Fair Scratch Tickets: Finding Fair Sparse Networks without Weight Training

Pengwei Tang[1,2,†] **Wei Yao**[1,2,†] **Zhicong Li**[1,2] **Yong Liu**[1,2,*]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing

tangpwei@ruc.edu.cn, busyweiyao@gmail.com, mrfive558@gmail.com, liuyonggsai@ruc.edu.cn

Abstract

Recent studies suggest that computer vision models come at the risk of compromising fairness. There are extensive works to alleviate unfairness in computer vision using pre-processing, in-processing, and post-processing methods. In this paper, we lead a novel fairness-aware learning paradigm for in-processing methods through the lens of the lottery ticket hypothesis (LTH) in the context of computer vision fairness. We randomly initialize a dense neural network and find appropriate binary masks for the weights to obtain a fair sparse subnetworks without any weight training. ***Interestingly, to the best of our knowledge, we are the first to discover that such sparse subnetworks with inborn fairness exist in randomly initialized networks, achieving an accuracy-fairness trade-off comparable to that of dense neural networks trained with existing fairness-aware in-processing approaches.*** *We term these fair subnetworks as Fair Scratch Tickets (FSTs). We also theoretically provide fairness and accuracy guarantees for them. In our experiments, we investigate the existence of FSTs on various datasets, target attributes, random initialization methods, sparsity patterns, and fairness surrogates. We also find that FSTs can transfer across datasets and investigate other properties of FSTs.*

1. Introduction

In recent years, deep neural networks (DNN) has become one of the core technologies in computer vision (CV). However, it has been observed that CV models learn spurious age, gender, and race correlations when trained for seemingly unrelated tasks [7, 65]. There are growing appeals for fairness-aware learning [56]. A model should not discriminate against any demographic group with sensitive attributes [3, 15, 58, 61, 74].

Extensive work has been done to alleviate unfairness in CV using pre-processing [35, 52, 62, 64], in-processing [5, 6, 12, 55], and post-processing methods [37, 72]. Only in-processing approaches can optimize notions of fairness during model training. Such methods have direct control over the optimization function of the model [8] and have attracted great attention in the research community. Popular in-processing ideas include fairness regularization [5, 12, 13, 33, 47, 50, 55, 67] and fairness-aware adversarial training [6, 19, 42, 70]. Fairness regularization is to introduce regularization terms to penalize unfairness. Fairness-aware adversarial training uses an adversary to predict the sensitive attribute and enforces the main classifier to prevent the adversary from predicting successfully. However, most in-processing methods leverage deep and dense neural networks so that they are computationally intensive during the inference phase [28].

In this paper, to fill the research gap, we raise an intriguing and challenging question: *Is there a learning paradigm without weight training that is plug-and-play for bias mitigation approaches in computer vision*? Intuitively, the recently proposed Lottery Ticket Hypothesis (LTH) [20] is a natural fit for our needs. LTH focuses on finding sparse trainable subnetworks (winning tickets) that reach test accuracy comparable to the original dense neural network. The primal training method in [20] is iteratively pruning and retraining the neural network. Interestingly, some researchers empirically discover that winning tickets can be found without weight training [51, 73], which is theoretically validated in [14, 43, 46, 48]. Both empirical observations and theoretical results have verified the feasibility of finding winning tickets without training the weights of the neural networks. Motivated by the above, we break down the original question into three sub-questions instead:

- Q1: Is there a fair winning ticket?
- Q2: How can we find it without weight training?
- Q3: Is it easy to generalize on various datasets, tar-

†Equal Contribution. *Corresponding author.

get attributes, random initialization methods, sparsity patterns and fairness surrogates?

For the first question, Proposition 1 states that a sufficiently over-parameterized neural network with random weights contains a subnetwork that can approximate any target neural network with high probability under some conditions. Furthermore, our Theorem 1 shows that if we successfully find a sparse neural network that approximates a fair and accurate neural network well, then the sparse neural network is also fair and accurate. Combining the results of Proposition 1 and Theorem 1, they answer our first question by clarifying the possibility of finding fair and accurate winning tickets without any weight training. To our best knowledge, LTH remains poorly understood in the context of fairness. **For the second question**, note that the proof of Theorem 2.1 in [43] follows a constructive routine for masking. Therefore, it sheds light on the feasibility of finding fair winning tickets without any weight training by designing an appropriate masking scheme, and that is exactly what we do. We randomly initialize a DNN and search for masks to iteratively find Fair Scratch Tickets (FSTs). In particular, following [51], we search for the best binary masks by optimizing a continuously updated learnable score for each weight. **For the third question**, to verify the generality of FST, we demonstrate its effectiveness in two famous types of in-processing approaches in CV fairness: fairness regularization [5] and fairness-aware adversarial training [70]. Extensive experiments verify the existence of FSTs on various datasets, target attributes, random initialization methods, sparsity patterns and fairness surrogates. We further show the properties of fine-tuning and transferability of FSTs.

Overall, our contributions are threefold:

- We are the first to theoretically and empirically confirm the existence of *winning tickets with inborn fairness*. And we extend the application scenario of LTH to CV fairness.
- We are the first to propose a brand new *plug-and-play* learning paradigm that does not require weight training for the CV fairness community.
- Extensive experiments verify the existence of FSTs on various datasets, target attributes, random initialization methods, sparsity patterns and fairness surrogates. Furthermore, we show the properties of fine-tuning and transferability of FSTs.

2. Related Work

2.1. Fairness in Computer Vision

In the past few years, based on the observation that facial image analysis systems cause substantial accuracy disparities for different sensitive groups [7], there has been a growing number of papers on fairness in computer vision [59, 60]. Most of the existing work in this field falls into three categories: pre-processing, in-processing, and post-processing. Similar categories also appear in the fair machine learning literature, which is exhaustively surveyed in [8, 44].

Pre-processing methods are data operations that focus on changing the data itself to mitigate unwanted bias. Most of them use deep models to incorporate techniques such as image generation [17, 35, 52, 71], sampling [54, 57], reweighing [2, 36], masking [62], perturbation [64], etc. As a result, the pre-processed or augmented images can be used to train fairer models. **Post-processing** methods try to modify the prediction results to satisfy the fairness definitions, e.g., [30, 37, 72]. **In-processing** is the research emphasis of this paper. Such approaches learn sensitive-free features from data during training. Popular ideas include fairness regularization [5,12,13,33,47,50,55,67] and fairness-aware adversarial training [6, 19, 42, 70]. *Fairness regularization* incorporates unfairness penalty terms into the objective. The penalty can be designed according to intuitions from a specific fairness criterion [5, 12, 67], disentangling meaningful and sensitive representations [13,47,50,55], and others like [1,33]. *Fairness-aware adversarial training* uses an adversary [6, 19, 42, 70] to predict the sensitive attribute of the training set. Then the main classifier should act in opposition to fool the adversary and at the same time accomplish the main prediction task. Among pre-processing, in-processing, and post-processing, a key advantage of in-processing is that it can easily incorporate fairness considerations into the optimization objective. Consequently, there is a high flexibility in picking the accuracy-fairness trade-off, and in-processing has attracted great attention in the research community. However, deep and dense neural networks are commonly used in in-processing models and thus making the inference phase time-consuming.

In contrast to many methods mentioned above that require training a neural network from scratch, our FSTs suffer from less computational burden because they are sparse and do not require any weight training. Furthermore, FSTs also serve as a universally adaptable plug-in for any DNN-based approaches in CV fairness so that it can be naturally combined with existing DNN-based fair CV models.

2.2. Lottery Ticket Hypothesis

A recently proposed technique called Lottery Tickets Hypothesis (LTH) [20] leads a fast-rising field that investigates sparse trainable subnetworks within fully dense networks [14, 21–23, 39, 41, 43, 46, 48, 53, 63, 73]. The original lottery ticket hypothesis states that in a randomly initialized dense neural network, there is a sparse subnetwork that can achieve similar test accuracy when trained in isolation [20]. The sparse neural network is called "winning tickets" and can be found by iteratively pruning the dense network. In the follow-up work [22, 53], the authors introduce LTH with rewinding to enable LTH for deeper models and larger datasets. The robustness, learning dynamics, and underlying condition of LTH are also dissected in [21, 23, 39], respectively. LTH has been extensively explored in various application scenarios like image classification [9, 25], natural language processing [10, 49] and graph neural networks [11]. In addition, winning tickets can be found with some inborn characteristics, such as robustness [24] and differential privacy [27].

Going a step further, in particular, there is a refreshing line of work empirically discovering that winning tickets can be found with little training [68] or even no training [51, 73]. From a theoretical perspective, the researchers even prove that winning tickets can be found without any training under some conditions [43]. And this result is further improved by [46, 48], which shows that logarithmic over-parameterization is sufficient. It is extended to convolutional neural networks in the follow-up work [14]. In general, both empirical observations and theoretical results have verified the feasibility of finding winning tickets without weight training. In support of the above observations and theory, an orthogonal work [24] to ours successfully finds robust winning tickets without training the weights. A piece of related work is [29]. They empirically study the impact of some pruning strategies on fairness in natural language processing. Distributionally robust optimization loss [38] is considered to find a fair winning ticket. By comparison, our approach differs from their work in that our FSTs do not require training the weights of the neural network, and we focus more on CV fairness.

Notably, although extensive research has been done on LTH, to the best of our knowledge, there has been no previous research that provides evidence for fair winning tickets without weight training in the field of computer vision. Therefore, in the perspective of application scenario of LTH, we motivate the research community that it is possible to obtain a fair winning ticket without weight training in computer vision.

3. Preliminaries

3.1. Fair Classification

$\mathcal{X}$ is the feature space. $\mathcal{Y} = \{-1, 1\}$ and $\mathcal{S} = \{a, b\}$ represent the space of class labels and sensitive attributes, respectively. The training set $\widehat{\mathcal{D}}_{\mathcal{Z}} = \{(x_i, s_i, y_i)\}_{i=1}^N$ is drawn from the distribution $\mathcal{D}_{\mathcal{Z}}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$. It consists of three parts: predictive features $x \in \mathcal{X}$, sensitive attribute $s \in \mathcal{S}$ and target attribute $y \in \mathcal{Y}$. There are N_{sy} data with sensitive attribute s and label y, $N_{s\cdot}$ data with sensitive attribute s and any label, and $N_{\cdot y}$ data with label y and any group. The predicted target label is $\hat{y} \in \mathcal{Y}$. A classifier $f(\theta, x) : \mathcal{X} \mapsto \mathbb{R}$ is parameterized by θ. If $f(\theta, x) > 0$, then $\hat{y} = 1$. The training set accuracy is

$$\mathrm{ACC}(f) = \frac{1}{N} \sum_{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}} \mathbb{I}_{y=\hat{y}},$$

where $\mathbb{I}_{[\cdot]}$ is the indicator function.

