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 how FSTs exist under different fairness regularization surrogates in Appendix D;
- 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,

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

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:

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

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:

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

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 $\times$ 256, and then center cropped to 224 $\times$ 224.

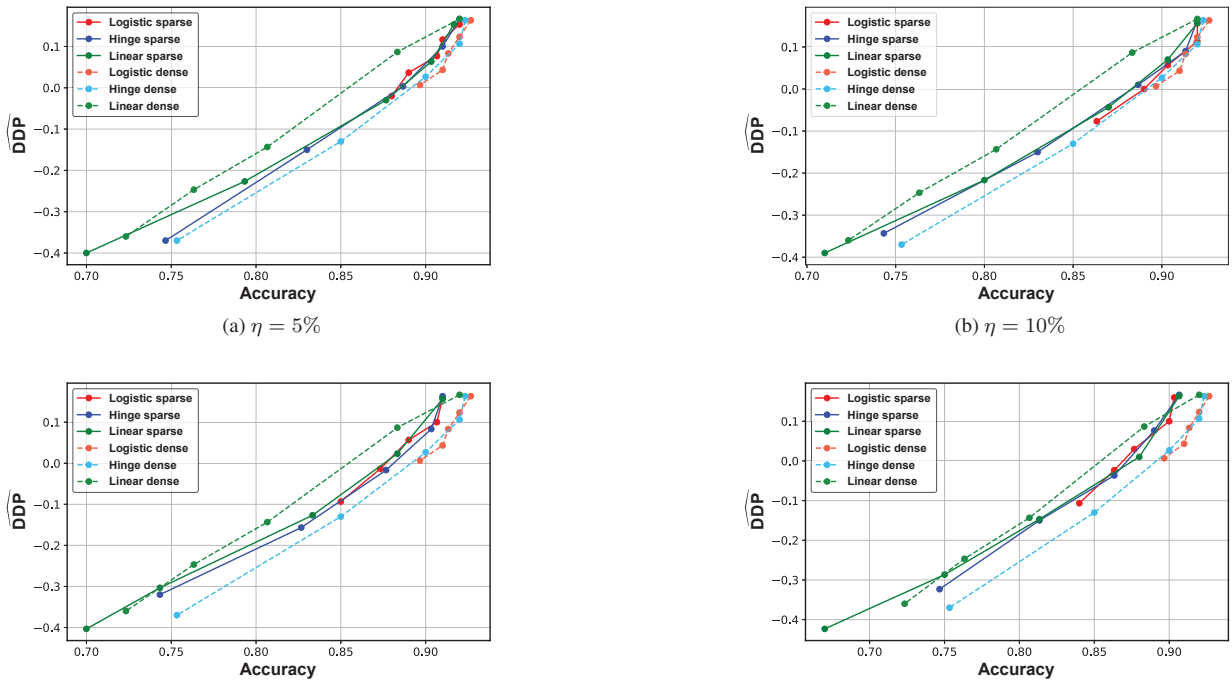
We find that, under fairness-aware adversarial training, when using the $Smiling$ targets on both CelebA and LFW, the model training suffers from model collapses. Thus, we only evaluate our FST search method on CelebA with *Blond Hair* targets and LFW with *Wary Hair* targets. Moreover, we find that 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.



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

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

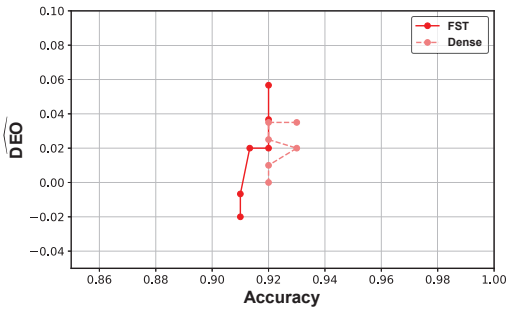
Figure 9. FSTs exist under R_{ddp} regularization with different sparsity patterns on CelebA with *Smiling* targets.

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} 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.

