Supplementary Material
Fair Domain Generalization with Heterogeneous Sensitive Attributes across Domains

This document consists of the following subsections:

Contents

1. Further Implementation Details

1.1. Hyperparameter Tuning and Analysis:

Our model utilizes 5 hyperparameters. Tab. 1 presents the range of hyperparameter values we tested and our final choice. γ determines the weight of the selective invariance loss, $\mathcal{L}_{DF}$. Generally, we observed better predictive performance when γ is higher and fairness when γ is lower. ϵ controls the distance between fairness representations of an attribute in $\mathcal{S}$ when one is specified as sensitive and the other is not. Generally, we noticed better predictive performance when $\epsilon \geq 1$ and better fairness when $\epsilon < 1$. $|\mathbf{C}|$ is the size of the random subset of $\mathcal{C}$ sampled at each iteration. We observed that the model has the best predictive performance when $|\mathbf{C}| = |\mathcal{C}|$. On average, we observed that $|\mathbf{C}| = |\mathcal{C}|/3$ samples were enough to reach 95% of the best predictive performance. α determines the weight of the domain invariance loss $\mathcal{L}_{DG}$ and ω determines the weight of the fairness loss $\mathcal{L}_{EO}$. We chose $\alpha = 0.1$ and $\omega = 1$ based on the best validation AUROC and MD respectively. While tuning each hyperparameter, we fixed the value of other hyperparameters. We also tuned α and ω values for the respective baselines. *More analysis on each hyperparameter is provided in Appendix Sec. 3.3 in Tabs. 4 to 8.*

Table 1. Hyperparameter Choices

Parameter	Used with	Range Tested	Final Choice
ϵ	$\mathcal{L}_{DF}$	$\{0.01, 0.1, 1, 10\}$	1
γ	$\mathcal{L}_{DF}$	$\{0.01, 0.1, 1, 10\}$	1
$\lvert\mathbf{C}\rvert$	$\mathcal{L}_{DF}$	$\{\lvert\mathcal{C}\rvert, \lvert\mathcal{C}\rvert/2, \lvert\mathcal{C}\rvert/3\}$	$\lvert\mathcal{C}\rvert/2$ (MIMIC) $\lvert\mathcal{C}\rvert$ (CelebA) $\lvert\mathcal{C}\rvert/3$ (FACET)
α	$\mathcal{L}_{DG}$	$\{0.01, 0.1, 1, 10\}$	0.1
ω	$\mathcal{L}_{EO}$	$\{0.1, 1, 5\}$ $\{0.01, 0.1, 1\}$	1 (MIMIC,CelebA) 0.01 (FACET)

1.2. Algorithm

Please refer to the Algorithm 1.

1.3. Dimensions of the Encoded Representations

The total dimension of the encoded representation ($\mathbf{z}$) of SISA is split between the generalization representation ($\mathbf{z}_g$) and the fairness representation ($\mathbf{z}_f$). When the sensitive attributes are on the lower side (for the MIMIC dataset when $n = 2$ or 3), the dimension of $\mathbf{z}$ is proportionately split between $\mathbf{z}_g$ and $\mathbf{z}_f$. However, when n was increased to 4, due to the drop in accuracy, we split the dimension in the ratio $3:1$ such that the dimension of $\mathbf{z}_g$ was 3 times that of $\mathbf{z}_f$. Dimension of $\mathbf{z}_f$ is equally split between $\mathbf{z}_{f_i}$'s for all datasets. We report the dimensions used for the represen-

Algorithm 1 **S**elective **I**nvariance under **S**ensitive **A**ttributes

Require: Training data: $\mathcal{T}$, sensitive attribute set: $\mathcal{S}$, set of sensitive attribute encodings: $\mathcal{C}$, density translators: G'', G', batch size: B

