FairDomain: Achieving Fairness in Cross-Domain Medical Image Segmentation and Classification

Yu Tian[1*], Congcong Wen[2,3*], Min Shi[1*], Muhammad Muneeb Afzal[3], Hao Huang[2,3], Muhammad Osama Khan[3], Yan Luo[1], Yi Fang[2,3**], and Mengyu Wang[1**]

[1] Harvard Ophthalmology AI Lab, Harvard University
[2] Center for Artificial Intelligence and Robotics, New York University Abu Dhabi
[3] Embodied AI and Robotics (AIR) Lab, New York University

Abstract. Addressing fairness in artificial intelligence (AI), particularly in medical AI, is crucial for ensuring equitable healthcare outcomes. Recent efforts to enhance fairness have introduced new methodologies and datasets in medical AI. However, the fairness issue under the setting of domain transfer is almost unexplored, while it is common that clinics rely on different imaging technologies (e.g., different retinal imaging modalities) for patient diagnosis. This paper presents FairDomain, a pioneering systemic study into algorithmic fairness under domain shifts, employing state-of-the-art domain adaptation (DA) and generalization (DG) algorithms for both medical segmentation and classification tasks to understand how biases are transferred between different domains. We also introduce a novel plug-and-play fair identity attention (FIA) module that adapts to various DA and DG algorithms to improve fairness by using self-attention to adjust feature importance based on demographic attributes. Additionally, we curate the first fairness-focused dataset with two paired imaging modalities for the same patient cohort on medical segmentation and classification tasks, to rigorously assess fairness in domain-shift scenarios. Excluding the confounding impact of demographic distribution variation between source and target domains will allow clearer quantification of the performance of domain transfer models. Our extensive evaluations reveal that the proposed FIA significantly enhances both model performance accounted for fairness across all domain shift settings (i.e., DA and DG) with respect to different demographics, which outperforms existing methods on both segmentation and classification. The code and data can be accessed at `https://ophai.hms.harvard.edu/datasets/harvard-fairdomain20k`.

Keywords: Fairness Learning · Domain Shift · Medical Segmentation · Medical Classification

* Contributed equally as co-first authors.
** Contributed equally as co-senior authors.

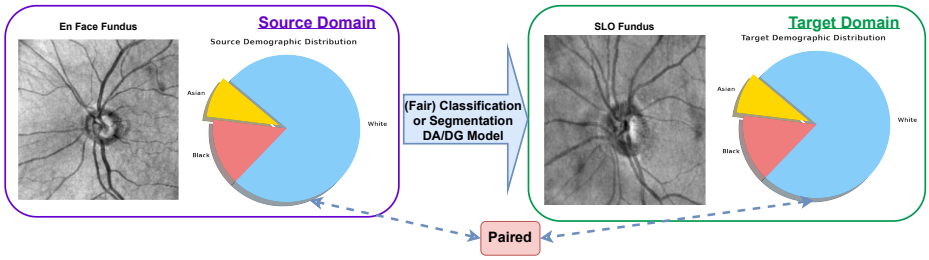


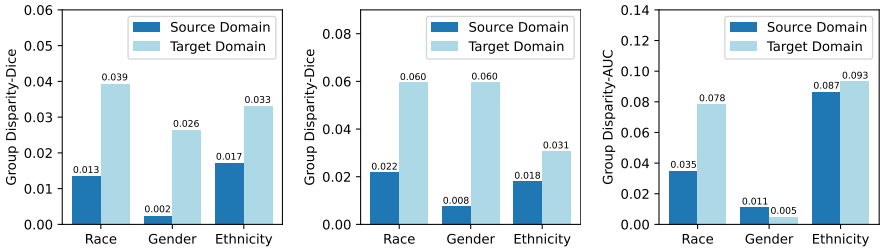
Fig. 1: FairDomain overview. An example of domain adaptation and generalization in healthcare: (fair) segmentation or classification model trained with patient data with one imaging modality, such as En Face Fundus images, can be adapted or generalized to another imaging modality such as SLO Fundus images, by maintaining both high accuracy and fairness.