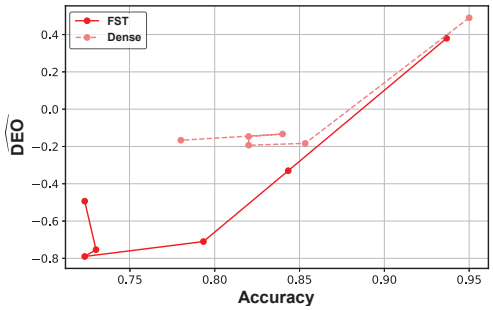
D. FSTs Exist under Different Fairness Surrogates

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, 71], hinge surrogate [68], 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 finding is that FSTs with linear surrogate outperform the dense counterparts trained with linear surrogate, which is different from other fairness surrogates.

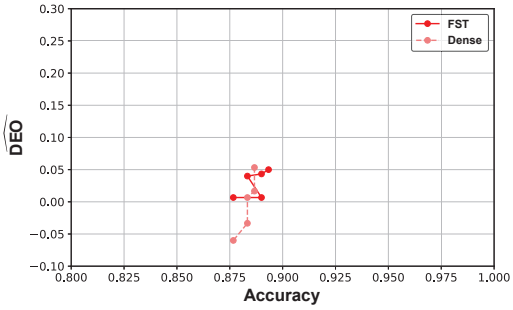
E. More Experiments



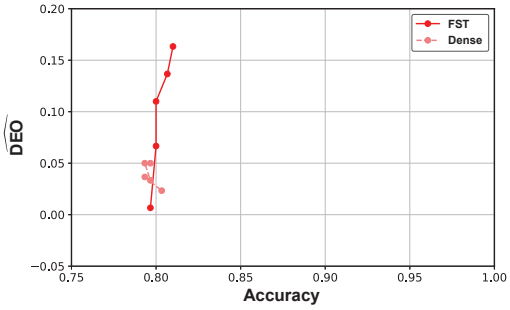
(a) CelebA with *Smiling* Targets



(b) CelebA with *Blond Hair* Targets

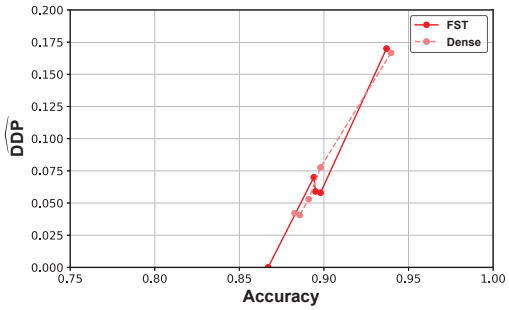


(c) LFW with *Smiling* Targets