Initialize θ, ϕ, and ψ (parameters of g_θ, f_ϕ and, h_ψ).
for epoch in MAX_EPOCHS **do**
  for each domain $d \in \mathcal{D}$ **do**
    Sample a batch $\{\mathbf{x}^k, y^k, \mathbf{s}^k\}_{k=1}^B \sim \mathrm{P}_d$ from $\mathcal{T}$
    for each $k \in (1, B)$ **do**
      $\mathbf{z}_g^k \leftarrow g_\theta(\mathbf{x}^k)$ *# Generalization representation*
      $d' \in \mathcal{D}$
      $\mathbf{x}'^k \leftarrow G(\mathbf{x}^k, d, d')$ *# Domain translated* $\mathbf{x}$
      $\mathbf{z}_g'^k \leftarrow g_\theta(\mathbf{x}'^k)$ *# Domain translated rep.*
      for $\mathbf{c} \in \mathbf{C} \subseteq \mathcal{C}$ **do**
        $[\mathbf{z}_{f_1}^k, \dots, \mathbf{z}_{f_n}^k] \leftarrow f_\phi(\mathbf{x}^k \oplus embed(\mathbf{c}))$ *# Fairness rep.*
        $\mathbf{x}'^{sk} \leftarrow G^s(\mathbf{x}^k, d, d')$
        $\mathbf{c}' = permute(\mathbf{c})$ *# Sample another encoding*
        $[\mathbf{z}_{f_1}'^{sk}, \dots, \mathbf{z}_{f_n}'^{sk}] \leftarrow f_\phi(\mathbf{x}'^{sk} \oplus embed(\mathbf{c}'))$
        $\mathbf{z}^k \leftarrow \mathbf{z}_g^k \oplus \mathbf{z}_{f_1}^k \oplus \dots \oplus \mathbf{z}_{f_n}^k$
      end for *# Repeat for all* $\mathbf{c} \in \mathbf{C}$
    end for
    $\mathcal{L}_{DG} \leftarrow \frac{1}{k} \sum_k \parallel \mathbf{z}_g^k - \mathbf{z}_g'^k \parallel_2$ *# Domain invariant loss*
    for each $i \in (1, n)$ **do**
      $\mathcal{L}_{DF} \leftarrow \frac{1}{k} \sum_k \mathbb{1}_{[\mathbf{c}[i]=\mathbf{c}'[i]]} \parallel \mathbf{z}_{f_i}^k - \mathbf{z}_{f_i}'^{\mathbf{s}k} \parallel_2 + \mathbb{1}_{[\mathbf{c}[i]\neq\mathbf{c}'[i]]} \max(0, \epsilon - \parallel \mathbf{z}_{f_i}^k - \mathbf{z}_{f_i}'^{\mathbf{s}k} \parallel_2)$
    end for *# Selective dom. inv. loss*
    $\mathcal{L}_{ER} \leftarrow \frac{1}{k} \sum_k l(h_\psi(\mathbf{z}^k), y^k)$ *# Classification loss*
    $\mathcal{L}_{EO} \leftarrow \frac{1}{k} \sum_k \frac{1}{y} \sum_y \frac{1}{|\mathbf{C}|} \sum_{\mathbf{c}\in\mathbf{C}} \frac{1}{\binom{|\mathcal{I}_\mathbf{c}|}{2}} \sum_{(\mathbf{i},\mathbf{j})\in\mathcal{I}_\mathbf{c}} \mid h_\psi(\mathbf{z}^k \mid y, \mathbf{i}) - h_\psi(\mathbf{z}^k \mid y, \mathbf{j}) \mid$ *# Fairness loss*
    $\mathcal{L}_{final} \leftarrow \mathcal{L}_{ER} + \omega\, \mathcal{L}_{EO} + \alpha\, \mathcal{L}_{DG} + \gamma\, \mathcal{L}_{DF}$
    $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}_{final}$ *# Gradient Descent*
    $\phi \leftarrow \phi - \nabla_\phi \mathcal{L}_{final}$
    $\psi \leftarrow \psi - \nabla_\psi \mathcal{L}_{final}$
    Optimize θ, ϕ, and ψ based on $\mathcal{L}_{final}$ via gradient descent
  end for
end for
return Trained θ, ϕ, ψ

Table 2. Dimension of $\mathbf{z}$

Model	CelebA ($n=4$) $\mathbf{z}$		Cardio./Pneu. ($n=2$) $\mathbf{z}$		Edema ($n=3$) $\mathbf{z}$		FACET ($n=3$)	
ERM	1024		1024		1280		1024	
DIRT	1024		1024		1280		1024	
FATDM	1024		1024		1024		1024	
	1024		1024		1280		1024	
SISA	768 ($\mathbf{z}_g$)	64 ($\mathbf{z}_{f_i}$)	512 ($\mathbf{z}_g$)	256 ($\mathbf{z}_{f_i}$)	512 ($\mathbf{z}_g$)	256 ($\mathbf{z}_{f_i}$)	640 ($\mathbf{z}_g$)	128 ($\mathbf{z}_{f_i}$)

tation $\mathbf{z}$ in Tab. 2. We consider the same dimensions for $\mathbf{z}$ across the baselines (ERM, ERM-F, DIRT, and FATDM) and our model for valid comparison.

1.4. Training Procedure of G and G^s

Models G and G^s are generators of a StarGAN [1] $G\colon \mathbb{R}^{w\times h\times c} \times \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}^{w\times h\times c}$. The GAN also contains a discriminator $D : \mathbb{R}^{w\times h\times c} \rightarrow \mathbb{N} \times 0, 1$. The generator takes in a real image $\mathbf{x}$ and a pair of domain labels d, d' as input and generates a fake image. The discriminator aims to predict the domain label of the image generated by the generator and distinguish whether it is fake or real. G and

Table 3. **CelebA** dataset - Comparison of SISA with FFVAE

Target	Domain	$\mathcal{S}$	Model	Fair?	Performance ↑	Unfairness ↓
					Accuracy	Demographic Parity - MD
Attractiveness	Hair color	{big nose, smiling male, young}	FFVAE SISA (ours)	Yes	64.11 $\pm$ 0.6 **72.40** $\pm$ 1.3	0.08 $\pm$ 0.02 **0.0173** $\pm$ 0.01

D are learned simultaneously as below:

$$\mathcal{L}_D^{\text{StarGAN}} = -\mathcal{L}_{\text{adv}}^{\text{StarGAN}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls(real)}}^{\text{StarGAN}} \quad (1)$$

$$\mathcal{L}_G^{\text{StarGAN}} = \mathcal{L}_{\text{adv}}^{\text{StarGAN}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls(fake)}}^{\text{StarGAN}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}^{\text{StarGAN}} \quad (2)$$

$\mathcal{L}_{\text{adv}}^{\text{StarGAN}}$ is the adversarial loss, $\mathcal{L}_{\text{cls(fake)}}^{\text{StarGAN}}$ and $\mathcal{L}_{\text{cls(real)}}^{\text{StarGAN}}$ are the domain classification losses of the fake and real images respectively, and $\mathcal{L}_{\text{rec}}^{\text{StarGAN}}$ is the reconstruction loss. We follow a similar procedure as outlined by FATDM [4](G^Y) and DIRT [3] to train domain invariant enabler G. To train G^s, we slightly deviate from the outlined method by FATDM [4]($G^{Y,A}$) due to multiple sensitive attributes. We partition the dataset corresponding to mini-batches (domains) where each batch is conditioned on the set of values taken by the sensitive attributes and the same label y. We perform the domain translation between these mini-batches instead of the combined batches. This is done to achieve domain invariant translations between all the sensitive attributes and the target label y.