In this paper, we focus on two widely used fairness metrics: demographic parity (DP) [18] and equality of opportunity (EO) [30]. The *difference in demographic parity* (DDP) is $\mathbb{P}(\hat{y} = 1|s = a) - \mathbb{P}(\hat{y} = 1|s = b)$. We use the empirical version of DDP to indicate the violation of DP:

$$\widehat{\mathrm{DDP}}(f) = \frac{1}{N_{a\cdot}} \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f(x)>0} - \frac{1}{N_{b\cdot}} \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f(x)>0}.$$

Similarly, the *difference in equality of opportunity* (DEO) is $\mathbb{P}(\hat{y} = 1|s = a, y = 1) - \mathbb{P}(\hat{y} = 1|s = b, y = 1)$. And its empirical version is

$$\widehat{\mathrm{DEO}}(f) = \frac{1}{N_{a1}} \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a \\ y=1}} \mathbb{I}_{f(x)>0} - \frac{1}{N_{b1}} \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b \\ y=1}} \mathbb{I}_{f(x)>0}.$$

For a fairness threshold $\delta > 0$, the fair classification task is to find a classifier f such that $\left|\widehat{\mathrm{DDP}}(f)\right| \leq \delta$ (or $\left|\widehat{\mathrm{DEO}}(f)\right| \leq \delta$). In the experiments, $\widehat{\mathrm{DDP}}$ and $\widehat{\mathrm{DEO}}$ are indicators to measure the violation of specific fairness metrics.

3.2. LTH without Weight Training

The original LTH iteratively prunes a small fraction of weights and retrains the remaining weights. However, in this work, we focus on finding winning tickets that do not require weight training. As a consequence, once the neural network $f(\theta)$ is randomly initialized, the weights $\theta \in \mathbb{R}^d$ are fixed. We search for binary masks $m \in \{0, 1\}^d$ to find a winning ticket $f(\theta \odot m)$, where $\odot$ is the element-wise product.

Previous theoretical work proves that winning tickets can be found without any weight training under some conditions [14, 43]. We briefly review their conclusions below.

Proposition 1. *To approximate any target neural network $f^*(\theta^*)$, from a randomly initialized deep and wide enough neural network $f(\theta)$, we can find a sparse subnetwork $f(\theta \odot m)$ such that $\forall x_i \in \mathcal{X}$ and some $\epsilon > 0$, the inequality $|f^*(\theta^*, x_i) - f(\theta \odot m, x_i)| \leq \epsilon$ holds with high probability.*

Proposition 1 is an informal version of the conclusions in [14, 43]. The detailed theorem and proof can be found in their papers. Thus, to approximate $f^*(\theta^*)$, it is quite possible to find a good approximation $f(\theta \odot m)$ from a deep and wide enough $f(\theta)$ without weight training.

4. Drawing Fair Scratch Tickets

4.1. Do FSTs Exist?

In Theorem 1, we extend the results in Proposition 1 and validate the existence of FSTs. We demonstrate that the FSTs are both fair and accurate.

Theorem 1. *Given the training set* $\widehat{\mathcal{D}}_{\mathcal{Z}} = \{(x_i, s_i, y_i)\}_{i=1}^N$, *approximation error threshold* $\epsilon > 0$, *fairness tolerance* $\delta_{f^*} > 0, \delta_{f'} > 0$, *accuracy lower bound* $\delta_{acc} > 0$. *Assume that the following conditions hold:*

(A) a sufficiently large training set: $N \geq \frac{\sum_{i=1}^N \mathbb{I}_{|f^*(x_i)| \leq \epsilon}}{\delta_{f'}}$,

(B) a fair and accurate neural network f^* *that satisfies* $\left|\widehat{DDP}(f^*)\right| \leq \delta_{f^*}$ *and* $ACC(f^*) \geq \delta_{acc}$,

(C) a neural network $f' = f(\theta \odot m)$ *such that* $\forall x_i \in \mathcal{X}$, *there holds* $|f^*(x_i) - f'(x_i)| \leq \epsilon$.

Then f' *is fair and accurate:*

$$\begin{cases} \left|\widehat{DDP}(f')\right| \leq \delta_{f^*} + \delta_{f'}, (\textit{Fairness}) \\ ACC(f') \geq \delta_{acc} - \delta_{f'}. (\textit{Accuracy}) \end{cases}$$

The proof and EO version of this theorem are given in the supplementary. Theorem 1 ensures that if a fair and accurate neural network f^* and $f(\theta \odot m)$ share similar results for any input feature, then for a sufficiently large training set, there are fairness and accuracy guarantees for the winning ticket $f(\theta \odot m)$, which is our FST. Notice that all of the three conditions are natural and not restrictive. For assumption *(A)*, $\sum_{i=1}^N \mathbb{I}_{|f^*(x_i)| \leq \epsilon}$ is the number of points that are close to the decision boundary. When ϵ is small, there holds $\sum_{i=1}^N \mathbb{I}_{|f^*(x_i)| \leq \epsilon} \ll N$. So the condition $N \geq \frac{\sum_{i=1}^N \mathbb{I}_{|f^*(x_i)| \leq \epsilon}}{\delta_{f'}}$ can be satisfied. For assumption *(B)*, although f^* is an ideal neural network, at least any fair and accurate neural networks in previous fairness-aware methods can be cases of f^*. So this assumption is naturally satisfied based on existing works. For assumption *(C)*, its reasonability has been validated by Proposition 1 and theoretical justifications [14, 43], which means that this assumption is also a mild one for our theorem.

In summary, we now establish the relation between our analysis and FST. We initialize $f(\theta)$ with random weights θ. We keep θ unchanged and only search for masks m to find the winning ticket $f(\theta \odot m)$. It can be found with high probability because of Proposition 1. According to Theorem 1, when we find the winning ticket, it is guaranteed to be fair and accurate. The fair and accurate winning ticket $f(\theta \odot m)$ is just our FST.

4.2. How to Search for FSTs?

Our method operates on each convolutional layer. In a randomly initialized dense network $f(\theta)$, θ_l denotes the weights of l-th layer of $f(\theta)$ and m_l denotes the binary masks associated with θ_l. Given a pre-defined weight remaining ratio η $(0 < \eta < 1)$, FST search is equivalent to finding appropriate binary masks m for untrained weights θ. Generally, FST search can be formulated as

$$\begin{aligned} \widehat{m} \in \arg\min_m \frac{1}{N} \sum_i \ell(f(\theta \odot m, x_i), y_i, s_i), \\ s.t. \|m_l\|_0 = \eta \cdot n_l, l = 1, \ldots, L \end{aligned} \tag{1}$$

where $\widehat{m}$ is the winning binary masks of FST, ℓ is the fairness loss function, n_l is the number of weights in layer l and L is the number of layers in $f(\theta)$.

Motivated and guided by prior works [24, 51] which find winning scratch tickets in randomly initialized neural networks, we search for winning binary masks m by iteratively updating learnable scores r attached to each randomly initialized weight. Given a pre-defined remaining ratio η, we obtain winning scratch tickets by retaining the weights in each layer which own the top-η highest scores and discarding the other weights. The learnable scores r is updated by gradient descent, which is written as

$$r = r - \frac{\partial \frac{1}{N} \sum_i \ell(f(\theta \odot m, x_i), y_i, s_i)}{\partial r}.$$

After each updating of r, the binary masks m_l of layer l are correspondingly updated by

$$m_{i,l} = \begin{cases} 1, & r_{i,l} \geq r_{\eta,l} \\ 0, & r_{i,l} < r_{\eta,l} \end{cases},$$

where $r_{i,l}$ denotes the i-th weight in layer l and $r_{\eta,l}$ represents the value of the score ranking exactly top-η in layer l. Our search only learns the attached scores r by gradient descent and obtains winning scratch tickets without any weight training.

Next, we introduce two specific search methods for FSTs under fairness regularization and fair adversarial training.

4.3. FST Search under Fairness Regularization

Fairness regularization improves the fairness of prediction by incorporating a fairness penalty into the objective function, which is formulated as

$$\arg\min_{m} \frac{1}{N} \sum_{i} \ell_c(f(\theta \odot m, x_i)), y_i) + \lambda R_g(x_i, y_i, s_i),$$
$$s.t. \|\widehat{m_l}\|_0 = \eta \cdot n_l, l = 1, \ldots, L \tag{2}$$

where R_g denotes the fairness regularization, ℓ_c is loss function and λ is the regularization coefficient.

Following [5], to optimize DDP and DEO, the regularization is given by

$$R_{ddp}(x, y, s) = \begin{cases} \frac{u(f(\theta,x))}{p_a}, & s = a \\ \frac{u(-f(\theta,x))}{p_b}, & s = b \end{cases}, (DDP) \tag{3}$$

$$R_{deo}(x, y, s) = \begin{cases} \frac{u(f(\theta,x))}{p_{a1}}, & s = a, y = 1 \\ \frac{u(-f(\theta,x))}{p_{b1}}, & s = b, y = 1 \\ 0, & \text{otherwise} \end{cases}, (DEO) \tag{4}$$

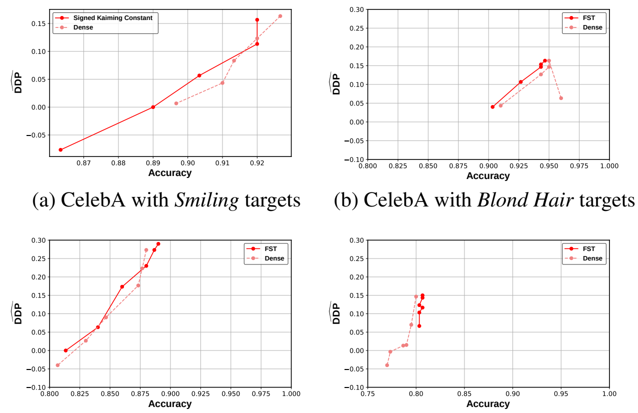
where $u(\cdot)$ is a smooth surrogate of the indicator function.

4.4. FST Search under Adversarial Training

Fairness-aware adversarial training aims to mitigate bias by avoiding the prediction of sensitive attributes from the representation or target output. We adopt the method proposed in [6] to verify the existence of FSTs under adversarial debiasing methods. The network in this method has three sub-components, including a shared representation encoder e, a target prediction head t, and an adversarial head o. We denote the parameters of these three sub-components as θ_e, θ_t and θ_o, respectively. The binary masks m also include three corresponding sub-components, *i.e.*, m_e , m_t and m_o. The goal of this method is to make $e(\theta_e, x)$ produce a fair representation, $t(\theta_t, e(\theta_e, x))$ can predict the targets , $o(\theta_o, e(\theta_e, x))$ can predict the sensitive attributes. This method adopts a special identity function $J_\lambda(\cdot)$ with negative gradient where $J_\lambda(x) = x$ and $\frac{\partial J_\lambda(e(\theta_e,x))}{\partial x} = -\lambda \frac{\partial e(\theta_e,x)}{\partial x}$. The objective function of the adversarial method can be formulated as

$$\arg\min_{m} [\frac{1}{N} \sum_{(x_i,y_i)} \ell_y(t(\theta_t \odot m_t, e(\theta_e \odot m_e, x_i)), y_i)$$
$$+ \lambda \frac{1}{N} \sum_{(x_i,y_i,s_i)} \ell_z(o(\theta_o \odot m_o, J_\lambda(e(\theta_e \odot m_e, x_i))), s_i)],$$
$$s.t. \|\widehat{m_l}\|_0 = \eta \cdot n_l, l = 1, \ldots, L \tag{5}$$

where both ℓ_y and ℓ_z are loss functions, and λ is the trade-off coefficient.