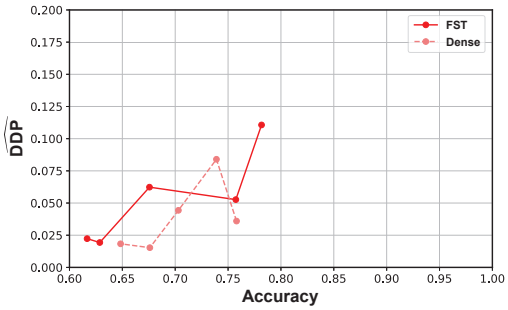


(d) LFW with *Wary Hair* Targets

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

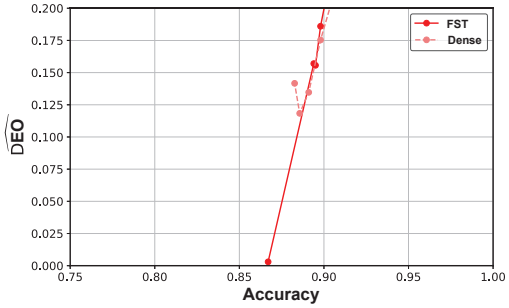


(a) LFW with *Blond Hair* targets

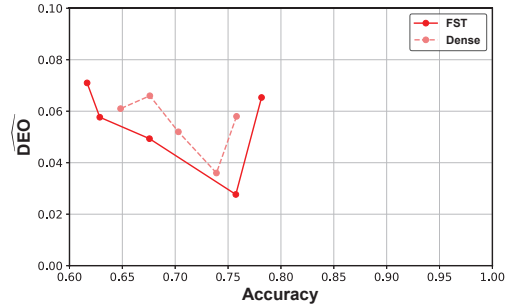


(b) LFW with *Wary Hair* targets

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

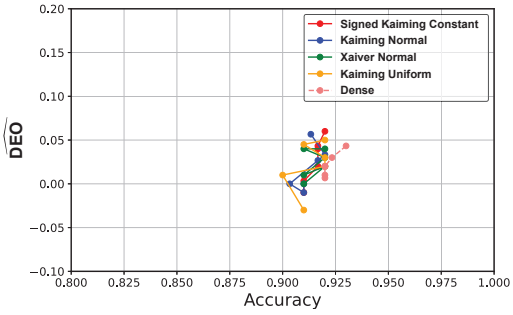


(a) LFW with *Blond Hair* targets

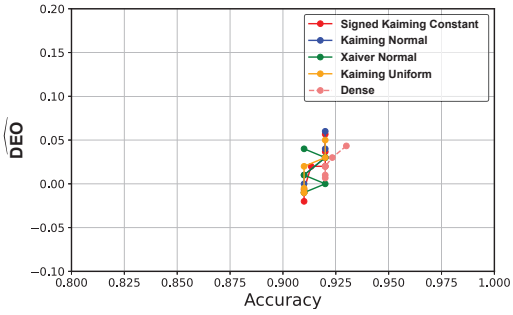


(b) LFW with *Wary Hair* targets

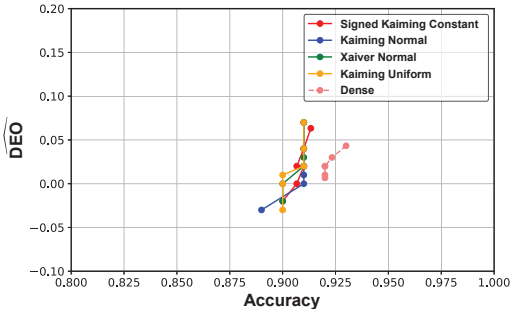
Figure 12. FSTs exist under advesarial training on CelebA and LFW datasets with remaining ratio $\eta = 10\%$ ($\widehat{\text{DEO}}$ metric).