2. Further Analysis on Datasets

2.1. Details on Dataset Size, Domain Splits, and Sensitive Attribute Splits

We have performed our experiments on data sizes ranging from 35078 (Pneumothorax prediction) to 255600 (Edema prediction) and *show that our model consistently performs well* across these ranges. Figs. 1, 3, 5, 7 and 9 show the number of training samples across CelebA, Cardiomegaly, Edema, Pneumothorax, and FACET datasets for each of their domains and also the total number of dataset samples. Figs. 2, 4, 6, 8 and 10 show how the values of the sensitive attributes we chose are distributed across the CelebA, Cardiomegaly, Edema, Pneumothorax, and FACET dataset respectively. It can be seen that a few of the values are imbalanced, and a few values are balanced, covering a wide spectrum. *From the graphs, it is evident that the datasets we have used have imbalanced samples with respect to domain, and sensitive attributes emulating a challenging real-world representative dataset.*

2.2. Effect of Increasing Number of Domains or Sensitive Attributes

We have performed our experiments on covariate shifts caused by the age of the patients, rotations of the input images, and hair color. We experiment with two to four sensitive attributes. We did not notice any pattern in the predictive performance or fairness (decrease/increase) due to low/high number of sensitive attributes. However, *SISA consistently performed better than FATDM across all these settings*.

- CelebA - Hair color, Domains:3, Sensitive Attributes:4
- Cardiomegaly - Age, Domains:4, Sensitive Attributes:2
- Edema - Rotations, Domains:5, Sensitive Attributes:3
- Pneumothorax - Age, Domains:3, Sensitive Attributes:2
- FACET - Visibility of Person, Domains:3, Sensitive Attributes:3

3. Further Analysis on Model

3.1. Comparison of SISA with FFVAE [2]

FFVAE [2] encodes multiple sensitive attributes to a flexible representation that can accommodate any subset of sensitive attributes at the test time. However, this method differs from our approach SISA in many ways. Their formulation is currently restricted to a single fairness metric, demographic parity. They also do not consider distribution shifts in the data. Moreover, they have conducted evaluations only for binary-sensitive attributes. We compare SISA with FFVAE on CelebA dataset as the sensitive attributes in CelebA are also binary. We follow the same DG setup discussed in in the main paper. We modified FFVAE code shared by [5] to suit our fair domain generalization with heterogeneous sensitive attributes. The results are available in Tab. 3.

3.2. TSNE Visualization of the Representations

We show TSNE plots of the predictive performance and the fairness representations in Figs. 11 and 12 for the CelebA dataset. Fig. 11 is the 2D representation of $\mathbf{z}_g$. The colors reflect whether the prediction is Attractive or Not Attractive. $\mathbf{z}_g$ can separate the classes well. On the other hand,

