Fair Domain Generalization with Heterogeneous Sensitive Attributes across Domains

Ragja Palakkadavath, Hung Le, Thanh Nguyen-Tang, Svetha Venkatesh, Sunil Gupta

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

APPLIED AI INSTITUTE
DEAKIN UNIVERSITY

Paper Code

PROBLEM STATEMENT

Problem Setting

- **Domain generalization(DG) techniques generalize classification to an unseen domain by leveraging data from multiple source domains. E.g. Generalizing to region E by training the model on regions A-D in Fig. 1.**
- **Recent methods (E.g., FATDM[1]) have started to equip Algorithmic Fairness into DG by considering a single sensitive attribute (Fig 2.) across all domains.**
- **We extended it to include multiple sensitive attributes (Fig. 3). But, this unnecessarily hampers predictive performance as fairness is enforced on all domains for every sensitive attribute.**
- **Each domain may require fairness on different sensitive attributes based on context.** E.g., Hiring requirements may differ across regions based on historical biases or social obligations.
- **We propose a novel setting Fair Domain Generalization with Heterogeneous Sensitive Attributes (Fig. 4) and an algorithm(SISA) to solve it.**
- **SISA maintains the tradeoff between predictive performance and fairness well.**

Domain Generalization

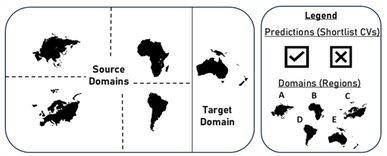


Fig 1.

Proposed Setting

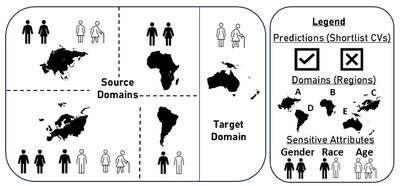


Fig 4. Fair Domain Generalization with Heterogeneous Sensitive Attributes

Existing approaches in Fair Domain Generalization

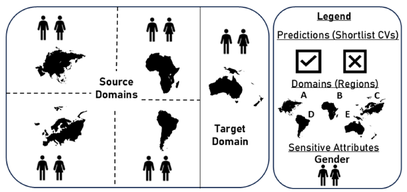


Fig 2.

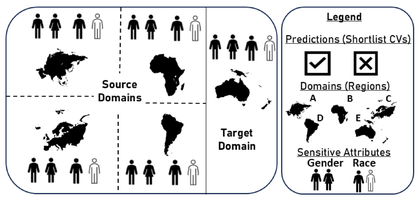


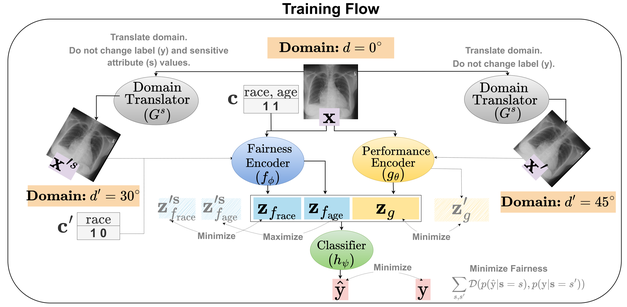
Fig 3.

METHODOLOGY and DATASETS

SISA Overview

- Data is in the form of image $\mathbf{x}$, set of sensitive attributes $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$, and label $\mathbf{y}$.
- Binary vector $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$ encodes the active and inactive sensitive attributes in each domain.
- Our method simultaneously encodes the input $\mathbf{x}$ into two latent representations as below:
  - $\mathbf{z}_g$ encodes predictive performance information.
  - $\mathbf{z}_f$ encodes fairness information.
- We employ domain invariance on $\mathbf{z}_g$ - $\mathbf{z}_g$ representations of a given class are brought closer in the latent space across domains.
- $\mathbf{z}_f$ is subdivided into distinct representations $\mathbf{z}_{f_i}$, one for each sensitive attributes in S.
  - We employ selective domain invariance on each $\mathbf{z}_{f_i}$ - if $\mathbf{z}_{f_i}$ representations from different domain share the same sensitive attribute, they are brought closer in the latent space and farther apart otherwise.
  - We encourage the representations of domains with the same sensitive attributes to lie closer as they must adhere to the same fairness constraint.
  - We encode attribute sensitivity in a domain via a through an indicator vector $\mathbf{c}$.
  - $\mathbf{c}$ is concatenated with each input image $\mathbf{x}$ before it is encoded as $\mathbf{z}_f$.