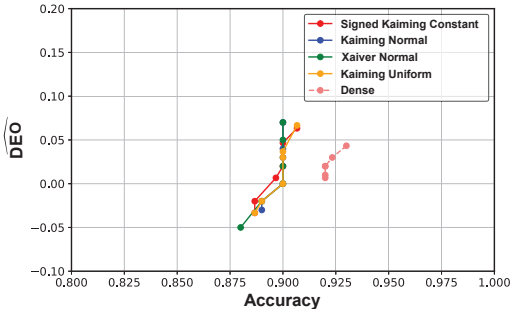
(a) $\eta = 5\%$



(b) $\eta = 10\%$

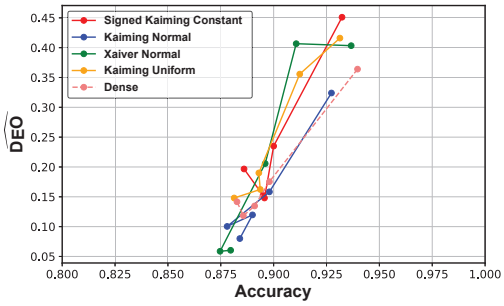


(c) $\eta = 40\%$

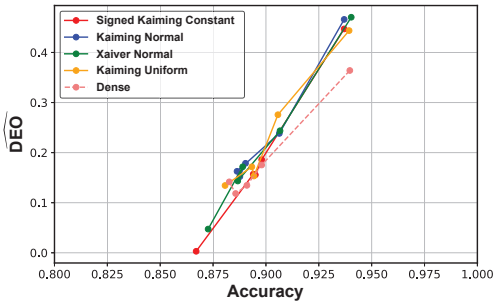


(d) $\eta = 80\%$

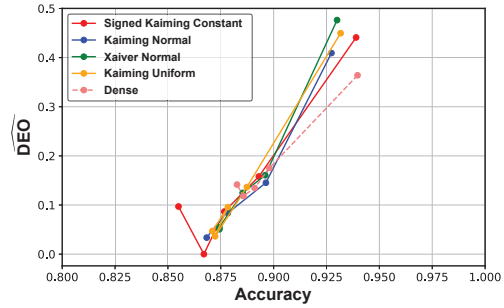
Figure 13. FSTs exist under R_{deo} regularization with four initialization methods on CelebA with *Smiling* targets.



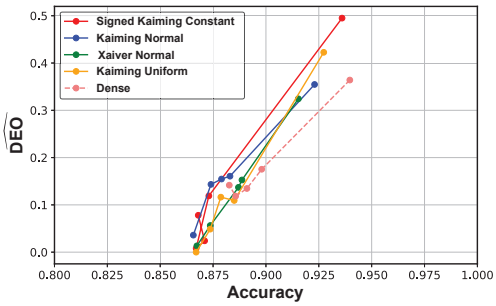
(a) $\eta = 5\%$



(b) $\eta = 10\%$

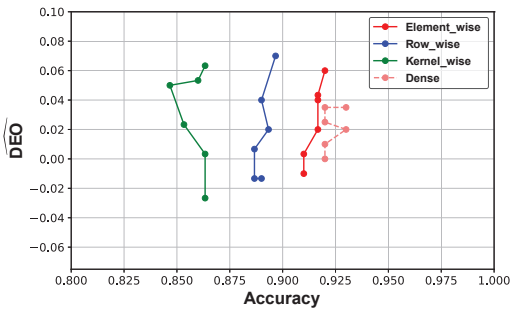


(c) $\eta = 40\%$

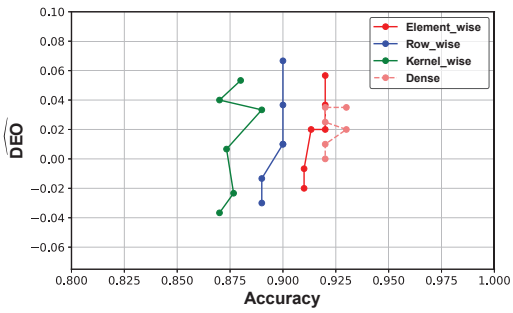


(d) $\eta = 80\%$

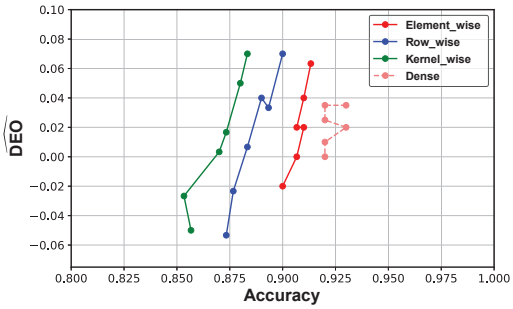
Figure 14. FSTs exist under adversarial training with four initialization methods on CelebA with *Blond Hair* targets ($\widehat{\text{DEO}}$ metric) .