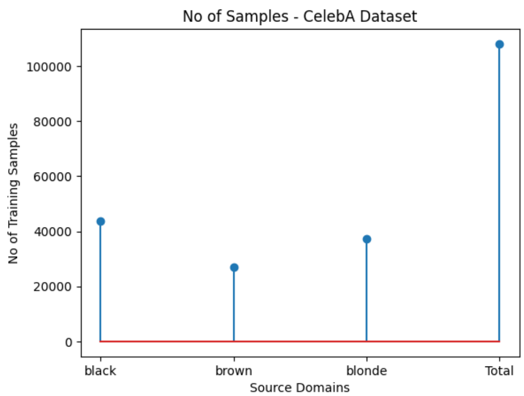

Figure 1. CelebA - Dataset Domain Distribution

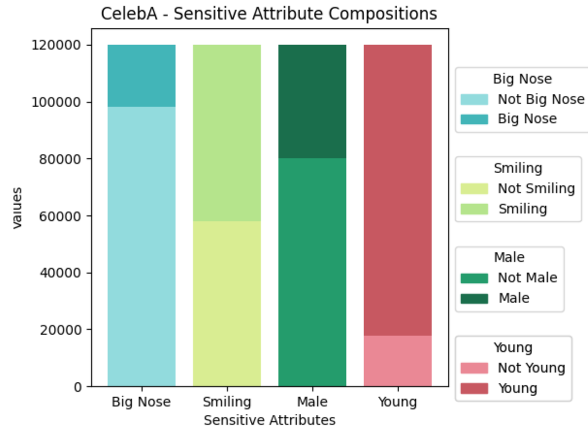

Figure 2. Celeba - Sensitive Attribute Distribution

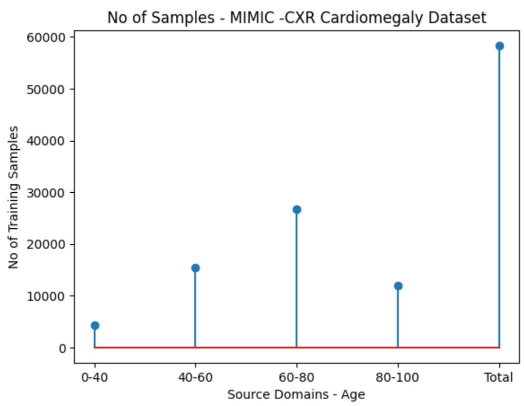

Figure 3. Cardiomegaly - Dataset Domain Distribution

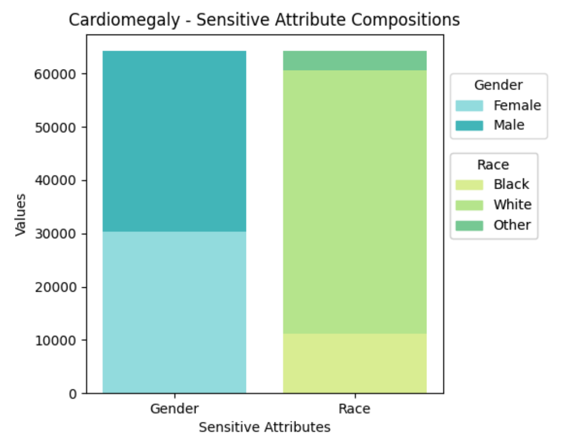

Figure 4. Cardiomegaly - Sensitive Attribute Distribution

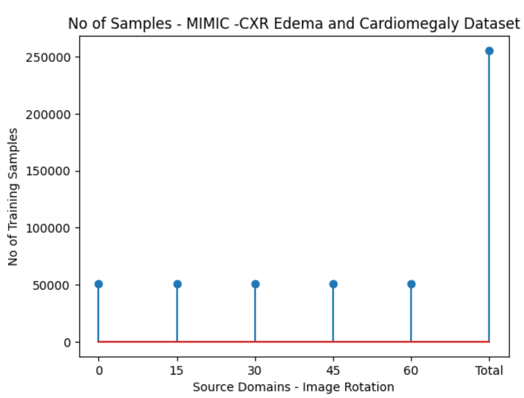

Figure 5. Edema - Dataset Domain Distribution

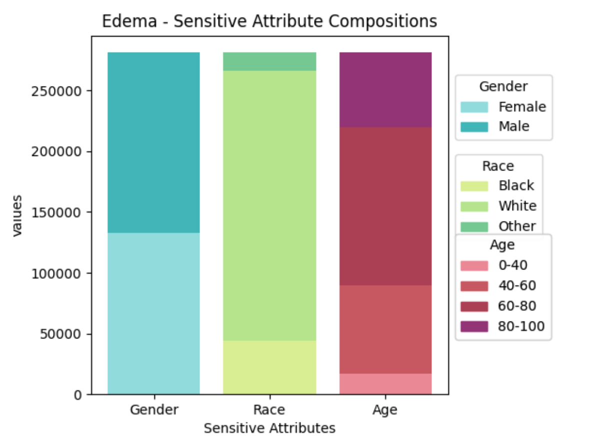

Figure 6. Edema - Sensitive Attribute Distribution

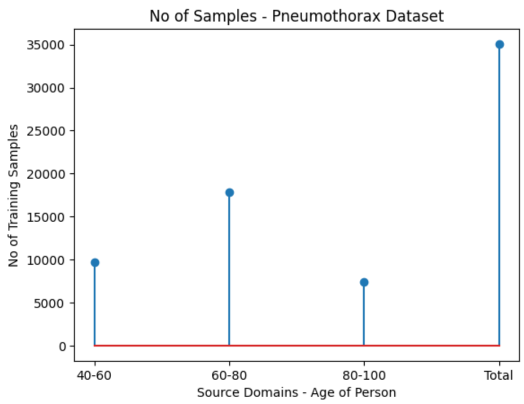

Figure 7. Pneumothorax - Dataset Domain Distribution

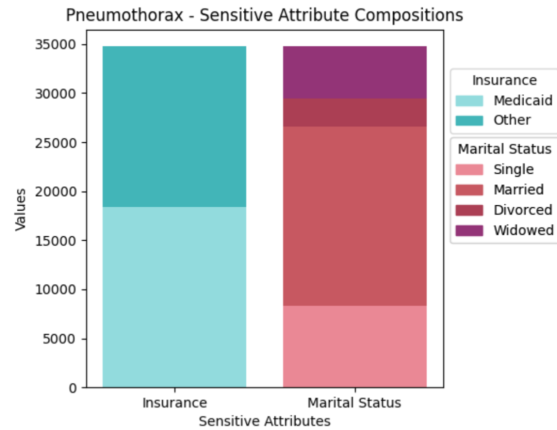

Figure 8. Pneumothorax - Sensitive Attribute Distribution

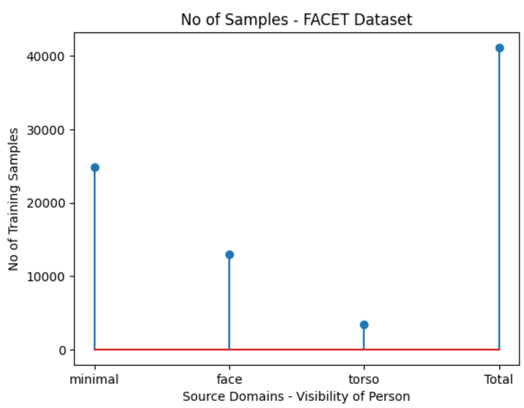

Figure 9. FACET - Dataset Domain Distribution

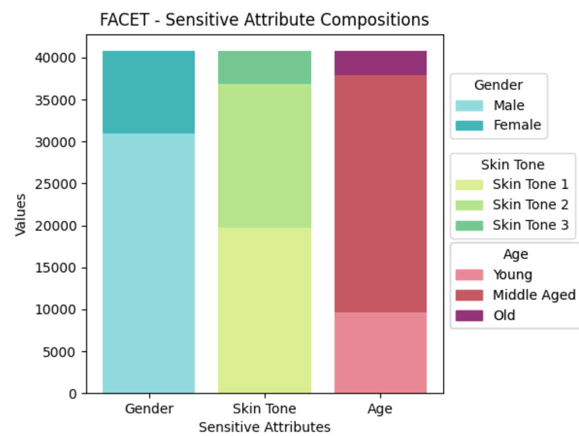

Figure 10. FACET - Sensitive Attribute Distribution

Fig. 12 is the 2D representation of $\mathbf{z}_f$. It shows that $\mathbf{z}_f$ is clustered based on attribute sensitivity. *E.g.*, red and blue clusters which share a sensitive attribute are close to each other. Similarly, yellow and lime green which do not share any sensitive attributes are far apart from each other.

3.3. Sensitivity Analysis on Different Model Components.

In this section, we report the variation of performance and fairness measures in the test set based on the different hyperparameter values of all hyperparameters in our model. The hyperparameters were tuned based on validation test accuracy for α, γ, and ϵ and validation mean distance for ω. While each hyperparameter was getting tuned, we fixed the values of all other hyperparameters. We observed that hyperparameter tuning could also help to find a good fairness-performance trade-off as each hyperparameter had control over predictive performance and fairness measures. *Finally, on average, the variation of most of the hyperparameters did not affect the model's predictive performance and fairness by a lot.* We report all sensitivity analysis on **CelebA** dataset (which had the highest number of sensitive attributes below. The other datasets also followed a similar trend.