(a) CelebA with *Smiling* targets (b) CelebA with *Blond Hair* targets

(c) LFW with *Smiling* targets (d) LFW with *Wary Hair* targets

Figure 1. FSTs exist under R_{ddp} regularization on CelebA and LFW datasets with remaining ratio $\eta = 10\%$.

5. Experiments

5.1. Experimental Setup

We briefly introduce some necessary experimental setup here. More details are provided in the supplementary.

Datasets. We evaluate the existence and property of FSTs on two real-world face image datasets, *i.e.*, CelebA [40] and LFW [34]. We adopt *gender* as the sensitive attribute. We use *Smiling* and *Blond Hair* as the target labels on CelebA and take *Smiling* and *Wary Hair* as the target labels on LFW.

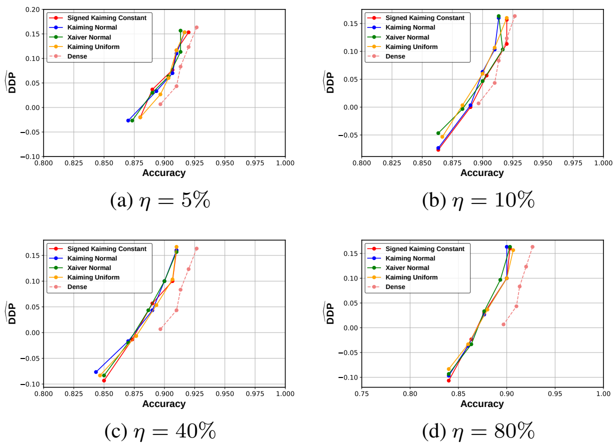
Model initialization. In our experiments, we consider four widely used initialization methods, *i.e.*, Kaiming Uniform [31], Kaiming Normal [31], Signed Kaiming Constant [51], Xavier Normal [26]. We use the Signed Kaiming Constant as the default initialization method.

Implementation details. We use ResNet18 [32] as the network architecture in our experiments. We train a network with training set, select the network weights with the best accuracy in validation set, and report the accuracy and unfairness in test set. The reported results are the average of three trials with different random seeds.

Evaluation metrics. For evaluation, we use the accuracy-fairness trade-off by varying the coefficient λ in the objective. A better accuracy-fairness trade-off means higher accuracy and fairness metrics closer to zero. We take accuracy as the x-axis and fairness metrics as the y-axis. In the experiments in our main paper, we only consider $\widehat{\text{DDP}}$. The corresponding experiments for $\widehat{\text{DEO}}$ are deferred to the supplementary.

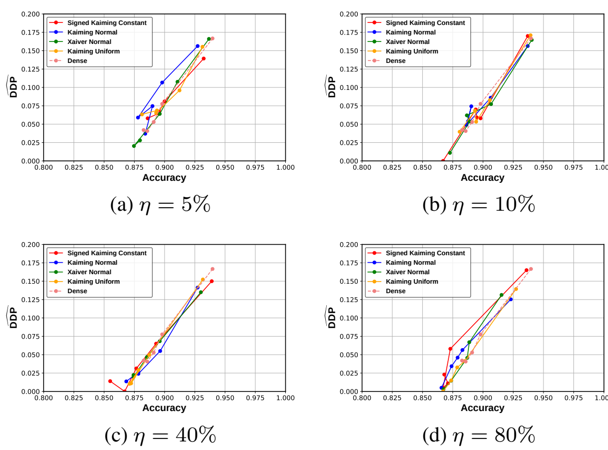
5.2. The Existence of Fair Scratch Tickets

We call the fair dense networks trained with existing fairness-aware in-processing methods "**dense counterparts**" for short. We plot the results of FSTs and their vari-



(a) $\eta = 5\%$ (b) $\eta = 10\%$ (c) $\eta = 40\%$ (d) $\eta = 80\%$

Figure 2. FSTs exist under R_{ddp} regularization with four initialization methods on CelebA with *Smiling* targets.



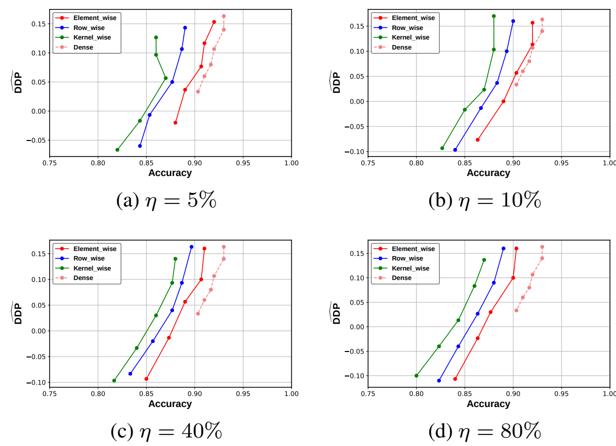
(a) $\eta = 5\%$ (b) $\eta = 10\%$ (c) $\eta = 40\%$ (d) $\eta = 80\%$

Figure 3. FSTs exist under adversarial training with four initialization methods on CelebA with *Blond Hair* targets.

ants using **solid lines** and the results of dense counterparts using **dashed lines**.

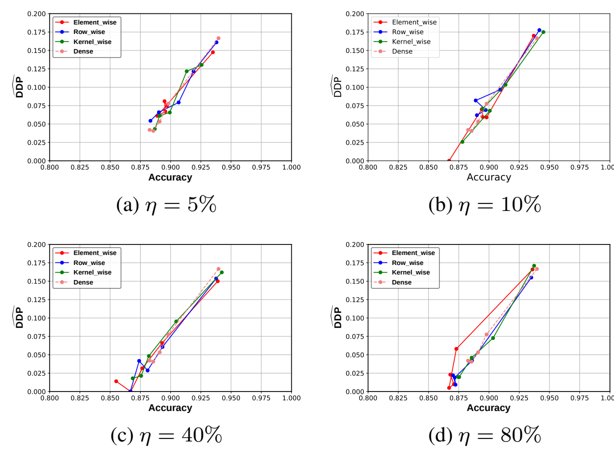
In Fig. 1, we show the empirical existence of FSTs under R_{ddp} regularization on CelebA and LFW with a widely used remaining ratio $\eta = 10\%$. The corresponding experiments for adversarial training are deferred to the supplementary. We can see that: (1) in Figs. 1a to 1d, the accuracy-fairness trade-off of FSTs are very close to the trade-off of the dense counterparts; (2) the accuracy-fairness trade-off of FSTs can outperform the dense counterparts in some cases; (3) in Fig. 1d, FSTs can even consistently outperform the dense counterparts.

Overall, it verifies that sparse subnetworks with inborn fairness do exist in randomly initialized dense networks and have comparable or even better accuracy-fairness trade-off than the dense counterparts, without any weight training.



(a) $\eta = 5\%$ (b) $\eta = 10\%$ (c) $\eta = 40\%$ (d) $\eta = 80\%$

Figure 4. FSTs exist under R_{ddp} regularization with different sparsity patterns on CelebA with *Smiling* targets.



(a) $\eta = 5\%$ (b) $\eta = 10\%$ (c) $\eta = 40\%$ (d) $\eta = 80\%$

Figure 5. FSTs exist under adversarial training with different sparsity patterns on CelebA with *Blond Hair* targets.

5.3. FSTs Exist under Different Remaining Ratios

In Figs. 2 and 3, we show the accuracy-fairness trade-off of FSTs on CelebA with *Smiling* targets (for fairness regularization) and *Blond Hair* targets (for adversarial training) under a wide range of remaining ratios (*i.e.*, $\eta = 5\% \sim 80\%$) with four different initialization methods.

In Fig. 2, under R_{ddp} regularization, we can observe that: (1) FSTs have comparable accuracy-fairness trade-off to the dense counterparts under a wide range of weight remaining ratios (*i.e.*, $\eta = 5\% \sim 80\%$), even without any weight training; (2) FSTs perform best under the remaining ratio $\eta = 10\%$, indicating that an appropriate remaining ratio plays an important role in FSTs. It shows that FSTs with low or high remaining ratio have relatively worse performance than FSTs with the best appropriate remaining ratio. When the weight remaining ratio is low, FSTs suffer from being under-parameterized due to the small capacity of the subnetworks. While the original randomly initialized weights are retained at high ratio level, FSTs are close to the randomly initialized networks and incline to make random predictions. In Fig. 3, the results also follow a similar trend under adversarial training: although some FSTs can outperform the dense in all reported remaining ratios, FSTs still suffer from performance drop when the remaining ratios are low (*e.g.*, $\eta = 5\%$) or high (*e.g.*, $\eta = 80\%$).

In summary, FSTs have comparable or even superior performance to the dense counterparts, and less inference time makes FSTs more advantageous.

5.4. FSTs Exist under Different Initialization

As shown in Figs. 2 and 3, when applying four different widely used distributions to randomly initialize the dense networks, FSTs consistently exist and achieve comparable or even better accuracy-fairness trade-off, showing that our FST search method is general.

5.5. FSTs Exist under Different Sparsity Patterns

We investigate the impact of structured sparsity patterns of FSTs and visualize their accuracy-fairness trade-off in Figs. 4 and 5. Besides element-wise sparsity, we consider other two structured sparsity patterns: row-wise sparsity and kernel-wise sparsity. We can observe that FSTs do exit under different sparsity patterns. Moreover, Fig. 4 shows that a more structured sparsity pattern leads to FSTs with more inferior performance under fairness regularization. In Fig. 5, the element-wise sparsity also suffers from a performance drop when the remaining ratio is low or high. However, the structured sparsity patterns (*i.e.*, row-wise sparsity and kernel-wise sparsity) show a different trend that FSTs can outperform the dense counterparts with considerably high remaining ratios (*e.g.*, even $\eta = 80\%$).

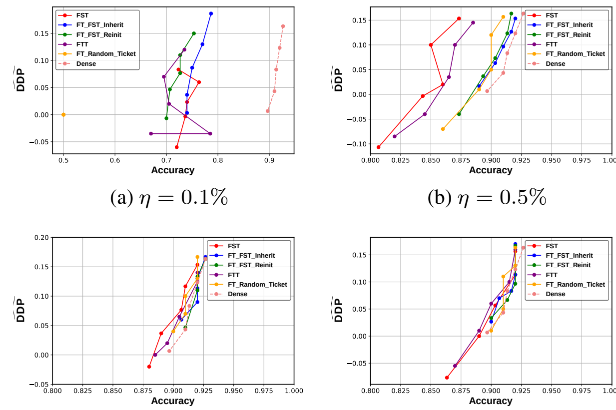
We also study how FSTs exist under different fairness surrogates, including linear [4, 16, 69], hinge [66], and logistic [5] surrogates in the supplementary.