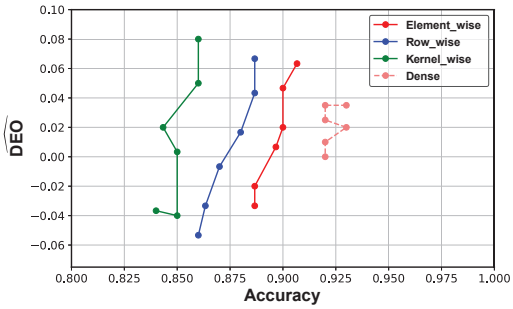
(a) $\eta = 5\%$



(b) $\eta = 10\%$

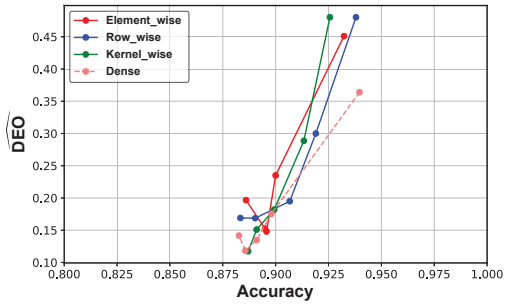


(c) $\eta = 40\%$

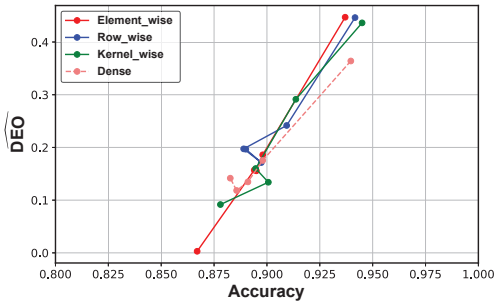


(d) $\eta = 80\%$

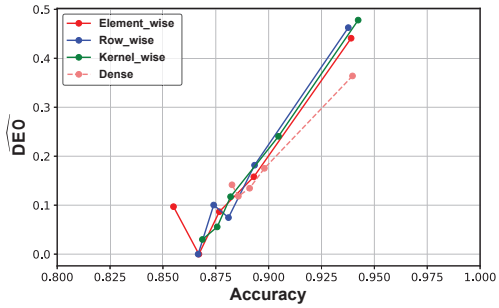
Figure 15. FSTs exist under R_{deo} regularization with different sparsity patterns on CelebA with *Smiling* targets.



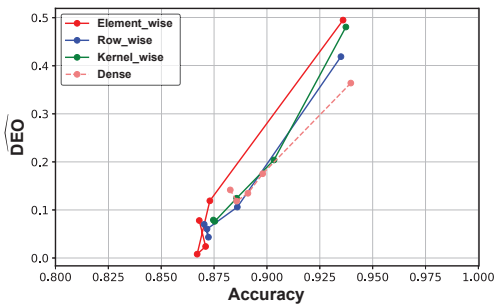
(a) $\eta = 5\%$



(b) $\eta = 10\%$

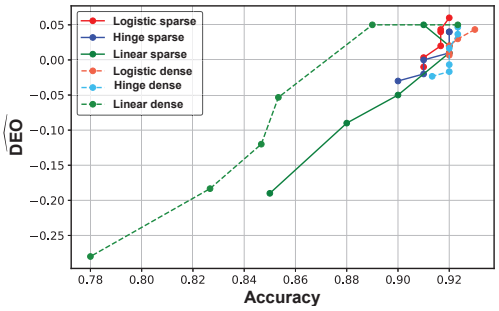


(c) $\eta = 40\%$

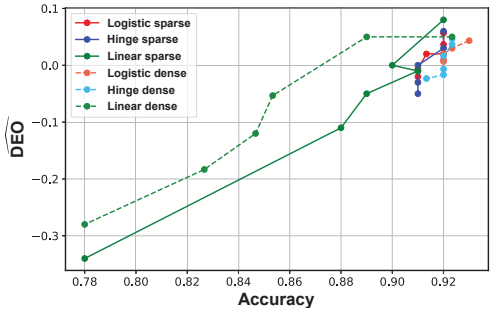


(d) $\eta = 80\%$

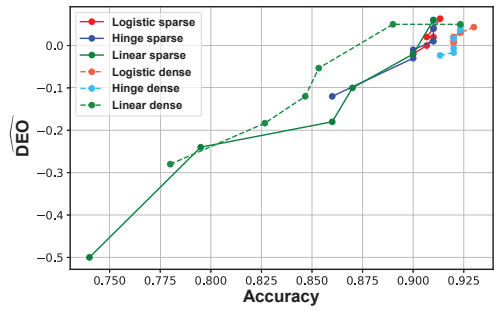
Figure 16. FSTs exist under adversarial training with different sparsity patterns on CelebA with *Blond Hair* targets ($\widehat{\text{DEO}}$ metric) .



