2020.3005987. URL https://doi.org/10.1109/ACCESS.2020.3005987.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/zafar17a.html.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/zemel13.html.

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*, pp. 335–340, 2018.

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*, pp. 1–1, 2020. doi: 10.1109/TKDE.2020.3002567.

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*, pp. 245–256. Springer, 2021a.

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

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.

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

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

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. URL https://openreview.net/forum?id=Hkekl0NFPr.

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.

Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DXPftn5kjQK.

## A Related Work

**Fair machine learning** Generally, fair machine learning methods fall into three categories: pre-processing, in-processing and post-processing (Mehrabi et al., 2019; Caton & Haas, 2020). In this paper, we focus on in-processing methods which modify learning algorithms to remove discrimination during the training process. As for fair classification, seminal approaches have been proposed including fair representation learning (Zemel et al., 2013; Louizos et al., 2016; Beutel et al., 2017; Zhang et al., 2018; Madras et al., 2018; Song et al., 2019; Creager et al., 2019; Zhao et al., 2020), fairness-constrained optimization (Donini et al., 2018; Agarwal et al., 2018), causal methods (Kusner et al., 2017; Glymour & Herington, 2019; Oneto & Chiappa, 2020), and many other approaches with different techniques (Celis et al., 2019; Chuang & Mroueh, 2021; Goel et al., 2021). All of those work are for in-distribution fairness and we investigate out-of-distribution fairness in this paper. We use LAFTR (Madras et al., 2018), an adversarial learning method which shows advanced performance on fairness (Reddy et al., 2021), to learn a fair model in a source domain and adapt it to a target domain. Many metrics of fairness have been proposed (Corbett-Davies & Goel, 2018) including demographic parity (Calders et al., 2009), equalized opportunity, and equalized odds (Hardt et al., 2016) which are most widely adopted.

**Domain adaptation and self-training** Inspired by seminal theoretical work (Ben-David et al., 2010), numerous distribution matching approaches have emerged for domain adaption over the past decade. Domain-adversarial training (Ganin et al., 2016) and many of its variants (Tzeng et al., 2017; Long et al., 2018; Hoffman et al., 2018; Tachet des Combes et al., 2020) that aim at matching the distribution of two domains in the feature space have shown encouraging results in many applications. However, recent studies (Wu et al., 2019; Zhao et al., 2019b; Li et al., 2020) show that such methods may fail in many cases since they only optimize part of the theoretical bound. In our experiment, we also test DANN (Ganin et al., 2016) as a baseline. Recently, another line of work for domain adaptation arises which uses self-training similar to semi-supervised learning (Zhang et al., 2021b; Berthelot et al., 2021). Those methods enjoy guarantees (Wei et al., 2021; Cai et al., 2021), demonstrate superior empirical results and desirable properties such as robustness to spurious features (Kumar et al., 2020; Chen et al., 2020; Liu et al., 2021). However, all of those works on domain adaptation only aim at high target accuracy. Although there are works that study fairness issues in current domain adaptation methods (Lan & Huan, 2017) and propose to alleviate it by balancing the data (Jing et al., 2021; Wang et al., 2021; Zhu et al., 2022), fair domain adaptation is still under-explored. Based on the findings that model's consistency to input transformations is important to generalization (Zhu et al., 2021) and is a core component of self-training (Shu et al., 2018; Sohn et al., 2020; Grill et al., 2020), we improve the consistency regularization in Sohn et al. (2020) to achieve fair transfer.

**Transfer fairness** Out-of-distribution fairness has been an under-explored area. Prior works can be categorized into: 1) *Group-wise distribution matching*. Schumann et al. (2019) derive an upper bound for fairness in target domain which suggests to train a fair model in source domain and match the distributions of relevant groups from two domains in feature space at the same time. Yoon et al. (2020) also do group-wise distribution matching but with Wasserstein distance. Such methods are hard to achieve if we do not have supervisions 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. Coston et al. (2019) use reweighting to deal with fairness problem under covariate shift and Giguere et al. (2022) use reweighting together with a fairness test to guarantee the fairness under demographic shift. Reweighiting methods strongly rely on the support cover assumption which might satisfied under subpopulation shift, while is not applicable to 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 model that is robust to the worst-case shift (Rezaei et al., 2020; Mandal et al., 2020). These methods also assume subpopulation shift instead of domain shift. 4) *Causal inference*. Singh et al. (2021) do causal domain adaptation and DRO based on a well-characterized causal graph which 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, Schrouff et al. (2022) 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. There are also studies aim to maintain fairness under distribution shifts through online learning (Zhang et al., 2021a) or robustness (Wang et al., 2022). To the best of our knowledge, this is the first work that use self-training to transfer fairness. Some work also study self-supervised learning and fairness, yet they use unlabeled data and self-training to improve the in-distribution fairness (Chzhen et al., 2019; Zhang et al., 2020; Chakraborty et al., 2021) which is different from our research problem.