6. The Properties of FSTs

6.1. Fine-tuned Random Tickets and Fine-tuned FSTs

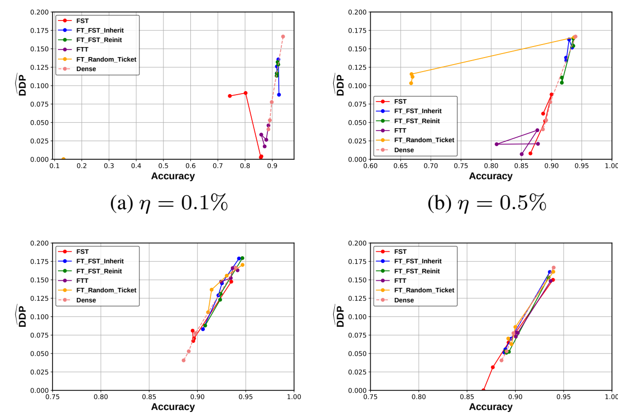
In randomly initialized networks, we randomly select weights of each convolutional layer with pre-defined remaining ratios to obtain random tickets. We fine-tune random tickets to obtain fine-tuned random tickets.

In Figs. 6 and 7, we first compare the fine-tuned random tickets with the dense counterparts. We can observe that: (1) fine-tuned random tickets suffer from model collapse under very low remaining ratios (*e.g.*, Figs. 6a and 7a); (2) fine-tuned random tickets can have comparable performance to



(a) $\eta = 0.1\%$ (b) $\eta = 0.5\%$ (c) $\eta = 5\%$ (d) $\eta = 10\%$

Figure 6. Comparisions of FST variants under R_{ddp} regularization on CelebA with *Smiling* targets.



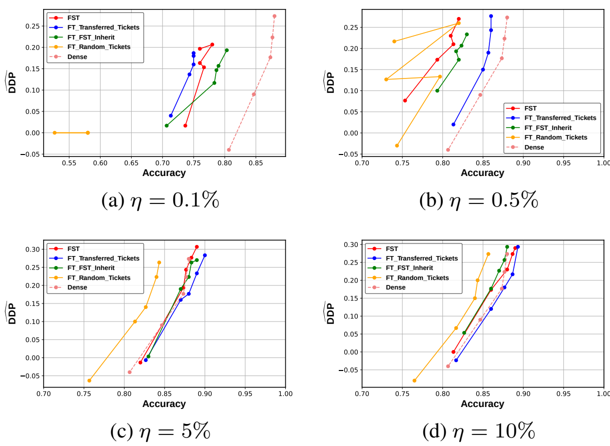
(a) $\eta = 0.1\%$ (b) $\eta = 0.5\%$ (c) $\eta = 5\%$ (d) $\eta = 40\%$

Figure 7. Comparisions of FST variants under adversarial training on CelebA with *Smiling* targets.

the dense counterparts under relatively high remaining ratios (*e.g.*, Figs. 6d and 7d), which is expected due to the large capacity of subnetworks under relatively high remaining ratios. Thus, when studying the fine-tuning properties, we only consider the relatively low remaining ratios (*e.g.*, $\eta \leq 10\%$ in Fig. 6 and $\eta \leq 40\%$ in Fig. 7).

Following [24, 51], we also consider two fine-tuning settings: (1) fine-tuning FSTs with initialization inherited from the vanilla FSTs, and (2) fine-tuning FSTs with random re-initialization of the vanilla FSTs.

In Fig. 6, under fairness regularization, we can find: (1) fine-tuned FSTs can improve performance of the vanilla FSTs under relatively high remaining ratios (*e.g.*, $\eta \geq 0.5\%$); (2) fine-tuned FSTs under high remaining ratios (*e.g.*, $\eta = 5\%$ and 10%) have performance very close to the dense counterparts, which is expected due to large



(a) $\eta = 0.1\%$ (b) $\eta = 0.5\%$

(c) $\eta = 5\%$ (d) $\eta = 10\%$

Figure 8. Comparisons between fine-tuned transferred FSTs and other methods under R_{ddp} on LFW with *Smiling* targets.

capacity of networks; (3) fine-tuned FSTs with inherited weights outperform fine-tuned FSTs with randomly reinitialized weights when weights remaining ratios are low (*e.g.*, $\eta = 0.1\%$ and $\eta = 0.5\%$) and these two fine-tuned FSTs have comparable performance when the remaining ratios are high (*i.e.*, $\eta \geq 5\%$), indicating that FSTs can find initialization particularly adept at further fairness learning; (4) fine-tuned FSTs outperform fine-tuned random tickets under low remaining ratios, *i.e.*, under-parameterization, showing that FSTs find good network architectures that are adept at fairness learning.

In Fig. 7, under adversarial training, fine-tuned FSTs have different properties: although fine-tuned FSTs can improve the performance of FSTs under low remaining ratios (*e.g.*, $\eta = 10\%$ and $\eta = 0.5\%$), the fine-tuned FSTs even have inferior performance to the vanilla FSTs (*e.g.*, $\eta = 5\%$ and $\eta = 10\%$). It shows that under fair adversarial training, FSTs without weight training is really a good approach to fairness.

Overall, FSTs can find combinations of sparse architectures and initialization that are with inborn fairness and even particularly adept at further fairness learning.

6.2. FTTs Drawn from Trained Dense Networks

Here, we investigate the winning tickets drawn from dense networks trained with existing in-processing fairness method, which is called Fair Trained Tickets (FTTs). The accuracy-fairness trade-off of FTTs are also shown in Figs. 6 and 7. We can find that FTTs have inferior performance to the fine-tuned FSTs in the vast majority of cases, except Figs. 7c and 7d, suggesting that firstly finding untrained tickets from randomly initialized networks then fine-tuning the remaining weights is better than firstly training weights then finding tickets from the trained networks.

6.3. Transferability of FSTs across Datasets

Inspired by [45], we conduct experiments to study the transferability of FSTs. As shown in Fig. 8, we fine-tune the FSTs drawn from large dataset to small dataset, *i.e.*, from CelebA with *Smiling* targets to LFW also with *Smiling* targets. We can see that when the remaining ratios are relatively high (*e.g.*, $\eta = 0.5\%$, 5% and 10%), the fine-tuned transferred FSTs perform better than other methods, including the vanilla FSTs and the fine-tuned FSTs, and even better than the dense counterparts (*e.g.*, $\eta = 5\%$, 10%). It verifies that our FSTs have good transferability. Although the weights of FSTs are untrained and selected from randomly initialized dense networks, our FST search method really have good understanding of training set.

7. Conclusion

In this work, we propose a novel fairness-aware learning paradigm for in-processing methods in computer vision from the perspective of the lottery ticket hypothesis. We are the first to theoretically and empirically verify that subnetworks drawn from randomly initialized neural networks can achieve comparable or even better accuracy-fairness trade-off than the existing in-processing methods, without any weight training. We provide theoretical guarantees for the fairness and accuracy of FSTs. Extensive experiments show that FSTs can generalize on various datasets, target attributes, random initialization methods, sparsity patterns, and fairness surrogates. Furthermore, we study the properties of fine-tuning and transferability of FSTs. Throughout the theoretical justification and extensive experiments, we show that our FSTs are effective, and we believe that our study can provide new insights into the CV fairness community.

Acknowledgement

This work is supported by the National Natural Science Foundation of China No.62076234; the Beijing Natural Science Foundation No.4222029; the "Intelligent Social Governance Interdisciplinary Platform, Major Innovation Planning Interdisciplinary Platform for the "Double First Class" Initiative, Renmin University of China"; the Beijing Outstanding Young Scientist Program No.BJJWZYJH012019100020098; the Public Computing Cloud, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China No.2021030199; the Huawei-Renmin University joint program on Information Retrieval; the Unicom Innovation Ecological Cooperation Plan; the CCF-Huawei Populus Grove Fund; and the National Key Research and Development Project (Grant No.2022YFB2703102)

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Workshop on the European Conference on Computer Vision*, 2018. 2

[2] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019. 2

[3] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016. 1

[4] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017. 7, 16

[5] Henry C Bendekgey and Erik Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. In *Advances in Neural Information Processing Systems*, 2021. 1, 2, 5, 7, 16

[6] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017. 1, 2, 5

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018. 1, 2

[8] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020. 1, 2

[9] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[10] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. In *Advances in Neural Information Processing Systems*, 2020. 3

[11] Tianlong Chen, Yongduo Sui, Xuxi Chen, Aston Zhang, and Zhangyang Wang. A unified lottery ticket hypothesis for graph neural networks. In *Proceedings of the International Conference on Machine Learning*, 2021. 3

[12] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *Proceedings of the International Conference on Machine Learning*, 2021. 1, 2

[13] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the International Conference on Machine Learning*, 2019. 1, 2

[14] Arthur da Cunha, Emanuele Natale, and Laurent Viennot. Proving the strong lottery ticket hypothesis for convolutional neural networks. In *Proceedings of the International Conference on Learning Representations*, 2022. 1, 2, 3, 4

[15] Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. Fifa: Making fairness more generalizable in classifiers trained on imbalanced data. In *Proceedings of the International Conference on Learning Representations*, 2023. 1

[16] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018. 7, 16

[17] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. In *Advances in Neural Information Processing Systems*, 2021. 2

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012. 3

[19] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *Proceedings of the International Conference on Learning Representations*, 2015. 1, 2

[20] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of the International Conference on Learning Representations*, 2019. 1, 2, 3

[21] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the International Conference on Machine Learning*, 2020. 2, 3

[22] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*, 2019. 2, 3

[23] Jonathan Frankle, David J Schwab, and Ari S Morcos. The early phase of neural network training. In *Proceedings of the International Conference on Learning Representations*, 2020. 2, 3

[24] Yonggan Fu, Qixuan Yu, Yang Zhang, Shang Wu, Xu Ouyang, David Cox, and Yingyan Lin. Drawing robust scratch tickets: Subnetworks with inborn robustness are found within randomly initialized networks. In *Advances in Neural Information Processing Systems*, 2021. 3, 4, 7

[25] Sharath Girish, Shishira R Maiya, Kamal Gupta, Hao Chen, Larry S Davis, and Abhinav Shrivastava. The lottery ticket hypothesis for object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[26] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010. 5

[27] Lovedeep Gondara, Ricardo Silva Carvalho, and Ke Wang. Training differentially private neural networks with lottery tickets. In *Proceedings of the European Symposium on Research in Computer Security*, 2021. 3

[28] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In

Advances in Neural Information Processing Systems, 2015. 1

[29] Victor Petrén Bach Hansen and Anders Søgaard. Is the lottery fair? evaluating winning tickets across demographics. In *Proceedings of the Findings of the Association for Computational Linguistics*, 2021. 3

[30] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016. 2, 3

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015. 5

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[33] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, 2018. 1, 2

[34] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 5

[35] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*, 2020. 1, 2

[36] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2020. 2

[37] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019. 1, 2

[38] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2020. 3

[39] Ning Liu, Geng Yuan, Zhengping Che, Xuan Shen, Xiaolong Ma, Qing Jin, Jian Ren, Jian Tang, Sijia Liu, and Yanzhi Wang. Lottery ticket preserves weight correlation: Is it desirable or not? In *Proceedings of the International Conference on Machine Learning*, 2021. 2, 3

[40] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 5

[41] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[42] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the International Conference on Machine Learning*, 2018. 1, 2