Training Flow



Testing Flow

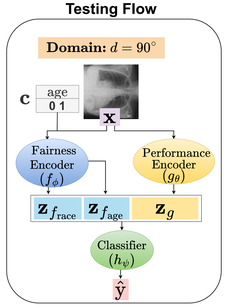


Figure 5.

Dataset Details

Dataset	Prediction		Domains		Sensitive Attributes
CelebA	Attractive	yes, no	hair color	brown, black, blonde	big nose: {yes, no}, male: {yes, no}, smiling: {yes, no}, young: {yes, no}
MIMIC CXR	Cardiomegaly	yes, no	age	0-40, 40-60, 60-80, 80-100	gender: {M, F}, race: {B, W, Other}
	Edema	yes, no	image rotations	$0°, 15°, 30°, 45°, 60°$	gender: {M, F}, race: {B, W, Other}, age: {0-40, 40-60, 60-80, 80-100}
	Pneumothorax	yes, no	age	40-60, 60-80, 80-100	marital status: {S, M, W, D}, insurance: {medic-aid, other}
FACET	Occupation	doctor, ...	person visibility	minimal, face, torso	skin tone: {tone1, tone2, tone3}, gender: {M,F}, age: {Y, M, O}

EXPERIMENTS

Experimental Results

Predict	Domain	$\mathcal{S}$	Model	Fair?	Predictive Performance Measures ↑ AUROC	AUPR	Accuracy	F1	Unfairness Measures ↓ MD	EMD
Attractive? (CelebA)	hair color	{big nose, smiling, male, young}	ERM	No	87.00±0.1	91.36±0.1	78.07±0.4	80.91±0.6	2.62±0.47	1.24±0.11
			DIRT		86.92±0.1	91.30±0.1	78.52±0.5	81.42±0.7	2.56±0.46	1.29±0.11
			ERMF	Yes	83.11±0.2	88.77±0.1	74.27±0.8	77.01±0.9	0.42±0.09	0.31±0.04
			FATDM		83.36±0.3	89.00±0.2	74.46±0.4	77.15±0.6	0.24±0.06	0.26±0.03
			SISA (ours)		**84.82**±0.5	**90.02**±0.3	**75.86**±0.4	**78.67**±0.5	**0.00**±0.01	**0.02**±0.00
Cardiomegaly?	age	{gender, race}	ERM	No	85.19±0.1	92.59±0.1	78.27±0.5	82.19±0.5	0.12±0.01	0.29±0.02
			DIRT		85.20±0.1	92.62±0.1	78.69±0.5	82.65±0.5	0.12±0.02	0.31±0.02
			ERMF	Yes	83.58±0.2	91.74±0.1	76.60±0.6	80.72±0.6	0.03±0.01	0.12±0.02
			FATDM		83.66±0.2	91.85±0.1	76.05±0.4	81.08±0.4	0.05±0.01	0.12±0.01
			SISA (ours)		**84.71**±0.2	**92.35**±0.2	**76.99**±0.4	**81.89**±0.4	**0.02**±0.00	**0.09**±0.01
Edema?	image rotation	{gender, race, age}	ERM	No	87.92±0.3	86.54±0.4	79.51±0.4	79.34±0.5	0.29±0.04	0.51±0.04
			DIRT		88.10±0.3	86.82±0.3	79.67±0.4	79.32±0.4	0.31±0.04	0.53±0.04
			ERMF	Yes	84.66±0.6	82.73±0.7	76.60±0.7	76.32±1.0	**0.00**±0.00	0.02±0.00
			FATDM		85.02±0.6	83.13±0.7	76.93±0.6	77.03±0.6	0.00±0.00	0.03±0.00
			SISA (ours)		**87.06**±0.5	**85.66**±0.5	**79.62**±0.5	**79.08**±0.5	0.00±0.00	**0.02**±0.00
Pneumothorax?	age	{insurance, marital-status}	ERM	No	61.25±0.5	27.70±0.4	58.66±2.6	34.21±0.6	0.21 ±0.08	0.20 ±0.06
			DIRT		61.56±0.6	58.47±0.6	28.06±1.9	34.80±0.4	0.06±0.00	0.15±0.01
			ERMF	Yes	60.16±0.5	25.95±0.4	61.44±1.9	32.61±0.9	0.00±0.00	**0.00**±0.01
			FATDM		60.20±0.5	26.00±0.5	60.78±2.1	33.21±0.6	0.00±0.00	0.01±0.05
			SISA (ours)		**60.60**±0.6	**26.22**±0.5	**63.44**±1.6	**33.93**±1.7	0.00±0.00	0.01±0.06