Sensitivity analysis of ϵ: ϵ is the hyperparameter that decides how apart $\mathbf{z}_{f_i}$ and $\mathbf{z}'_{f_i}$ should be if sensitive attribute i is not equal. We train our models with $\epsilon = \{0.01, 0.1, 1, 10\}$ and report the results in Tab. 4 for CelebA dataset. We did not observe much difference in the performance and fairness as we changed ϵ. In general, we noticed slightly better test results for performance when $\epsilon \geq 1$ and for fairness when $\epsilon = 0.01$. We chose $\epsilon = 1$ as it had the best validation predictive performance.

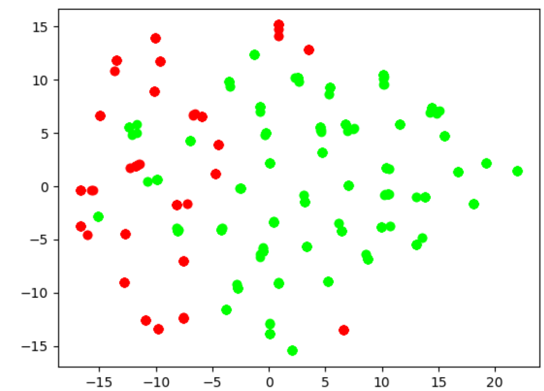

Figure 11. TSNE visualization of Representations $\mathbf{z}_g$

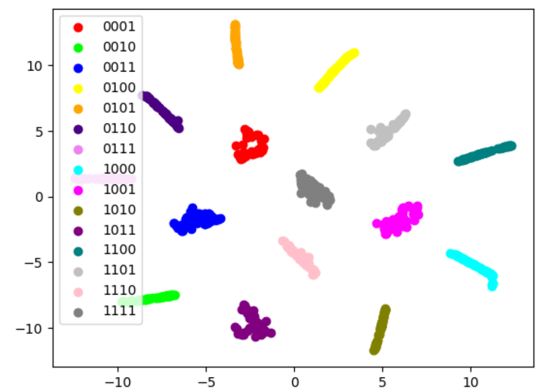

Figure 12. TSNE visualization of Representations $\mathbf{z}_f$

Table 4. Sensitivity analysis of ϵ using CelebA dataset

ϵ	Predictive Performance Measures (↑)				Unfairness Measures (↓)	
	AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
0.01	83.73	88.82	74.65	77.73	**0.0011**	**0.0148**
0.1	84.26	89.50	75.19	78.20	0.0020	0.0201
1	**84.82**	**90.02**	75.86	78.67	0.0017	0.0195
10	84.79	89.92	**76.16**	**79.17**	0.0045	0.0288

Sensitivity analysis of γ: γ is the hyperparameter that decides the weight of $\mathcal{L}_{DF}$ loss in our model. We trained our models with $\gamma = \{0.01, 0.1, 1, 10\}$ and report the results in Tab. 5 for CelebA dataset. We did not observe a lot of difference in the performance and fairness as we changed γ. In general, we noticed slightly better results for performance when γ values were higher and fairness when γ values were lower. We chose $\gamma = 1$ as it had the best validation predictive performance.

Table 5. Sensitivity analysis of γ using CelebA dataset

γ	Predictive Performance Measures (↑)				Unfairness Measures (↓)	
	AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
0.01	84.44	89.88	75.42	78.25	**0.0009**	**0.0165**
0.1	84.36	89.72	75.17	77.89	0.0016	0.0194
1	**84.82**	**90.02**	**75.86**	78.67	0.0017	0.0195
10	84.76	89.98	75.85	**78.78**	0.0014	0.0216

Sensitivity analysis on the cardinality of C ($|\mathbf{C}|$): The $|\mathbf{C}|$ is the size of the random subset of $\mathcal{C}$ sampled at each iteration. From our experiments, we observed that the best performance (validation and test) is when $|\mathcal{C}| = |\mathcal{C}|$ for all datasets. However, in the case of MIMIC dataset, we chose to go for a lower value for $|\mathbf{C}|$ as the predictive performance did not deteriorate much due to the huge training data size.