1 Introduction

Advancements in deep learning have revolutionized the field of medical imaging, enabling significant improvements in tasks such as classification [9–11, 32, 33, 63–66, 68, 75] and segmentation [7, 30, 34, 34, 67, 77, 90]. These technologies have the potential to enhance diagnostic accuracy, streamline treatment planning, and ultimately improve patient outcomes. Despite these advancements, the deployment of deep learning models across varied healthcare settings has unearthed a pivotal challenge: the risk of inherent algorithmic bias and discrimination against certain demographic groups, which could undermine the fairness of medical diagnostics and treatments.

Recent studies have begun to address the issues of algorithmic biases in medical imaging by developing methodologies aimed at enhancing the fairness of deep learning models [4, 14, 23, 24, 26, 27, 36–38, 45, 59, 62, 69, 79, 87, 91]. These methodologies, while pioneering, commonly presuppose that the distribution of data during training and testing remains constant, thereby assuming that fairness measures implemented during training will suffice to ensure equitable decisions during testing within identical domains. This presumption, however, frequently does not hold in practical healthcare scenarios. For instance, primary care clinics and specialty hospitals may rely on different imaging technologies (e.g., different retinal imaging modalities) for patient diagnosis, leading to significant domain shifts that can adversely affect model performance and fairness when models trained on one type of imaging data are deployed on another. Therefore, it is critical to account for domain shifts and learn fair models that are robust to potential cross-domain scenarios in real-world deployment environments.

The previous literature extensively explores domain adaptation and domain generalization as methodologies to counteract the challenges posed by domain shifts, aiming to develop models that perform reliably across diverse but related domains. Domain adaptation, particularly in its unsupervised form, leverages both labeled data from a source domain and unlabeled data from a target domain to facilitate model generalization to new, unseen data [88]. Conversely, domain



(a) Cup segmentation **(b)** Rim segmentation **(c)** Glaucoma classification

Fig. 2: Group performance disparities in the source and target domains. It is evident that as domain shift occurs, algorithmic fairness significantly deteriorates throughout this process.

generalization operates under the premise that target domain data remain inaccessible during training, relying solely on source domain data to foster model applicability to novel, unseen domains [55]. Despite extensive research into enhancing model **accuracy** across domain shifts, the critical aspect of **fairness** — ensuring that models provide equitable predictions across different demographic groups — has been notably underexplored. This insufficiency is particularly critical in medical domains, where decision-making models directly impact human health, well-being, and safety. Only limited studies have begun to explore the transfer of fairness across domains [49, 50, 54]. However, all these previous studies lack systematically comprehensive investigations of fairness issues under domain shift. They predominantly focus on either domain adaptation or generalization, but rarely both. Furthermore, existing studies primarily address medical classification challenges, overlooking the critical task of medical segmentation, which is equally if not more significant under domain shifts in healthcare scenarios. Additionally, many studies adopt impractical assumptions about domain shifts, such as considering changes in patient age groups as domain shifts, which contradicts established understandings of medical domain shifts that are typically caused by the use of different imaging technologies [20, 29].

To address these gaps, our work introduces the first comprehensive systematic investigation of algorithmic fairness under domain shift in the medical imaging field, named FairDomain. We conduct extensive experiments across multiple state-of-the-art (SOTA) domain adaptation and generalization algorithms with three common demographic attributes, to assess both the accuracy and fairness of those algorithms and understand how fairness transfers across different domains in terms of both medical segmentation and classification tasks, as shown in Fig. 1. In Fig. 2, we investigate the transfer of algorithmic fairness across domains. Our observations reveal a significant exacerbation of group performance disparities between source and target domains (i.e., increased disparity performance values) when subjected to domain shifts in different medical classification and segmentation tasks. This indicates the critical necessity for devising fairness-oriented algorithms to address this pressing issue effectively. Recognizing the limitations of previous bias mitigation efforts, which were confined to single domains and lacked adaptability to diverse domain shift scenarios, we introduce a novel, versatile fair identity