[43] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *Proceedings of the International Conference on Machine Learning*, 2020. 1, 2, 3, 4

[44] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. 2

[45] Ari Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. In *Advances in Neural Information Processing Systems*, 2019. 8

[46] Laurent Orseau, Marcus Hutter, and Omar Rivasplata. Logarithmic pruning is all you need. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3

[47] Sungho Park, Dohyung Kim, Sunhee Hwang, and Hyeran Byun. Readme: Representation learning by fairness-aware disentangling method. *arXiv preprint arXiv:2007.03775*, 2020. 1, 2

[48] Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3

[49] Sai Prasanna, Anna Rogers, and Anna Rumshisky. When bert plays the lottery, all tickets are winning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020. 3

[50] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[51] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 4, 5, 7

[52] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[53] Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. In *Proceedings of the International Conference on Learning Representations*, 2020. 2, 3

[54] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[55] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 2

[56] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the*

ACM Conference on Fairness, Accountability, and Transparency, 2019. 1

[57] Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. In *Advances in Neural Information Processing Systems*, 2021. 2

[58] Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. In *Advances in Neural Information Processing Systems*, 2022. 1

[59] Richa Singh, Puspita Majumdar, Surbhi Mittal, and Mayank Vatsa. Anatomizing bias in facial analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2

[60] Huan Tian, Tianqing Zhu, Wei Liu, and Wanlei Zhou. Image fairness in deep learning: problems, models, and challenges. *Neural Computing and Applications*, 34:12875–12893, 2022. 2

[61] Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, and Rakshit Naidu. Pruning has a disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, 2022. 1

[62] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1, 2

[63] Yulong Wang, Xiaolu Zhang, Lingxi Xie, Jun Zhou, Hang Su, Bo Zhang, and Xiaolin Hu. Pruning from scratch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 2

[64] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[65] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[66] Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *Proceedings of the International Conference on World Wide Web*, 2019. 7, 16

[67] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[68] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G Baraniuk, Zhangyang Wang, and Yingyan Lin. Drawing early-bird tickets: Towards more efficient training of deep networks. In *Proceedings of the International Conference on Learning Representations*, 2020. 3

[69] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2017. 7, 16

[70] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2018. 1, 2

[71] Yi Zhang and Jitao Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 2

[72] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2017. 1, 2

[73] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, 2019. 1, 2, 3

[74] Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. Medfair: Benchmarking fairness for medical imaging. In *Proceedings of the International Conference on Learning Representations*, 2023. 1

A. Overview and Outline

In this supplement, we provide a complement to the main content as outlined as below:

- We provide the proof for the Theorem 1 and EO version of Theorem 1 in Appendix B;
- We provide detalied experimental setup in Appendix C;
- We provide more experiments in Appendix E;

B. Proof and EO version of Theorem 1

B.1. Proof of Theorem 1

Proof. We provide the proof for fairness and accuracy, respectively.
Fairness. Notice that $\forall x, |f^*(x) - f'(x)| \leq \epsilon$. So we denote T_a, T_b, t_a, t_b as follows:

- $\sum_{i=1}^{N} \mathbb{I}_{|f^*(x_i)|\leq\epsilon, s=a} = T_a$,
- $\sum_{i=1}^{N} \mathbb{I}_{|f^*(x_i)|\leq\epsilon, s=b} = T_b$.
- $\sum\limits_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f'(x)>0} = t_a + \sum\limits_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f^*(x)>0}$,
- $\sum\limits_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f'(x)>0} = t_b + \sum\limits_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f^*(x)>0}$

So we can derive that

- $T_a + T_b = T$,
- $|t_a| \leq T_a$,
- $|t_b| \leq T_b$.

The last two inequalities are because the point x_i that satisfies $f^*(x_i)f'(x_i) < 0$ is obviously in the range $|f^*(x_i)| \leq \epsilon$ because the assumption $\forall x_i, |f^*(x_i) - f'(x_i)| \leq \epsilon$.

Therefore,

$$
\begin{aligned}
\left| \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f'(x)>0} - \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f'(x)>0} \right| =& \left| \left(t_a + \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f^*(x)>0} \right) - \left(t_b + \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f^*(x)>0} \right) \right| \\
\leq& \left| \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f^*(x)>0} - \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f^*(x)>0} \right| + |t_a - t_b| \\
\leq& \left| \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f^*(x)>0} - \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f^*(x)>0} \right| + |T_a + T_b| \\
=& N \left| \widehat{\text{DDP}}(f^*) \right| + T \\
\leq& N\delta_{f^*} + T.
\end{aligned}
$$

Finally,

$$
\left| \widehat{\text{DDP}}(f') \right| = \frac{1}{N} \left| \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a}} \mathbb{I}_{f'(x)>0} - \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b}} \mathbb{I}_{f'(x)>0} \right| \leq \delta_{f^*} + \frac{T}{N} \leq \delta_{f^*} + \delta_{f'}.
$$

Accuracy. We have

$$\text{ACC}(f^*) = \frac{1}{N} \sum_{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}} \mathbb{I}_{y=\hat{y}}.$$

Notice that for the worst case, all of the T points change their labels and are misclassified, causing an accuracy drop of $\frac{T}{N}$. So $\text{ACC}(f')$ is not worse than the worst case:

$$\text{ACC}(f') \geq \frac{1}{N}\left(\sum_{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}} \mathbb{I}_{y=\hat{y}} - T\right) = \text{ACC}(f^*) - \frac{T}{N} \geq \text{ACC}(f^*) - \delta_{f'} \geq \delta_{acc} - \delta_{f'}.$$

The proof is complete.

□

B.2. EO Version of Theorem 1

Both the theorem and the proof are similar to that of DP. Just by conditioning on $y = 1$, the proof is complete.

Theorem 2. *Given the training set $\widehat{\mathcal{D}}_{\mathcal{Z}} = \{(x_i, s_i, y_i)\}_{i=1}^N$, approximation error threshold $\epsilon > 0$, fairness tolerance $\delta_{f^*} > 0, \delta_{f'} > 0$, accuracy lower bound $\delta_{acc} > 0$. Assume that the following conditions hold:*

(A) a sufficiently large training set: $N \geq \frac{\sum_{i=1}^N \mathbb{I}_{|f^*(x_i)|\leq\epsilon}}{\delta_{f'}}$,

(B) a fair and accurate neural network f^* that satisfies $\left|\widehat{DEO}(f^*)\right| \leq \delta_{f^*}$ and $ACC(f^*) \geq \delta_{acc}$,

(C) a neural network $f' = f(\theta \odot m)$ such that $\forall x_i \in \mathcal{X}$, there holds $|f^*(x_i) - f'(x_i)| \leq \epsilon$.

Then f' is fair and accurate:

$$\begin{cases} \left|\widehat{DEO}(f')\right| \leq \delta_{f^*} + \delta_{f'}, (\textit{Fairness}) \\ ACC(f') \geq \delta_{acc} - \delta_{f'}. (\textit{Accuracy}) \end{cases}$$

Proof. **Fairness.** Notice that $\forall x, |f^*(x) - f'(x)| \leq \epsilon$. So we denote T_a, T_b, t_a, t_b as follows:

- $\sum_{i=1}^N \mathbb{I}_{|f^*(x_i)|\leq\epsilon, s=a, y=1} = T_a$,
- $\sum_{i=1}^N \mathbb{I}_{|f^*(x_i)|\leq\epsilon, s=b, y=1} = T_b$.
- $\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a \\ y=1}} \mathbb{I}_{f'(x)>0} = t_a + \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=a \\ y=1}} \mathbb{I}_{f^*(x)>0}$,
- $\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b \\ y=1}} \mathbb{I}_{f'(x)>0} = t_b + \sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}} \\ s=b \\ y=1}} \mathbb{I}_{f^*(x)>0}$

So we can derive that

- $T_a + T_b = T$,
- $|t_a| \leq T_a$,
- $|t_b| \leq T_b$.

The last two inequalities are because the point x_i that satisfies $f^*(x_i)f'(x_i) < 0$ is obviously in the range $|f^*(x_i)| \leq \epsilon$ because the assumption $\forall x_i, |f^*(x_i) - f'(x_i)| \leq \epsilon$.

Therefore,

$$\begin{aligned}\left|\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=a\\ y=1}}\mathbb{I}_{f'(x)>0}-\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=b\\ y=1}}\mathbb{I}_{f'(x)>0}\right| &= \left|\left(t_a+\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=a\\ y=1}}\mathbb{I}_{f^*(x)>0}\right)-\left(t_b+\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=b\\ y=1}}\mathbb{I}_{f^*(x)>0}\right)\right| \\ &\leq \left|\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=a\\ y=1}}\mathbb{I}_{f^*(x)>0}-\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=b\\ y=1}}\mathbb{I}_{f^*(x)>0}\right|+|t_a-t_b| \\ &\leq \left|\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=a\\ y=1}}\mathbb{I}_{f^*(x)>0}-\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=b\\ y=1}}\mathbb{I}_{f^*(x)>0}\right|+|T_a+T_b| \\ &= N\left|\widehat{\text{DEO}}(f^*)\right|+T \\ &\leq N\delta_{f^*}+T.\end{aligned}$$

Finally,

$$\left|\widehat{\text{DEO}}(f')\right| = \frac{1}{N}\left|\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=a\\ y=1}}\mathbb{I}_{f'(x)>0}-\sum_{\substack{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}\\ s=b\\ y=1}}\mathbb{I}_{f'(x)>0}\right| \leq \delta_{f^*}+\frac{T}{N}\leq\delta_{f^*}+\delta_{f'}.$$

Accuracy. We have

$$\text{ACC}(f^*)=\frac{1}{N}\sum_{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}}\mathbb{I}_{y=\hat{y}}.$$

Notice that for the worst case, all of the T points change their labels and are misclassified, causing an accuracy drop of $\frac{T}{N}$. So $\text{ACC}(f')$ is not worse than the worst case:

$$\text{ACC}(f')\geq\frac{1}{N}\left(\sum_{(x,s,y)\sim\widehat{\mathcal{D}}_{\mathcal{Z}}}\mathbb{I}_{y=\hat{y}}-T\right)=\text{ACC}(f^*)-\frac{T}{N}\geq\text{ACC}(f^*)-\delta_{f'}\geq\delta_{acc}-\delta_{f'}.$$

The proof is complete. □

C. Detailed Experiment Setup

C.1. Datasets

We conduct experiments on two real-world face image datasets, *i.e.*, CelebA and LFW. The CelebA dataset consists of 202,599 images along with 40 annotated binary attributes per image, and LFW dataset consists of 13,244 images along with 73 annotated binary attributes per image. We adopt *gender* as the sensitive attribute. We use *Smiling* and *Blond Hair* as the target labels on CelebA , and we take *Smiling* and *Wavy Hair* as the target labels on LFW. We split each dataset into training set, validation set and test set. We use the torchvision, a library of Pytorch for computer vision to split the original dataset of CelebA into training set, validation set and test set. We randomly divide the original dataset of LFW into training set with 6,000 images, validation set with 3,600 images and test set with the remaining images. All the images are first resized to 256 × 256, and then center cropped to 224 × 224.