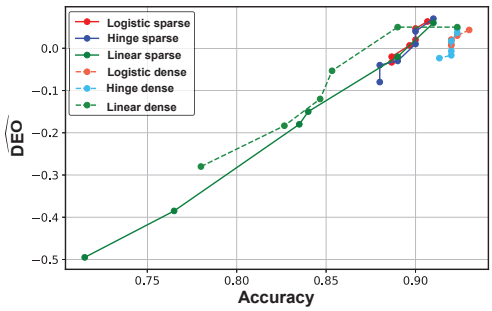
(a) $\eta = 5\%$



(b) $\eta = 10\%$

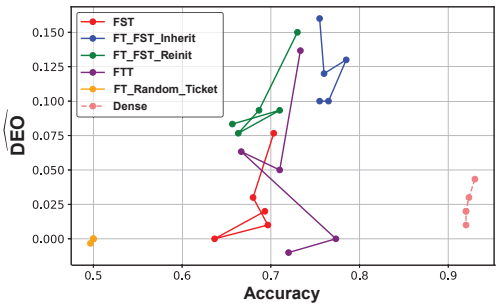


(c) $\eta = 40\%$

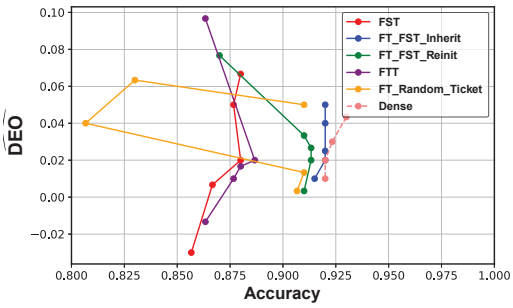


(d) $\eta = 80\%$

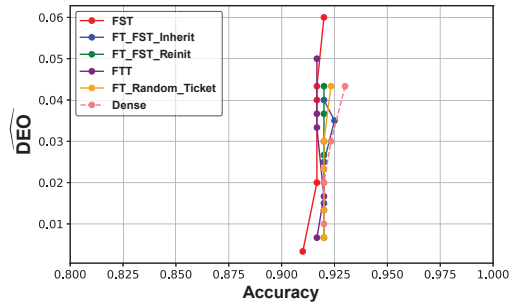
Figure 17. FSTs exist under R_{deo} regularization with different fairness surrogates on CelebA with *Smiling* targets.