attention (FIA) mechanism. This technique, designed for seamless integration with various domain adaptation and generalization strategies, utilizes self-attention derived from demographic attributes (e.g., racial groups) to harmonize feature significance across demographics, which will promote fairness. A pivotal challenge in developing the FairDomain benchmark was the absence of a medical imaging dataset that realistically encapsulates domain shifts in real-world medical domains, where domain shifts are typically induced by disparate imaging technologies, without conflating with demographic distribution shifts. Existing medical datasets with differences in patient demographics between source and target domains introduce ambiguity regarding the root causes of observed algorithmic bias — is it the performance fairness change in domain transfer due to demographic distributions or the inherent domain shift? To address this, we curated a unique dataset with paired retinal fundus images from the same patient cohort, captured through two distinct imaging modalities (En face and SLO fundus images), specifically for analyzing algorithmic bias in domain-shifted scenarios. To our knowledge, this is the first dataset meticulously designed to study fairness in domain shifts with consistent patient distributions, thus offering a clearer analysis of algorithmic biases attributable solely to domain shifts. Compared with previous medical datasets, our FairDomain dataset is distinguished by three key features: (1) It includes cross-domain data for both medical **segmentation** and **classification** tasks, offering a holistic view of algorithmic fairness under domain shifts; (2) The pairing of source and target domain images for the same patient cohort allows a clean-cut setting for studying the model fairness change due to domain shift excluding the confounding impact of demographic distribution variation on model fairness change; and (3) The dataset is enriched with six real-world demographic attributes from large eye hospitals, facilitating a more nuanced study of fairness. This new dataset serves as an invaluable resource not only for fairness research in cross-domain tasks but also for general cross-domain tasks as excluding the confounding impact of demographic distribution variation between source and target domains will allow clearer quantification of the performance of domain transfer models.

To summarize, our contributions are threefold:

- The first systematic exploration of algorithmic fairness under domain shifts in medical imaging.
- The introduction of fair identity attention techniques to improve accuracy and fairness across domain adaptations and generalizations.
- The creation of a large-scale, paired medical segmentation and classification dataset for fairness studies under domain shifts.

2 Related Work

Domain Adaptation and Generalization: Unsupervised Domain Adaptation (UDA) leverages unlabeled target data has become essential, with adversarial learning [12,21,71,72] and self-supervised training [43,93] standing out as principal strategies to harmonize feature distributions across different domains. These

methods, complemented by innovations like pseudo labeling [8], depth-aware adaptation [74], and entropy minimization [73], have significantly advanced UDA. In parallel, Domain Generalization (DG) seeks to cultivate domain-invariant features through adversarial techniques [17, 31], regularization [18, 48], instance reweighting [42], and meta-learning [5, 28]. Recent shifts towards style-based learning, which distinguishes between the content and style of images to mitigate domain gaps [13, 52, 76, 80, 82, 89]. Nevertheless, these advancements primarily focus on enhancing model accuracy and overlook incorporating fairness into adaptation and generalization frameworks.

Fairness Learning in Medical Imaging: The pursuit of fairness in machine learning, particularly in medical imaging, aims to mitigate biases and ensure equitable outcomes across diverse patient groups. Recent works have introduced fairness datasets [19, 23, 24, 38, 40, 69] and methodologies [6, 25, 35, 53, 56–58, 61, 78, 83–85] to study and improve algorithmic fairness. However, these efforts often overlook the dynamic nature of healthcare settings, where various medical imaging modalities may used from different healthcare domains, and therefore model fairness issues should be addressed in the context of cross-domain modeling.