Table 6. Sensitivity analysis of cardinality of $\mathbf{C}$ using CelebA dataset

$\mathbf{\vert C\vert}$	Predictive Performance Measures (↑)				Unfairness Measures (↓)	
	AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
$\vert\mathcal{C}\vert = 15$	**84.82**	90.02	**75.86**	78.67	**0.0017**	**0.0195**
$\vert\mathcal{C}\vert/2 = 8$	84.71	**90.04**	75.74	**78.68**	0.0051	0.0352
$\vert\mathcal{C}\vert/3 = 5$	83.65	89.37	74.77	77.79	0.0133	0.0544

Table 7. Sensitivity analysis of α using CelebA dataset

α	Predictive Performance Measures (↑)				Unfairness Measures (↓)	
	AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
0.01	84.65	89.78	75.77	78.70	0.0025	0.0232
0.1	**84.82**	**90.02**	**75.86**	78.67	0.0017	**0.0195**
1	84.78	89.97	75.85	**78.85**	0.0021	0.0218
10	84.37	89.61	75.65	78.79	**0.0013**	0.0209

The sensitivity analysis of $\mathcal{C}$ is reported in Tab. 6.

Sensitivity analysis of α: α determines the weight of the domain invariance loss $\mathcal{L}_{DG}$. Initially, we chose it as 0.1 based on the original FATDM paper. Then we tuned it between $\alpha = \{0.01, 0.1, 1, 10\}$ and found that 0.1 gave the best validation accuracy for our model too. We report the results on the test set in Tab. 7.

Sensitivity analysis of ω: ω determines the weight of the fairness loss $\mathcal{L}_{EO}$. We observed that varying ω resulted in higher variance in the performance and fairness measures than other hyperparameters. Hence, the value of ω needed to be carefully chosen to have a good performance-fairness trade-off. Initially, we chose it as 1 based on the original FATDM paper. Then we tuned it between $\omega = \{0.1, 1, 5\}$ and found that $\omega = 5$ gave the best validation and test Mean

Table 8. Sensitivity analysis of parameter ω using CelebA dataset

ω	Predictive Performance Measures (↑)				Unfairness Measures (↓)	
	AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
0.1	**85.01**	**90.29**	**76.30**	**79.29**	0.3929	0.2568
1	84.82	90.02	75.86	78.67	0.0017	0.0195
5	81.22	86.49	70.70	72.37	**0.0000**	**0.0024**

Table 9. Ablation on the number of encoders used using CelebA dataset

Encoders	Performance Measures (↑)				Fairness Measures (↓)	
Used	**AUROC**	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
one	82.36	87.82	73.47	76.10	0.0003	0.0109
two	84.82	90.02	75.86	78.67	0.0017	0.0195

for our model. We report the results on the test set in Tab. 8. However, the validation accuracy of the model was very low, so the performance was getting hampered. Hence, we went with $\omega = 1$, which had an adequate Mean (< 0.002) and EMD (< 0.02) measure but good validation accuracy too (fairness - performance trade-off).

Ablation study on the representations: We conduct a study using only a single encoder to validate the efficacy of two separate encoders to model the domain shift (generalization) and fairness. We report the results in Tab. 9. We find that having multiple (two) encoders to model the representations improved the predictive performance while a single encoder improved the fairness.

In the case of a single-encoder model, a single representation $\mathbf{z}$ denotes the fairness and the generalization information. Hence, it is implicitly equally divided among the loss for the n sensitive attribute ($\mathcal{L}_{DF}$) and the generalization loss ($\mathcal{L}_{DG}$). As there are n sensitive attributes, it overshadows the generalization information due to being the same representation.

In the case of two-encoders model, where one encoder stands for fairness and the other for generalization performance, $\mathbf{z}$ is explicitly split between $\mathbf{z}_g$ and $\mathbf{z}_f$, giving $\mathbf{z}_g$ a good enough representation in $\mathbf{z}$ and not get overshadowed by $\mathbf{z}_f$. Hence, the generalization performance (accuracy) is better with two encoders.

3.4. Balancing Fairness and Predictive Performance

Based on the empirical results, we can use the various hyperparameters in the model to achieve a good predictive performance versus fairness trade-off. In general, we can obtain better fairness from higher ω. In most cases, we can get better predictive performance from higher values of γ and ϵ. We can generally obtain better predictive performance and fairness from high values of $|\mathcal{C}|$. Additionally, we also found that balancing the dimensions of the generalization ($\mathbf{z}_g$) and fairness ($\mathbf{z}_f$) representations can also be a good way to maintain the fairness and predictive performance trade-off.

3.5. Correlation between Model Outputs and Sensitive Attributes

Our model has a lower correlation between the target variable and the sensitive attributes due to the loss function $\mathcal{L}_{EO}$ which while achieving fairness also tries to remove the correlation between the attribute and the target variable. This can be viewed in Tab. 10.

Table 10. Pearson Correlation between Model Outputs and Sensitive Attributes.