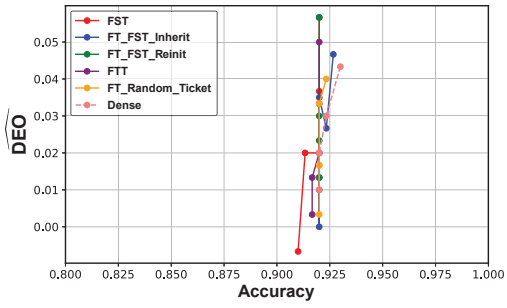
(a) $\eta = 0.1\%$



(b) $\eta = 0.5\%$

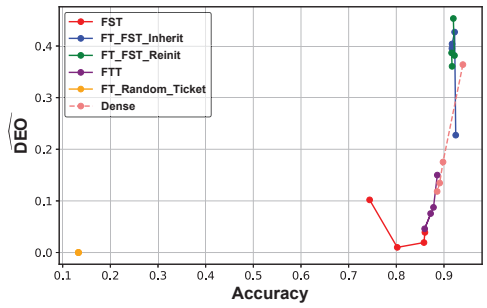


(c) $\eta = 5\%$

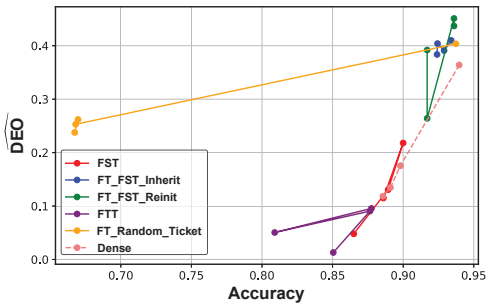


(d) $\eta = 10\%$

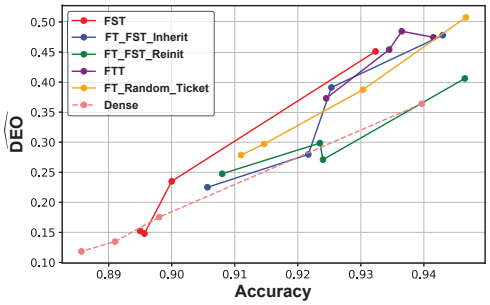
Figure 18. Comparisions of FST variants under R_{deo} regularization on CelebA with *Smiling* targets.



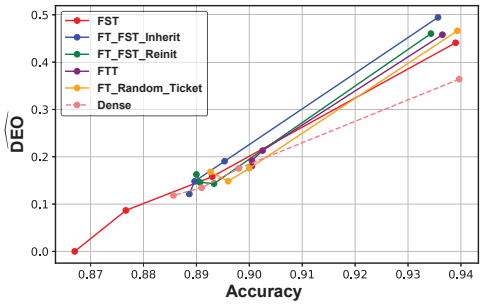
(a) $\eta = 0.1\%$



(b) $\eta = 0.5\%$

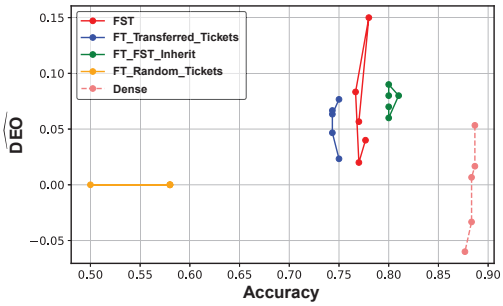


(c) $\eta = 5\%$

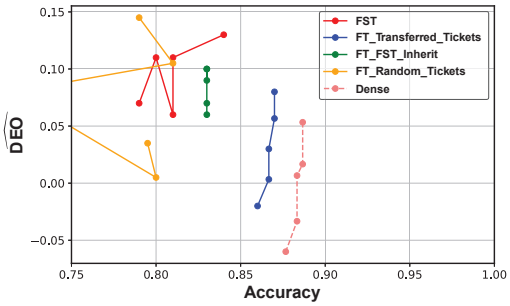


(d) $\eta = 40\%$

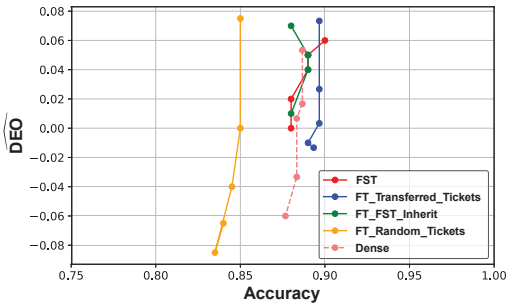
Figure 19. Comparisions of FST variants under adversarial training on CelebA with *Smiling* targets ($\widehat{\text{DEO}}$ metric) .



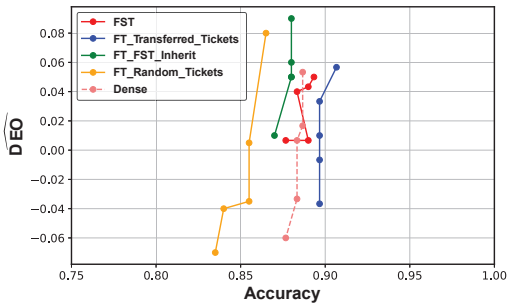
(a) $\eta = 0.1\%$



(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.