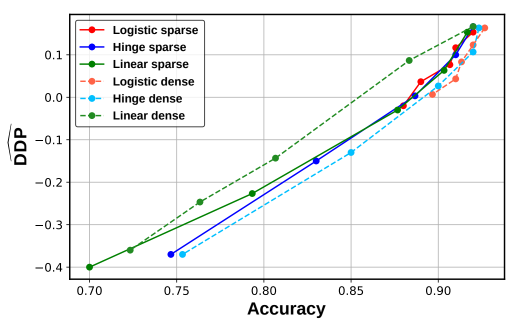
We find that, under fairness-aware adversarial training, when using the $Smiling$ targets on both CelebA and LFW, the model training suffers from model collapses. Thus, we only evaluate our FST search method on CelebA with *Blond Hair* targets and LFW with *Wary Hair* targets. Moreover, we find that employing the all training set under fairness-aware adversarial training on CelebA leads to model collapse. Thus, under fairness-aware adversarial training on CelebA, we only use the 10% images of CelebA training set, and the validation set and test set remain unchanged. **Although we have to adopt some special settings for fairness-aware adversarial training due to overcoming model collapses, we believe that our experiments for adversarial training is enough to prove the generality of our FST search method under fairness-aware adversarial training. In addition, we would like to emphasize that, the model collapses occur on both the fair dense networks trained with existing fairness-aware in-processing methods and our FST methods, which to some extent can also be considered comparable.**

Dataset	Method	Optimizer	Epochs	Learning Rate
CelebA	Regularization	SGD	3	0.01
CelebA	Adversarial	Adam	10	0.01
LFW	Regularization	Adam	10	0.0005
LFW	Adversarial	Adam	10	0.01

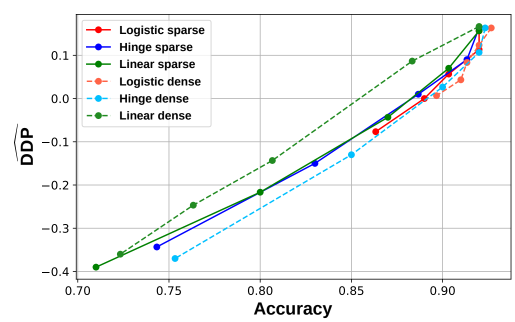
Table 1. Optimizers, Epochs and Learning Rates for Datasets and Methods

C.2. Implementation details

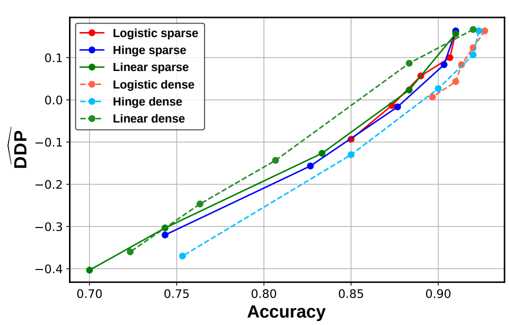
We implement all experiments by Pytorch. We use ResNet18 as the network architecture under fairness regularization. As for fairness-aware adversarial training, we use ResNet18 as the shared representation encoder, a fully connected layers with dimensions of 512-512-1 and ReLU activate function as the target prediction head, a fully connected layers with dimensions of 256-64-1 and LeakyReLU (negative slope = 0.1) activation function as the target prediction head as the adversarial head. In Tab. 1, we show the selection of optimizer, epochs and learning rate when specifying the dataset and method. The policy of learning rate decay is set to cosine annealing, and the mini-batch size is set to 128 except the experiments under R_{deo}



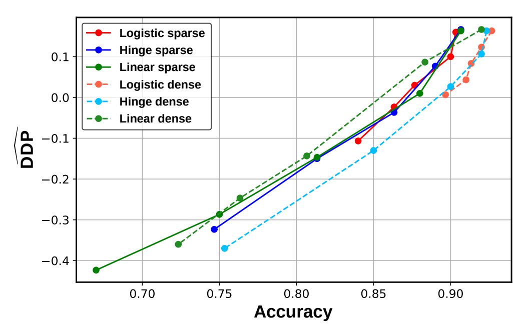
(a) $\eta = 5\%$



(b) $\eta = 10\%$



(c) $\eta = 40\%$



(d) $\eta = 80\%$

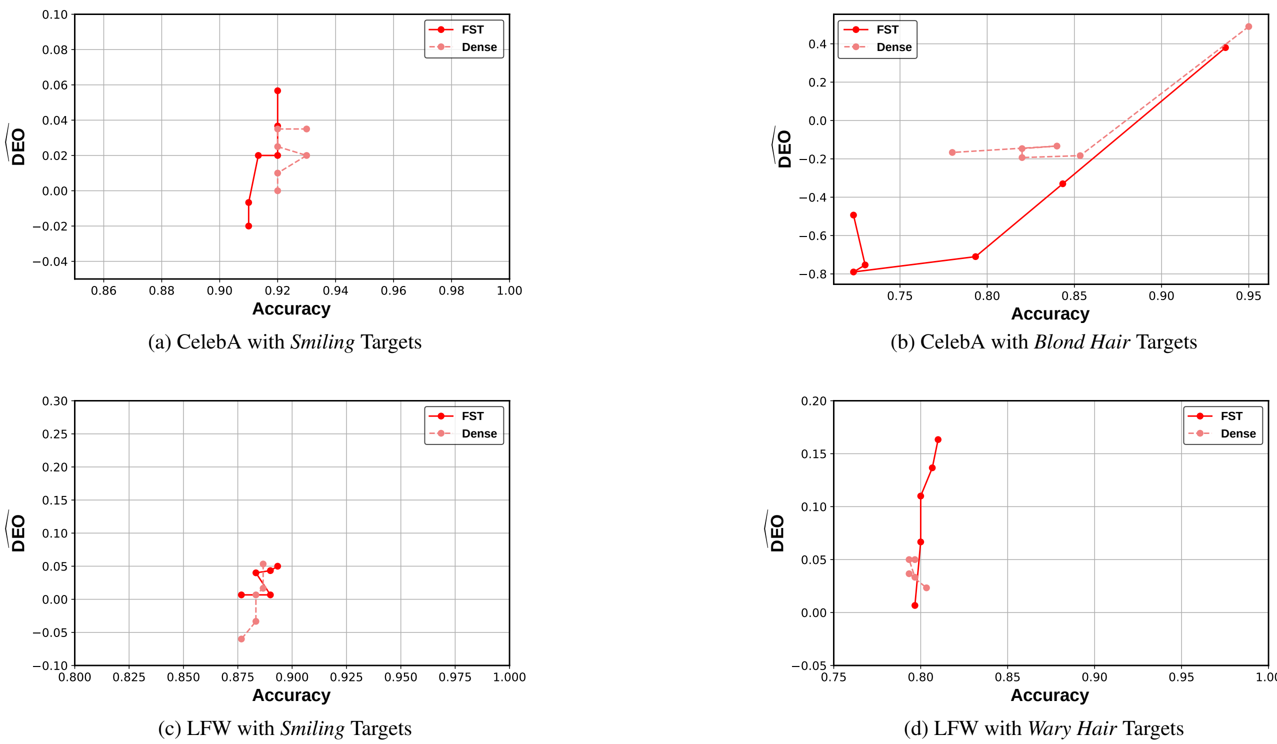
Figure 9. FSTs exist under R_{ddp} regularization with different sparsity patterns on CelebA with *Smiling* targets.

regularization on CelebA with *Blond Hair* targets is set to 512. For experiments whose optimizer is SGD, we use momentum of 0.9 and weight decay of 0.0001. For experiments whose optimizer is Adam, we use betas of 0.9 and 0.999 and weight decay of 0.0001. We train network with training set, select the network weights with the best accuracy in validation set, and report the accuracy and unfairness in test set. The reported results are the average of three trials with different random seeds.

D. FSTs Exist under Different Fairness Surrogates

In Fig. 9, we show the accuracy-fairness trade-off of FSTs under different fairness surrogates $u(\cdot)$. We consider three kinds of surrogates: linear surrogate [4, 16, 69], hinge surrogate [66], and logistic surrogate [5]. We can find that the FSTs exist under different fairness surrogates. The best surrogate is the logistic surrogate, which is consistent with [5]. An interesting fact is that FSTs with linear surrogate outperform the dense counterparts trained with linear surrogate, which is different from other fairness surrogates.

E. More Experiments



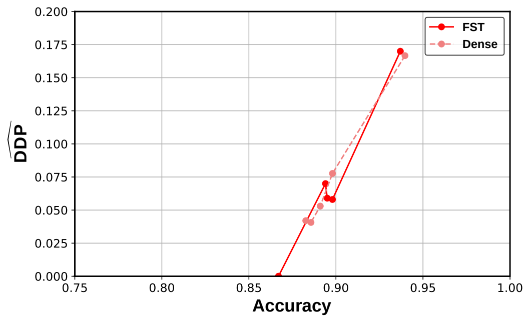
(a) CelebA with *Smiling* Targets

(b) CelebA with *Blond Hair* Targets

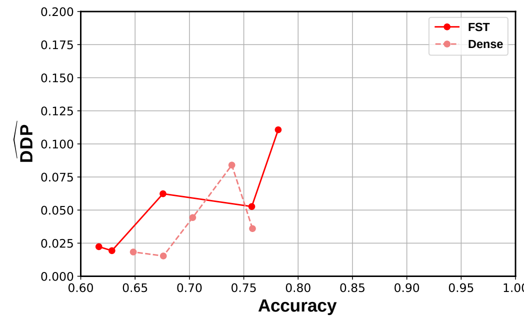
(c) LFW with *Smiling* Targets

(d) LFW with *Wary Hair* Targets

Figure 10. FSTs exist under R_{deo} regularization on CelebA and LFW datasets with remaining ratio $\eta = 10\%$.

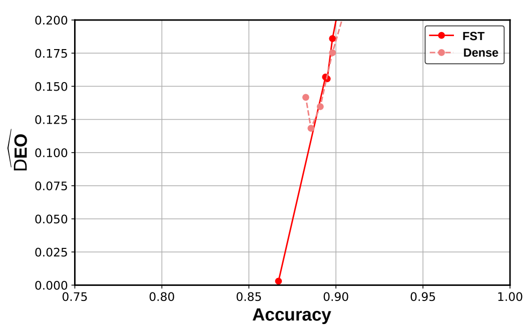


(a) LFW with *Blond Hair* targets

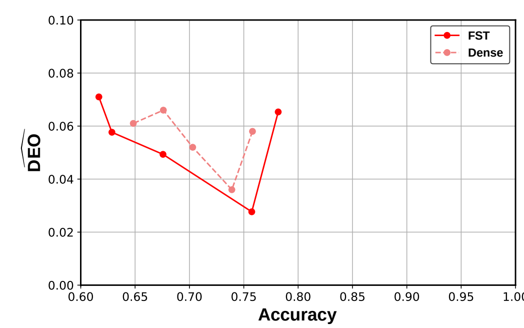


(b) LFW with *Wary Hair* targets

Figure 11. FSTs exist under fairness-aware advesarial training on CelebA and LFW datasets with remaining ratio $\eta = 10\%$ ($\widehat{\text{DDP}}$ metric).