## B PROOFS

The proof of theorem 2.1 is based on theorem B.2 and B.3.

**Theorem B.1.** *(Restatement of Lemma A.8 in Wei et al. (2021)) We assume that $U_a^i$ satisfies $(\bar{\alpha}, \bar{c})$-multiplicative expansion for $\mathbb{P}_{U_a^i}(\mathcal{M}(g_{tc})) \leq \bar{\alpha} < 1/3$ 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 subpopulation $U_a^i$ is upper bounded as:*

$$\varepsilon_{U_a^i}(g) \leq \frac{c+1}{c-1} L_{U_a^i}(g, g_{tc}) + \frac{2c}{c-1} R_{U_a^i}(g) - \varepsilon_{U_a^i}(g_{tc}) \tag{8}$$

**Theorem B.2.** *(A constraint version of the above theorem) We assume that $U_a^i$ satisfies $(\bar{\alpha}, \bar{c})$-multiplicative expansion for $\mathbb{P}_{U_a^i}(\mathcal{M}(g_{tc})) \leq \bar{\alpha} < 1/3$ 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 $L_{U_a^i}(g, g_{tc}) \leq \varepsilon_{U_a^i}(g_{tc})$, the error of it on the subpopulation $U_a^i$ is upper bounded as:*

$$\varepsilon_{U_a^i}(g) \leq \frac{2}{c-1} \varepsilon_{U_a^i}(g_{tc}) + \frac{2c}{c-1} R_{U_a^i}(g) \tag{9}$$

*Proof.*

$$\varepsilon_{U_a^i}(g) \leq \frac{c+1}{c-1} L_{U_a^i}(g, g_{tc}) + \frac{2c}{c-1} R_{U_a^i}(g) - \varepsilon_{U_a^i}(g_{tc}) \tag{10}$$

$$\varepsilon_{U_a^i}(g) \leq \varepsilon_{U_a^i}(g_{tc}) - \frac{c-3}{c-1} L_{U_a^i}(g, g_{tc}) + \frac{2c}{c-1} R_{U_a^i}(g) \qquad \text{(because } L_{U_a^i}(g, g_{tc}) \leq \varepsilon_{U_a^i}(g_{tc})\text{)}$$

$$\varepsilon_{U_a^i}(g) \leq \frac{2}{c-1} \varepsilon_{U_a^i}(g_{tc}) + \frac{2c}{c-1} R_{U_a^i}(g) \tag{11}$$

□

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

$$\varepsilon_{U_a^i}(g) \geq \varepsilon_{U_a^i}(g_{tc}) - L_{U_a^i}(g, g_{tc}) \tag{12}$$

*Proof.* By triangle inequality. □

**Theorem B.4.** *(same as theorem 2.1.) If we have a teacher classifier $g_{tc}$ with bounded unfairness such that $|\varepsilon_{U_a^i}(g_{tc}) - \varepsilon_{U_{a'}^{i'}}(g_{tc})| \leq \gamma, \forall a, a' \in \mathcal{A}$ and $i, i' \in \mathcal{Y}$. We assume that $U_a^i$ satisfies $(\bar{\alpha}, \bar{c})$-multiplicative expansion and $\varepsilon_{U_a^i}(g_{tc}) \leq \bar{\alpha} < 1/3$ and $\bar{c} > 3, \forall a, i$. We define $c \triangleq \min\{1/\bar{\alpha}, \bar{c}\}$. Set $\mu \leq \varepsilon_{U_a^i}(g_{tc}), \forall a, i$. If we train our classifier using the following algorithm*

$$\min_{g \in G} \max_{a,i} R_{U_a^i}(g) \tag{13}$$

$$\text{s.t.} \quad L_{U_a^i}(g, g_{tc}) \leq \mu \quad \forall a, i \tag{14}$$

*and denote the optimal solution as $\hat{g}$. Then the error and unfairness of $\hat{g}$ on the population distribution are bounded as*

$$\varepsilon(\hat{g}) \leq \frac{2}{c-1} \varepsilon(g_{tc}) + \frac{2c}{c-1} R(\hat{g}) \tag{15}$$

$$\Delta_{odds} \leq \frac{2}{c-1} (\gamma + \mu + c \max_{a,i} R_{U_a^i}(\hat{g})) \tag{16}$$