Dataset	Domain	Model	Pearson Correlation (Target, S.A.)			
Edema	Image Rotation		Gender		Race	Age
		ERM	0.027		0.007	0.177
		SISA	0.016		0.000	0.103
CelebA	Hair Color		Big Nose	Male	Smiling	Young
		ERM	-0.186	0.188	-0.383	0.332
		SISA	-0.136	0.239	-0.329	0.257

3.6. Empirical Time and Space Complexity of SISA and FATDM

We have provided the running time analysis of our model SISA and the baseline FATDM (average computed over 2^n models) Tab. 11. The training time of a single model of SISA is higher than that of a single model of FATDM by roughly 4 times. However, FATDM needs to train 2^n models to be able to generalize fairness across all the target domain-sensitive attributes. Hence, SISA is more efficient than FATDM especially as n goes higher.

Regarding the space complexity, we have an additional encoder model f compared to FATDM's architecture. However, FATDM needs to train 2^n models and ends up needing more resources. For example, a model of SISA for analysing CelebA dataset had 23537857 trainable parameters. A single model of FATDM for training the same dataset had 12358209 trainable parameters instead. A single model of FATDM only needs half the parameters, but we need to train 2^n models of FATDM to accomplish the task a single model of SISA achieves.

3.7. Performance and Fairness Metrics on Each Target Sensitive Attribute Subset

We report the results for each subset of the set of sensitive attributes $\mathcal{S}$ in Tab. 12 for CelebA, Tab. 13 for Car-

Table 11. Running Time Analysis for SISA and FATDM

Model	Dataset (No of Sensitive Attributes)				
	CelebA (4)	**Cardiomegaly** (2)	**Edema** (3)	**Pneumothorax** (2)	**FACET** (3)
FATDM - 2^n Models	2d:14h:52m $\pm$ 5h:04m	4h:08m $\pm$ 0h:12m	2d:17h:44m $\pm$ 1h:20m	3h:40m $\pm$ 0h:48m	3d:7h:36 $\pm$ 5h:44m
SISA - Single	14h:30m $\pm$ 1h:19m	5h:37m $\pm$ 0h:54m	21h:46m $\pm$ 0h:29m	2h:13m $\pm$ 0h:24m	14h:44m $\pm$ 1h:16m

diomegaly, Tab. 14 for Edema, Tab. 15 for Pneumothorax, and Tab. 16 for FACET datasets. In general, the prediction performance of FATDM dropped with the introduction of more sensitive attributes. It reduced from 83.73 to 80.26 for Cardiomegaly prediction, 87.91 to 82.30 for Edema prediction, and 86.41 to 81.46 for Attractiveness prediction. In the case of SISA, the performance drop was lower, 83.60 to 82.59 for Cardiomegaly prediction, 87.73 to 86.71 for Edema prediction, and 85.14 to 84.57 for Attractiveness prediction. *On average, SISA maintained an adequate prediction performance while not compromising on the fairness metrics.* Additionally, for SISA we report the results for the None attribute where we do not consider any fairness attributes. However, we did not include this result while averaging to get a fair comparison with the FATDM baseline as FATDM does not have this configuration.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2018. 2

[2] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR, 09–15 Jun 2019. 1, 3

[3] A. Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. In *Advances in Neural Information Processing Systems*, volume 34, pages 5264–5275, 2021. 3

[4] Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[5] Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero Soriano, Samira Shabanian, and Sina Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. 3

Target Sensitive Attributes	Model	Predictive Performance Measures ↑				Unfairness Measures ↓	
		AUROC	AUPR	Acc	F1	Mean	EMD
big nose	FATDM-B	**86.17**	**90.83**	**77.30**	**80.03**	0.4845	0.5866
	SISA	85.14	90.41	76.34	79.32	**0.0035**	**0.0439**
smiling	FATDM-S	**86.60**	**91.12**	**77.18**	79.64	0.0957	0.3001
	SISA	86.27	90.87	77.08	**80.31**	**0.0006**	**0.0313**
male	FATDM-M	**85.45**	**90.31**	**75.75**	**78.59**	1.2440	1.0787
	SISA	83.88	89.44	74.97	77.98	**0.0160**	**0.1046**
young	FATDM-Y	**85.00**	90.16	**76.30**	**79.06**	1.0293	1.0202
	SISA	84.59	**90.18**	75.61	78.46	**0.0031**	**0.0578**
big nose, smiling	FATDM-BS	83.76	89.35	74.96	78.02	0.0695	0.1038
	SISA	**85.47**	**90.31**	**76.64**	**79.62**	**0.0002**	**0.0057**
big nose, male	FATDM-BM	82.82	88.43	74.11	77.11	0.0701	0.0920
	SISA	**84.63**	**89.89**	**75.88**	**78.80**	**0.0007**	**0.0071**
big nose, young	FATDM-BY	82.56	88.66	74.54	77.10	0.1002	0.1156
	SISA	**84.95**	**90.28**	**76.18**	**79.05**	**0.0002**	**0.0064**
smiling, male	FATDM-SM	84.06	89.56	73.34	75.84	0.2604	0.1776
	SISA	**84.53**	**89.80**	**75.57**	**78.45**	**0.0005**	**0.0080**
smiling, young	FATDM-SY	82.49	88.61	74.31	76.98	0.1067	0.1340
	SISA	**85.41**	**90.46**	**76.42**	**79.34**	**0.0002**	**0.0071**
male, young	FATDM-MY	81.93	88.37	73.14	75.79	0.1736	0.1404
	SISA	**84.09**	**89.80**	**75.36**	**78.08**	**0.0006**	**0.0077**
big nose, smiling, male	FATDM-BSM	82.18	87.99	73.00	75.89	0.0035	0.0214
	SISA	**84.81**	**89.79**	**75.72**	**78.44**	**0.0001**	**0.0025**
big nose, smiling, young	FATDM-BSY	82.46	88.42	73.85	76.13	0.0033	0.0215
	SISA	**84.96**	**89.94**	**76.07**	**78.72**	**0.0000**	**0.0025**
big nose, male, young	FATDM-BMY	80.96	86.96	72.62	75.20	0.0027	0.0198
	SISA	**84.51**	**89.84**	**75.67**	**78.24**	**0.0001**	**0.0031**
smiling, male, young	FATDM-SMY	81.65	88.16	72.68	75.25	0.0060	0.0238
	SISA	**84.43**	**89.66**	**74.92**	**77.39**	**0.0000**	**0.0025**
big nose, smiling, male, young	FATDM-BSMY	82.25	88.08	73.77	76.67	0.0002	0.0059
	SISA	**84.57**	**89.59**	**75.43**	**77.83**	**0.0000**	**0.0018**
None	SISA	86.10	90.79	76.91	80.22	-	-