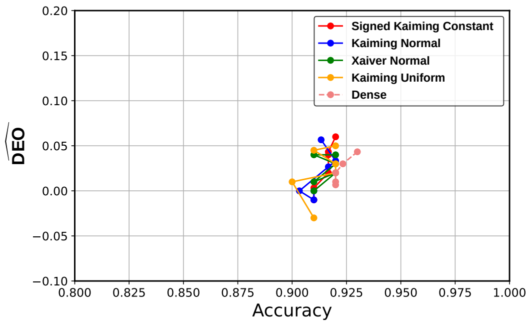


(a) LFW with *Blond Hair* targets

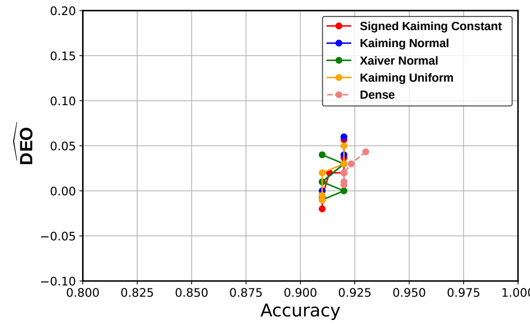


(b) LFW with *Wary Hair* targets

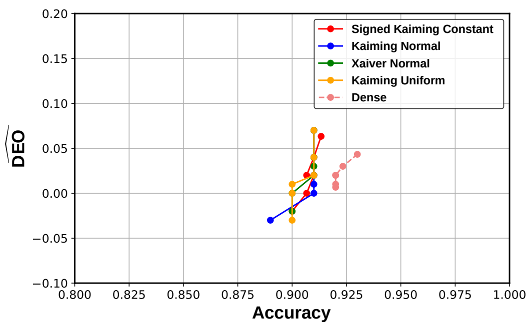
Figure 12. FSTs exist under advesarial training on CelebA and LFW datasets with remaining ratio $\eta = 10\%$ ($\widehat{\text{DEO}}$ metric).



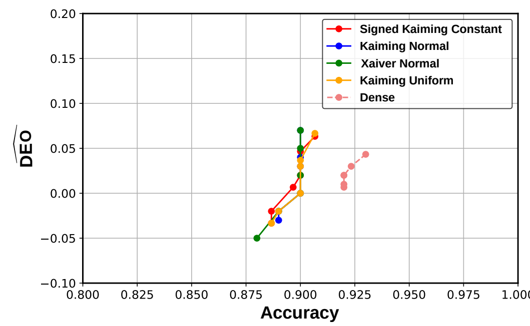
(a) $\eta = 5\%$



(b) $\eta = 10\%$

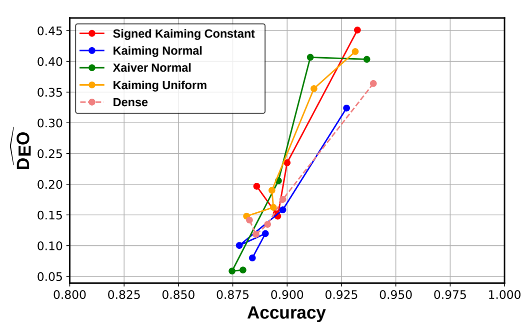


(c) $\eta = 40\%$

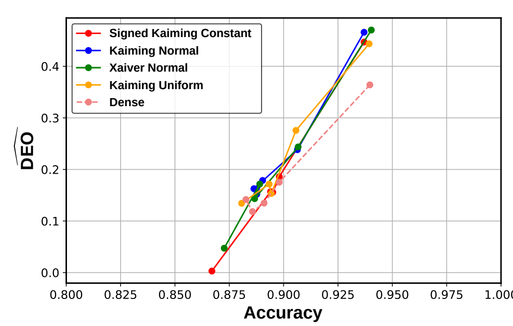


(d) $\eta = 80\%$

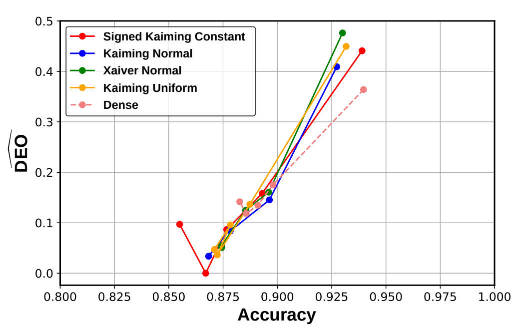
Figure 13. FSTs exist under R_{deo} regularization with four initialization methods on CelebA with *Smiling* targets.



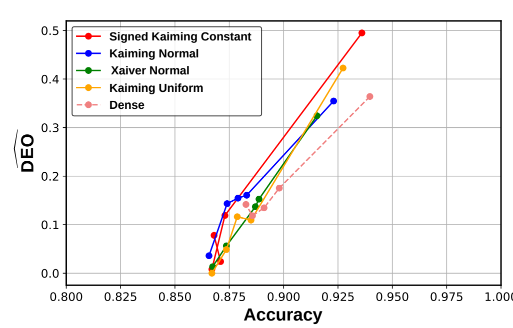
(a) $\eta = 5\%$



(b) $\eta = 10\%$

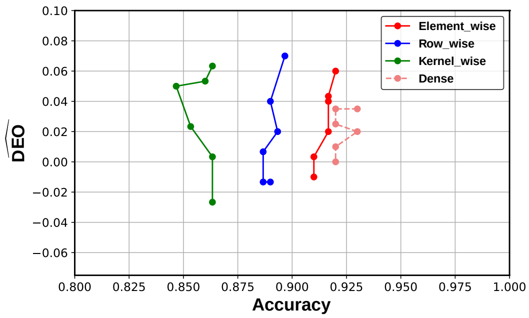


(c) $\eta = 40\%$

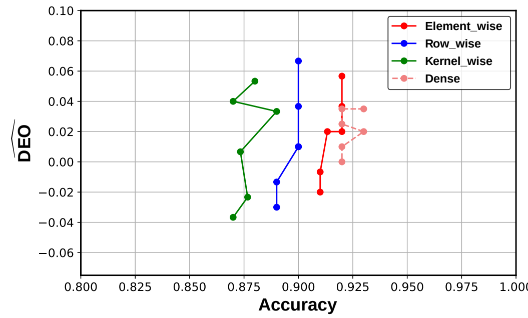


(d) $\eta = 80\%$

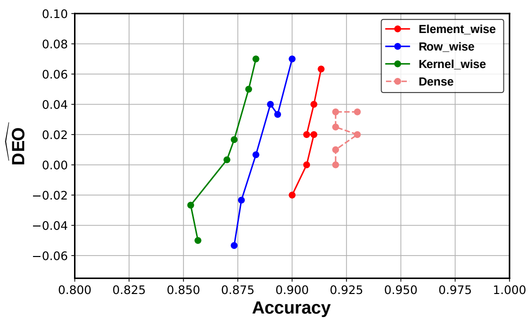
Figure 14. FSTs exist under adversarial training with four initialization methods on CelebA with *Blond Hair* targets ($\widehat{\text{DEO}}$ metric) .



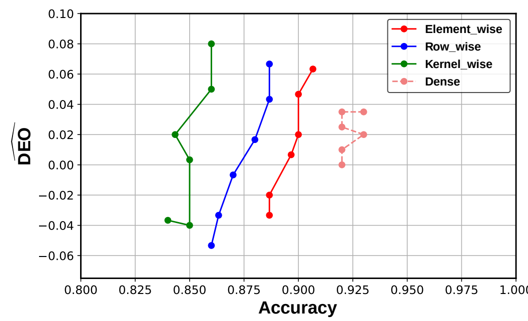
(a) $\eta = 5\%$



(b) $\eta = 10\%$

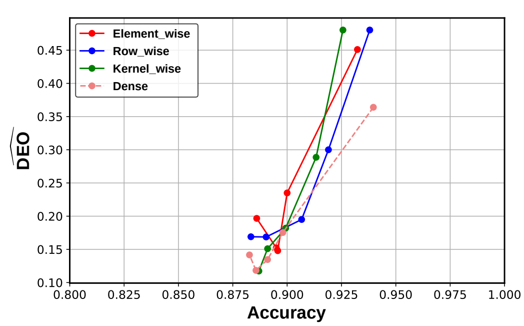


(c) $\eta = 40\%$

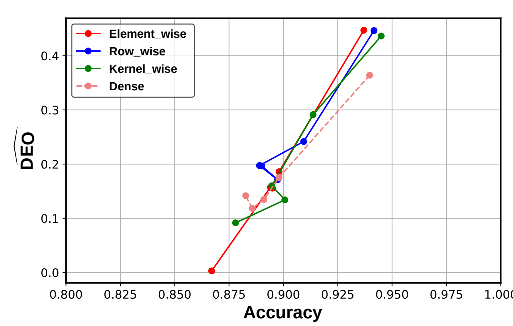


(d) $\eta = 80\%$

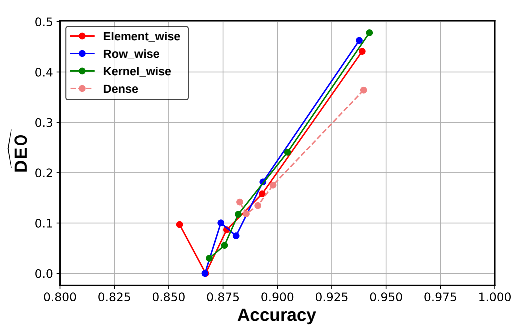
Figure 15. FSTs exist under R_{deo} regularization with different sparsity patterns on CelebA with *Smiling* targets.