Bridging Fairness under Domain Shift: The intersection of fairness under domain shift represents an emerging research frontier. Only a few existing works study the transfer of algorithmic bias under domain shift [54,54,70,86]. Pham et al. [54] focus on proposing a theoretical framework to understand how fairness and accuracy transfer by density matching and Truong [70] exemplify efforts to integrate fairness into domain adaptation processes, aiming to ensure equitable predictions across class distributions. Additionally, Zhang et al. [86] focus on using domain adaptation approaches to improve individual fairness. However, these studies predominantly focus on the theoretical frameworks of domain shift on limited tasks and lack comprehensive systemic empirical studies of fairness issues under domain shift. Moreover, prior methodologies often rely on restrictive assumptions regarding domain shifts due to a lack of high-quality domain shift datasets for a better study of this problem. The existing datasets and settings might not accurately reflect real-world scenarios, particularly in the context of medical imaging settings.

3 Dataset Analysis

Data Collection and Quality Control. Our institute's institutional review board (IRB) approved this study, which followed the principles of the Declaration of Helsinki. Since the study was retrospective, the IRB waived the requirement for informed consent from patients.

The subjects tested between 2010 and 2021 are from a large academic eye hospital of Harvard Medical School. Two cross-domain tasks, namely medical segmentation and classification, are studied in this work. For medical segmentation, there are five types of data: (1) En-face fundus imaging scans; (2) SLO fundus imaging scans; (3) patient demographics; (4) glaucoma diagnosis; and (5) cup-disc masks. Particularly, the pixel annotations of cup and disc regions are first acquired from the OCT machine, where the disc border in 3D OCT is segmented

as the Bruch's membrane opening by the OCT manufacturer software, and the cup border is detected as the intersection between the inner limiting membrane (ILM) and the plane that results in minimum surface area from the intersection and disc border on the plane [16, 46]. Approximately, the cup border can be considered as the closest location on the ILM to the disc border, which is defined as the Bruch's membrane opening. Both Bruch's membrane opening and the internal limiting membrane can be easily segmented due to the high contrast between them and the background. Since the OCT manufacturer software leverages 3D information, the cup and disc segmentation is generally reliable. Given the limited availability and high cost of OCT machines in primary care, we propose a method to transfer annotations from 3D OCT to 2D SLO fundus images, aiming to enhance early-stage glaucoma screening. By utilizing the registration tool NiftyReg [47], we accurately align SLO fundus images with OCT-derived pixel-wise annotations, generating a vast set of high-quality SLO fundus mask annotations. This process, verified by a panel of medical experts, shows an 80% success rate in registrations, streamlining the annotation process for broader applications in primary care settings. Upon the alignment and manual examination of those annotations, we leverage pixel-wise masks from both the SLO and En face fundus images to examine the transfer of algorithmic fairness in segmentation models under domain shifts.

For medical classification, there are four types of data: (1) En-face fundus imaging scans; (2) SLO fundus imaging scans; (3) patient demographics; and (4) glaucoma diagnosis. The subjects in the medical classification dataset are categorized into two classes including normal and glaucoma defined based on visual field tests.

Data Characteristics. The medical segmentation dataset contains 10,000 samples from 10,000 subjects. We divide our data into the training set with 8,000 samples, and the test set with 2,000 samples. The patient age average is 60.3 $\pm$ 16.5 years. Within this dataset, we have six demographic attributes including age, gender, race, ethnicity, preferred language, and marital status. The demogra-phic distributions are as follows: **Gender:** Female: 58.5%, and Male: 41.5%; **Race:** Asian: 9.2%, Black: 14.7%, and White: 76.1%. **Ethnicity:** Non-Hispanic: 90.6%, Hispanics: 3.7%, and Unknown: 5.7%. **Preferred Language:** English: 92.4%, Spanish: 1.5%, Others: 1%, and Unknown: 5.1%. **Marital Status:** Married or Partnered: 57.7%, Single: 27.1%, Divorced: 6.8%, Legally Separated: 0.8%, Windowed: 5.2%, and Unknown: 2.4%.