Table 1. Averaged Target Domain Predictive Performance and Fairness across Datasets

Single model of SISA v/s 2^n models of FATDM

Target Sensitive Attributes	Model	Predictive Performance Measures ↑ AUROC	AUPR	Acc	F1	Unfairness Measures ↓ 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 2. Individual Target Domain Predictive Performance and Fairness on CelebA Dataset

Target	Domain	$\mathcal{S}$	Model	Fair?	Performance ↑ Accuracy	Unfairness ↓ MD
Occupation	visibility	gender, skin-tone, age	ERM	No	65.60 ± 0.2	26.91±9.41
			FATDM	Yes	64.20 ± 0.5	9.57 ± 3.23
			SISA (ours)		**69.40** ± 0.1	**9.27** ± 3.69

Table 3. Average Predictive Performance and Fairness FACET Datasets

Dataset	Predicted using	Performance ↑ AUROC	F1	Unfairness ↓ Mean	EMD
CelebA	$\mathbf{z}_g$	**86.08**	**80.22**	1.13	0.68
	$\mathbf{z}_f$	51.91	43.83	**0.00**	**0.00**
	$\mathbf{z}$	84.82	78.67	0.00	0.02
Cardiomegaly	$\mathbf{z}_g$	84.30	81.79	0.04	0.12
	$\mathbf{z}_f$	49.71	26.04	**0.00**	**0.00**
	$\mathbf{z}$	**84.71**	**81.89**	0.02	0.09
Edema	$\mathbf{z}_g$	**87.71**	**79.32**	0.06	0.23
	$\mathbf{z}_f$	50.00	36.81	**0.00**	**0.00**
	$\mathbf{z}$	87.06	79.08	0.00	0.02
Pneumothorax	$\mathbf{z}_g$	**61.38**	**34.94**	0.00	0.02
	$\mathbf{z}_f$	56.04	23.48	**0.00**	**0.00**
	$\mathbf{z}$	60.60	33.93	0.00	0.01

Table 4. Ablation Study on Individual Representations

Qualitative Results

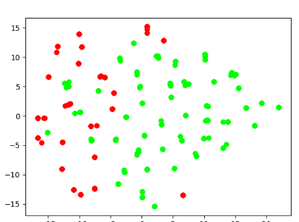


Fig 6. TSNE visualization of representation $\mathbf{z}_g$ The colors reflect whether the prediction is Attractive or Not Attractive. $\mathbf{z}_g$ partitions the classes well.

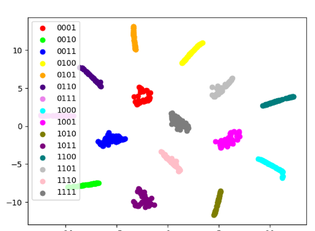


Fig 7. TSNE visualization of representation $\mathbf{z}_f$ $\mathbf{z}_f$ is clustered based on attribute sensitivity $\mathbf{c}$. Red and blue clusters that share a sensitive attribute are close to each other. Yellow and lime green that do not share any sensitive attributes are far apart.