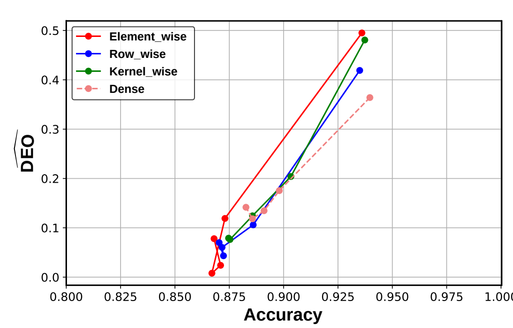
(a) $\eta = 5\%$



(b) $\eta = 10\%$

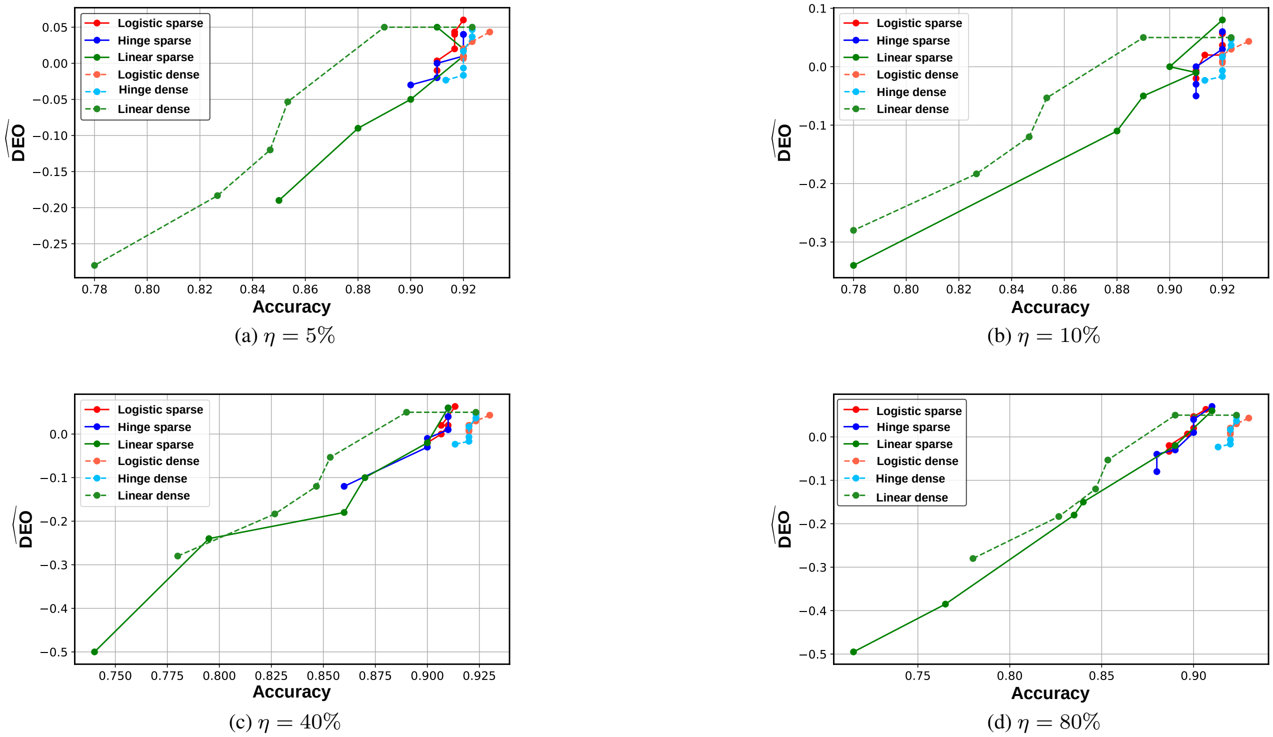


(c) $\eta = 40\%$



(d) $\eta = 80\%$

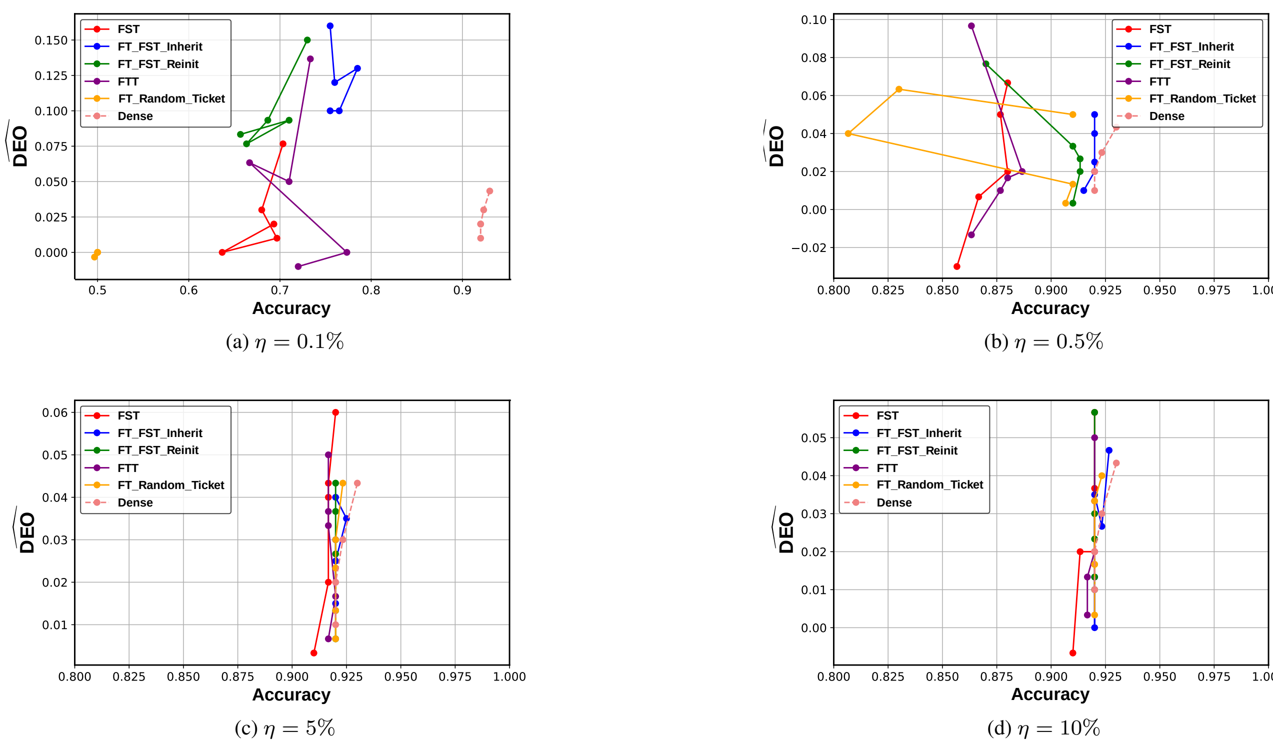
Figure 16. FSTs exist under adversarial training with different sparsity patterns on CelebA with *Blond Hair* targets ($\widehat{\text{DEO}}$ metric) .



(a) $\eta = 5\%$ (b) $\eta = 10\%$

(c) $\eta = 40\%$ (d) $\eta = 80\%$

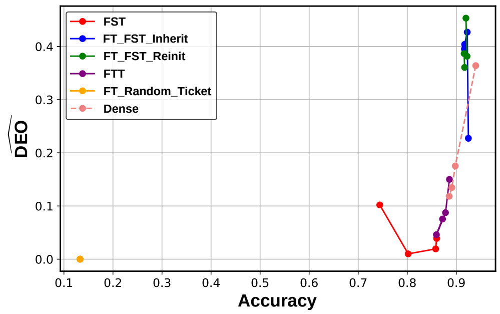
Figure 17. FSTs exist under R_{deo} regularization with different fairness surrogates on CelebA with *Smiling* targets.



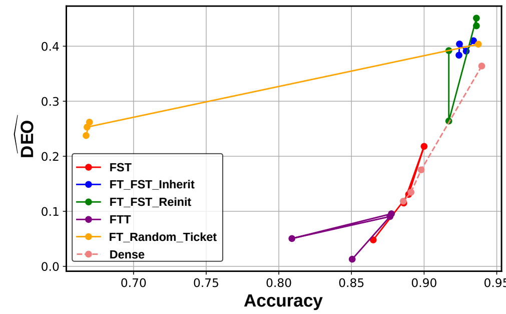
(a) $\eta = 0.1\%$ (b) $\eta = 0.5\%$

(c) $\eta = 5\%$ (d) $\eta = 10\%$

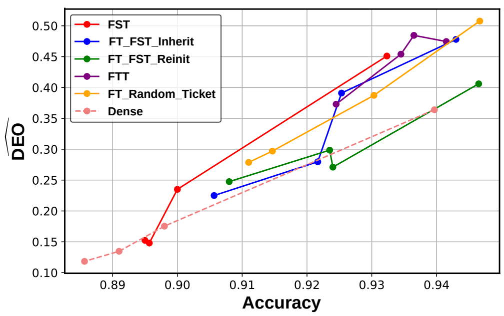
Figure 18. Comparisions of FST variants under R_{deo} regularization on CelebA with *Smiling* targets.



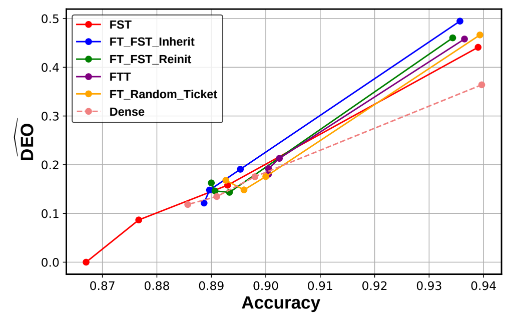
(a) $\eta = 0.1\%$



(b) $\eta = 0.5\%$

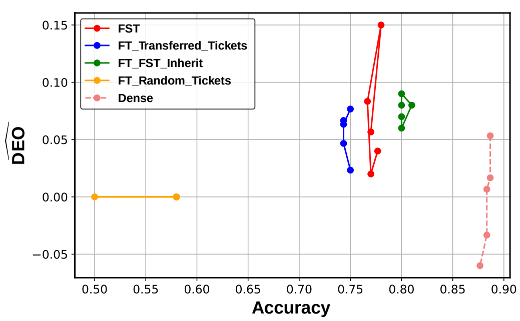


(c) $\eta = 5\%$

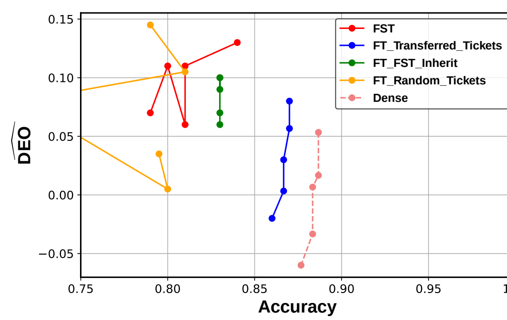


(d) $\eta = 40\%$

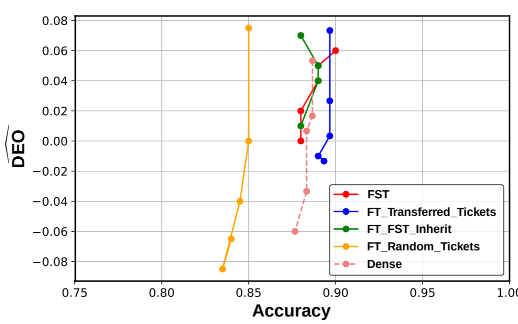
Figure 19. Comparisions of FST variants under adversarial training on CelebA with *Smiling* targets ($\widehat{\text{DEO}}$ metric) .



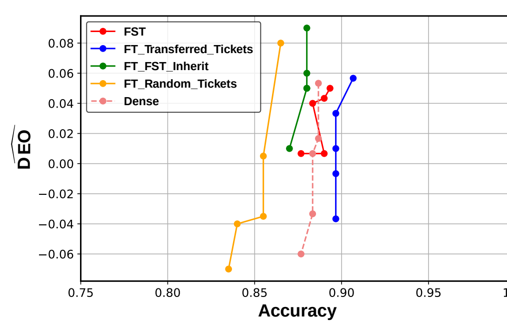
(a) $\eta = 0.1\%$



(b) $\eta = 0.5\%$



(c) $\eta = 5\%$



(d) $\eta = 10\%$

Figure 20. Comparisons between fine-tuned transferred FSTs and other methods under R_{deo} on LFW with *Smiling* targets.