Similarly, the medical classification dataset contains 10,000 samples from 10,000 subjects with an average age of 60.9 $\pm$ 16.1 years. We divide our data into the training set with 8,000 samples, and the test set with 2,000 samples. Within this dataset, we have six demographic attributes including age, gender, race, ethnicity, preferred language, and marital status. The demographic distributions are as follows: **Gender:** Female: 72.5%, and Male: 27.5%; **Race:** Asian: 8.7%, Black: 14.5%, and White: 76.8%. **Ethnicity:** Non-Hispanic: 96.0%, Hispanics: 4.0%. **Preferred Language:** English: 92.6%, Spanish: 1.7%, Others: 3.6%, and Unknown: 2.1%. **Marital Status:** Married or Partnered: 58.5%, Singe: 26.1%, Divorced: 6.9%, Legally Separated: 0.8%, Windowed: 1.9%, and Unknown: 5.8%.

4 Method

4.1 Problem Statement

Domain Adaptation (DA) and Domain Generalization (DG) are pivotal techniques in the development of machine learning models, aimed at addressing the variability that can occur when a model trained in one specific domain is applied to another. In the field of medical imaging, the techniques of DA and DG are critical in creating models that can robustly handle the variability present across different medical institutions, imaging devices, and patient populations. In this paper, we aim to explore the dynamics of fairness within the context of domain shift and develop methodologies to ensure that models remain fair and reliable as they are adapted or generalized to new domains. The problem definitions of DA and DG are as follows:

Domain Adaptation (DA): Given a domain data $D_S = \{(x_m^S, y_m^S, a_m^S)\}_{m=1}^{N_S}$, where $x_m^S \in \mathbb{R}^{H\times W}$ represents the m-th image in the source domain. For image classification, $y_m^S \in \mathbb{R}^K$ is the category label indicating the class to which the image belongs. For image segmentation, $y_m^S \in \mathbb{R}^{K\times H\times W}$ is the label matrix with each element specifying the class for each pixel. In both contexts, $a_m^S \in \mathbb{R}^d$ denotes an identity attribute associated with the patient, such as gender, race, or ethnicity. Additionally, there is an unlabeled target domain $D_T = \{x_m^T\}_{m=1}^{N_T}$, where the data distribution typically differs from the source domain. In addition to minimizing the distribution discrepancy between the source and target domains for effective performance on the target domain, a significant concern in DA is to ensure that the adaptation process respects fairness. This entails developing adaptation strategies that prevent the amplification of any biases present in the source domain and promote equity in treatment and outcomes across different groups defined by the identity attributes a_m^S in both the source and target domains.

Domain Generalization (DG): Consider a setting with labeled source domain data $D_S = \{(x_m^S, y_m^S, a_m^S)\}_{m=1}^{N_S}$. Unlike DA, the DG approach does not utilize data from the target domain during training. Instead, the objective is to learn a model f_θ from D_S that can generalize effectively to any related but unseen target domain T, characterized by a potential shift in data distribution. The challenge in DG is to extract robust domain-invariant features from D_S that are predictive of the target domain T while also ensuring that the model's predictions are fair and unbiased with respect to the identity attributes represented by a_m^S, without having any access to T during training.

4.2 Fair Identity Attention

We aim to develop a methodical function f that mitigates fairness deterioration, which is often observed during the transfer of models from a source domain to a target domain. Such deterioration is primarily due to domain shift that can amplify existing biases in the dataset, notably those associated with demographic attributes like gender, race, or ethnicity. To address this, we propose an attention

based mechanism designed to discern and utilize features from images that are pertinent to the downstream tasks, such as segmentation and classification, while also taking into account the demographic attributes.