*Proof.* The upper bound of $\varepsilon(\hat{g})$ could be directly get from theorem B.2. With theorem B.2 and B.3, we have

$$\Delta_{odds} = \frac{1}{2}\left|\varepsilon_{U_0^0}(g) - \varepsilon_{U_1^0}(g)\right| + \frac{1}{2}\left|\varepsilon_{U_0^1}(g) - \varepsilon_{U_1^1}(g)\right| \tag{17}$$

$$\leq \frac{1}{2}\max\left\{\frac{2}{c-1}(\gamma + L_{U_1^0}(g, g_{tc}) + cR_{U_0^0}(g)), \frac{2}{c-1}(\gamma + L_{U_0^0}(g, g_{tc}) + cR_{U_1^0}(g))\right\} \tag{18}$$

$$+ \frac{1}{2}\max\left\{\frac{2}{c-1}(\gamma + L_{U_1^1}(g, g_{tc}) + cR_{U_0^1}(g)), \frac{2}{c-1}(\gamma + L_{U_0^1}(g, g_{tc}) + cR_{U_1^1}(g))\right\} \tag{19}$$

$$\leq \mu + \frac{2}{c-1}(\gamma + \max_{a,i}\ L_{U_a^i}(\hat{g}, g_{tc})). \tag{20}$$

□

# C Details of Experiments

## C.1 Synthetic dataset

The 3D Shapes dataset (Kim & Mnih, 2018) contains 64 x 64 RGB images of 3D shapes with ground truth factors: shape[4], scale[8], orientation[15], floor colour[10], wall colour[10], object colour[10]. The ground truth factors are desirable for studying synthetic distribution shifts, where we choose one factor as the class label Y, one factor as the sensitive attribute A, and one factor as the domain D. In our experiments, we chose shape as Y, object colour as A, and scale as domain. We binarized Y and A by taking two choices of each factor, to satisfy the assumption in LAFTR.

We constructed source and target domains to have 4 different types of domain shifts: (a) Subpopulation shift with the same domain choices with different distributions $\mathbb{P}_S(Y, A) = \mathbb{P}_T(Y, A)$, $\mathbb{P}_S(D) \neq \mathbb{P}_T(D)$, (b) Subpopulation shift with the same domain and different joint distribution of class Y and sensitive attribute $\mathbb{P}_S(Y, A) \neq \mathbb{P}_T(Y, A)$, $\mathbb{P}_S(D) = \mathbb{P}_T(D)$, (c) Domain shift with different domain distributions and same joint distribution of class Y and sensitive attribute A $\mathbb{P}_S(Y, A) = \mathbb{P}_T(Y, A)$, $\mathbb{P}_S(D) \neq \mathbb{P}_T(D)$, and (d) Domain shift with different domain distributions and different joint distributions of class Y and sensitive attribute A $\mathbb{P}_S(Y, A) \neq \mathbb{P}_T(Y, A)$, $\mathbb{P}_S(D) \neq \mathbb{P}_T(D)$.

Scale (D) has 8 choices which are ordered. For (a), we experimented with domain of scale with distribution in source $[\frac{4}{16}, \frac{4}{16}, \frac{3}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}]$, and domain of scale distribution in target $[\frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{3}{16}, \frac{4}{16}, \frac{4}{16}]$. The joint distribution of shape (Y) and object color (A) are the same in two domains, $Y_0A_0 = 0.1$, $Y_0A_1 = 0.4$, $Y_1A_0 = 0.4$, $Y_1A_1 = 0.1$.

For (b), we experimented with the same domain distribution of scale in source and target, $[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}]$. The joint distribution of shape (Y) and object color (A) are shifted from $Y_0A_0 = 0.1$, $Y_0A_1 = 0.4$, $Y_1A_0 = 0.4$,$Y_1A_1 = 0.1$, to $Y_0A_0 = 0.4$, $Y_0A_1 = 0.1$, $Y_1A_0 = 0.1$,$Y_1A_1 = 0.4$.

For (c), we experimented with different domain distribution of scale, $[\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0, 0]$ in source and $[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}]$ in target. The joint distribution of shape (Y) and object color (A) are the same, $Y_0A_0 = 0.1$, $Y_0A_1 = 0.4$, $Y_1A_0 = 0.4$,$Y_1A_1 = 0.1$.

For (d), we experimented with different domain distribution of scale, $[\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0, 0]$ in source and $[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}]$ in target. The joint distribution of shape (Y) and object color (A) are shifted from, $Y_0A_0 = 0.1$, $Y_0A_1 = 0.4$, $Y_1A_0 = 0.4$,$Y_1A_1 = 0.1$, to $Y_0A_0 = 0.4$, $Y_0A_1 = 0.1$, $Y_1A_0 = 0.1$,$Y_1A_1 = 0.4$.