Table 12. **CelebA** - Performance and Fairness on each Target domain

Target Sensitive Attributes	Model	Predictive Performance Measures				Unfairness Measures	
		AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
gender	FATDM-G	**84.86**	**92.57**	**76.84**	**81.78**	0.0947	0.2385
	SISA	84.78	92.49	76.77	81.73	**0.0554**	**0.2117**
race	FATDM-R	83.71	91.98	75.79	80.94	0.0231	0.0893
	SISA	**84.78**	**92.47**	**77.17**	**82.03**	**0.0035**	**0.0362**
gender, race	FATDM-GR	82.40	91.01	75.91	80.26	0.0064	0.0522
	SISA	**84.58**	**92.09**	**77.03**	**81.92**	**0.0003**	**0.0142**
None	SISA	84.70	92.47	76.80	81.76	-	-

Table 13. **Cardiomegaly (Age)** - Performance and Fairness on each Target domain

Target Sensitive Attributes	Model	Predictive Performance Measures ↑				Unfairness Measures ↓	
		AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
gender	FATDM-G	**87.91**	**86.65**	79.67	**79.33**	**0.0068**	0.0887
	SISA	87.73	86.36	**79.76**	**79.33**	0.0078	**0.0886**
race	FATDM-R	85.35	83.34	77.24	77.19	0.0009	0.0302
	SISA	**87.39**	**85.89**	**79.78**	**79.36**	**0.0001**	**0.0092**
age	FATDM-A	84.81	82.94	77.23	76.75	0.0007	0.0276
	SISA	**86.44**	**85.25**	**79.50**	**78.83**	**0.0001**	**0.0116**
gender, race	FATDM-GR	85.31	83.29	76.99	77.51	**0.0000**	0.0078
	SISA	**87.27**	**85.67**	**79.75**	**79.31**	**0.0000**	**0.0018**
gender, age	FATDM-GA	85.54	84.06	77.96	77.37	**0.0000**	0.0060
	SISA	**86.61**	**85.45**	**79.54**	**78.88**	**0.0000**	**0.0019**
race, age	FATDM-RA	84.45	82.51	76.03	76.37	**0.0000**	0.0029
	SISA	**86.82**	**85.53**	**79.59**	**79.00**	**0.0000**	**0.0013**
gender, race, age	FATDM-GRA	82.30	80.07	74.43	75.02	**0.0000**	0.0008
	SISA	**86.71**	**85.29**	**79.34**	**78.62**	**0.0000**	**0.0005**
None	SISA	87.73	86.37	79.75	79.29	-	-

Table 14. **Edema (Image Rotation)** - Performance and Fairness on each Target domain

Target Sensitive Attributes	Model	Predictive Performance Measures ↑				Unfairness Measures ↓	
		AUROC	**AUPR**	**Acc**	**F1**	**Mean**	**EMD**
insurance	FATDM-N	60.57	26.59	58.97	**34.15**	**0.0002**	**0.0195**
	SISA	**61.45**	**27.54**	**59.82**	**35.04**	**0.0002**	0.0209
marital status	FATDM-M	**60.65**	**26.50**	60.90	**33.68**	**0.0000**	0.0019
	SISA	60.19	25.59	**64.37**	32.16	**0.0000**	**0.0004**
insurance, marital status	FATDM-NM	59.38	24.92	62.46	31.79	**0.0000**	0.0005
	SISA	**60.17**	**25.52**	**66.11**	**34.57**	**0.0000**	**0.0004**
None	SISA	61.59	27.59	60.20	35.09	-	-

Table 15. **Pneumothorax (Age)** - Performance and Fairness on each Target domain

Target Sensitive Attributes	Model	Predictive Performance Measures	Unfairness Measures
		Acc	**Mean**
gender	FATDM-G	65.00	**1.21**
	SISA	**69.34**	1.39
race	FATDM-R	63.66	**1.39**
	SISA	**69.18**	**2.13**
age	FATDM-A	64.45	4.46
	SISA	**69.13**	**4.30**
gender, race	FATDM-GR	63.63	**7.45**
	SISA	**69.41**	8.25
gender, age	FATDM-GA	64.94	**7.30**
	SISA	**69.53**	8.54
race, age	FATDM-RA	64.94	**15.09**
	SISA	**69.49**	17.64
gender, race, age	FATDM-GRA	62.60	27.54
	SISA	**69.57**	**25.20**
None	SISA	69.00	-

Table 16. **FACET (Person Visibility)** - Performance and Fairness on each Target domain