More formally, we seek to optimize a model $f_\theta : \mathbb{R}^{H\times W} \times \mathbb{R}^d \xrightarrow{\theta} \mathcal{Y}$, where the outcome space $\mathcal{Y}$ encompasses both $\mathbb{R}^K$ for classification or $\mathbb{R}^{K\times H\times W}$ for segmentation, reflecting the objectives of categorizing images into K classes or labeling each pixel in an image with one of these K classes, respectively.

Fig. 3 illustrates the architecture of the proposed fair identity attention module. The module initially processes the input image and attribute labels through an embedding layer to obtain the input embedding E_i and attribute embedding E_a. These embeddings are then added to the position embedding E_p. The sum of E_i and E_p is utilized to generate the attention keys K_{op} and values V_{op}, while the sum of E_a and E_p is used to produce the attention query Q_g. The specific formulas are as follows:

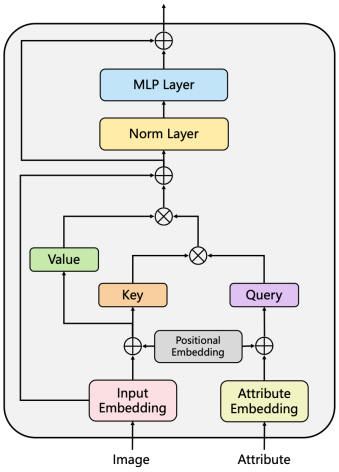


Fig. 3: Illustration of the proposed fair identity attention (FIA) Block. This block maps demographic attributes onto attention query vectors, thereby dynamically extracting relevant salient features associated with the current attributes for downstream tasks.

$$K = \text{Linear}(E_i{+}E_p), \quad V = \text{Linear}(E_i{+}E_p), \quad Q_a = \text{Linear}(E_a{+}E_p). \tag{1}$$

By performing the dot product of the query and keys, we extract a similarity matrix relevant to the current attribute. This matrix is then used in a dot product with the values, extracting features that are significant for each attribute with respect to downstream tasks. This process is expressed by the following equation:

$$FairAttention(Q_a, K, V) = \text{softmax}\left(\frac{Q_a \cdot K_T}{\sqrt{D}}\right) \cdot V, \tag{2}$$

where D is a scaling factor to avoid overly large values in the softmax function. Subsequently, a residual connection adds E_i to the output of the attention to maintain the integrity of the input information. Finally, a normalization layer and a multi-layer perceptron (MLP) layer are used to further extract features. After performing another residual operation on the output of these two layers, we obtain the final output of the fairness attention module.

The fair identity attention mechanism serves as a powerful and versatile tool designed to enhance model performance while addressing fairness concerns. By explicitly considering demographic attributes such as gender, race, or ethnicity, it ensures that the learned representations do not inadvertently amplify existing biases present in the data. Its architecture allows it to be seamlessly integrated as a plug-in component into any existing network. This modular nature empowers

researchers and practitioners to retrofit fair identity attention into their models without the need for extensive modifications to the underlying architecture. Consequently, the fair identity attention module not only contributes to the improvement of model accuracy and fairness in segmentation and classification tasks but also promotes ethical AI practices by facilitating more equitable treatment of diverse groups within datasets.

5 Experiments

5.1 Algorithmic Fairness Across Domain Shifts

In our experiments, we first analyze fairness in the context of domain shifts within the Cup-Disc Segmentation task. Cup-Disc Segmentation refers to the process of accurately delineating the optic cup and disc in fundus images, which is essential for calculating the cup-to-disc ratio (CDR) — a critical parameter for assessing the progression and risk of glaucoma. This task is pivotal in the field of medical imaging, particularly within the context of diagnosing and managing eye diseases such as glaucoma. Since the cup is a substantial subarea of the disc, we reformulate the segmentation task into cup and rim (the tissue area between the cup and disc border) segmentation to avoid misrepresented performance due to the large overlapped area between the cup and disc.

We investigate the performance of fairness across three different demographics, including gender, race, and ethnicity, across two distinct domains: En face fundus images derived from optical coherence tomography scans and SLO (Scanning Laser Ophthalmoscopy) fundus images. In the subsequent experiments, we selected en face fundus images as the source domain and SLO fundus images as the target domain. The rationale is that en face fundus images are routinely acquired in specialized ophthalmic care settings in comparison to SLO fundus images, leading to significantly greater availability of data. Therefore, we have chosen to position en face fundus images as the source domain, with SLO fundus images as the target domain. For classification, we utilize fundus images from those two domains as source and target domains, categorized into two classes: normal and glaucoma.

Evaluation Metrics: For evaluating model performance, the Dice coefficient and the intersection over union (IoU) metrics are employed for segmentation assessments, and the area under the receiver operating characteristic curve (AUC) is utilized for classification tasks. Those traditional metrics for segmentation and classification, while indicative of model performance, do not inherently account for fairness across demographic identity groups. Inspired by [39, 69], to address the potential tradeoff between model performance and fairness in medical imaging, we use novel equity-scaled performance (ESP) metrics to assess both performance and fairness for both segmentation and classification tasks.

Let $\mathcal{M} \in \{\text{Dice}, \text{IoU}, \text{AUC}, \ldots\}$ signify a generic performance metric applicable to either segmentation or classification. Traditional evaluation usually takes a

Table 1: Optic cup and rim segmentation performance on the **FairDomain-Segmentation** dataset using different Domain Adaptation (DA) and Domain Generalization (DG) methods with **gender** as the demographic attribute.

		Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Male Dice↑	Female Dice↑	Male IoU↑	Female IoU↑
	Cup	Baseline (Source)	0.885	0.888	0.806	0.808	0.886	0.889	0.807	0.810
	Cup	Baseline (Target)	0.688	0.700	0.535	0.555	0.693	0.711	0.557	0.574
	Rim	Baseline (Source)	0.854	0.861	0.753	0.762	0.864	0.856	0.767	0.755
	Rim	Baseline (Target)	0.485	0.495	0.336	0.342	0.486	0.507	0.334	0.353
DA	Cup	PixMatch [44]	0.768	0.775	0.650	0.660	0.772	0.769	0.645	0.660
DA	Cup	CBST [92]	0.791	0.797	0.682	0.686	0.794	0.802	0.684	0.690
DA	Cup	DAFormer [22]	0.781	0.785	0.676	0.680	0.783	0.789	0.678	0.684
DA	Cup	**DAFormer-FIA**	**0.802**	**0.810**	**0.692**	**0.700**	**0.806**	**0.816**	**0.695**	**0.706**
DA	Rim	PixMatch [44]	0.660	0.673	0.519	0.523	0.669	0.688	0.519	0.528
DA	Rim	CBST [92]	**0.693**	**0.696**	**0.541**	**0.544**	**0.694**	**0.698**	**0.542**	**0.548**
DA	Rim	DAFormer [22]	0.344	0.345	0.212	0.213	0.344	0.347	0.212	0.214
DA	Rim	**DAFormer-FIA**	0.528	0.531	0.367	0.369	0.533	0.528	0.372	0.366
DG	Cup	GIN+IPA [51]	0.741	0.750	0.590	0.594	0.740	0.752	0.590	0.597
DG	Cup	AADG [41]	0.711	0.715	0.595	0.598	0.713	0.717	0.596	0.601
DG	Cup	DAFormer [22]	0.787	0.795	0.677	0.684	0.791	0.802	0.680	0.691
DG	Cup	**DAFormer-FIA**	**0.816**	**0.820**	**0.712**	**0.716**	**0.818**	**0.823**	**0.714**	**0.719**
DG	Rim	GIN+IPA [51]	0.550	0.555	0.397	0.402	0.551	0.560	0.394	0.408
DG	Rim	AADG [41]	0.621	0.621	0.465	0.466	0.621	0.621	0.465	0.467
DG	Rim	DAFormer [22]	0.659	0.669	0.507	0.516	0.663	0.677	0.509	0.526
DG	Rim	**DAFormer-FIA**	**0.704**	**0.705**	**0.557**	**0.559**	**0.704**	**0.706**	**0.558**	**0.561**

Table 2: Optic cup and rim segmentation performance on the **FairDomain-Segmentation** dataset using different Domain Adaptation (DA) and Domain Generalization (DG) methods with **race** as the demographic attribute.

		Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Asian Dice↑	Black Dice↑	White Dice↑	Asian IoU↑	Black IoU↑	White IoU↑
	Cup	Baseline (Source)	0.868	0.888	0.784	0.808	0.881	0.901	0.886	0.802	0.830	0.805
	Cup	Baseline (Target)	0.673	0.700	0.535	0.564	0.703	0.731	0.694	0.567	0.606	0.555
	Rim	Baseline (Source)	0.820	0.861	0.716	0.762	0.841	0.838	0.868	0.737	0.731	0.771
	Rim	Baseline (Target)	0.481	0.495	0.334	0.342	0.496	0.472	0.499	0.347	0.326	0.345
DA	Cup	PixMatch [44]	0.739	0.775	0.614	0.660	0.752	0.793	0.783	0.633	0.693	0.673
DA	Cup	CBST [92]	0.770	0.791	0.662	0.679	0.773	0.785	0.794	0.657	0.680	0.681
DA	Cup	DAFormer [22]	0.785	0.804	0.669	0.691	0.794	0.816	0.802	0.682	0.712	0.688
DA	Cup	**DAFormer-FIA**	**0.796**	**0.810**	**0.682**	**0.700**	**0.804**	**0.821**	**0.809**	**0.694**	**0.717**	**0.697**
DA	Rim	PixMatch [44]	**0.627**	0.673	0.493	0.523	**0.653**	0.640	0.693	**0.522**	0.498	0.559
DA	Rim	CBST [92]	**0.627**	**0.702**	**0.496**	**0.556**	0.650	**0.650**	**0.719**	0.502	**0.506**	**0.573**
DA	Rim	DAFormer [22]	0.326	0.332	0.199	0.202	0.334	0.345	0.329	0.204	0.212	0.199
DA	Rim	**DAFormer-FIA**	0.521	0.531	0.363	0.369	0.522	0.539	0.530	0.364	0.380	0.368
DG	Cup	GIN+IPA [51]	0.714	0.750	0.568	0.594	0.732	0.762	0.730	0.590	0.635	0.595
DG	Cup	AADG [41]	0.694	0.715	0.586	0.598	0.723	0.696	0.717	0.607	0.588	0.600
DG	Cup	DAFormer [22]	0.769	0.798	0.673	0.688	0.771	0.809	0.799	0.662	0.706	0.687
DG	Cup	**DAFormer-FIA**	**0.787**	**0.810**	**0.675**	**0.701**	**0.790**	**0.820**	**0.811**	**0.678**	**0.717**	**0.700**
DG	Rim	GIN+IPA [51]	0.527	0.555	0.383	0.402	0.538	0.542	0.577	0.404	0.378	0.424
DG	Rim	AADG [41]	0.578	0.621	0.436	0.466	0.597	0.581	0.632	0.443	0.430	0.476
DG	Rim	DAFormer [22]	0.616	0.676	0.481	0.527	0.638	0.631	0.689	0.490	0.481	0.540
DG	Rim	**DAFormer-FIA**	**0.636**	**0.689**	**0.499**	**0.540**	**0.655**	**0.650**	**0.701**	**0.509**	**0.500**	**0.552**

set of triplets (z', a, y) as input to produce the metric score $\mathcal{M}(\{(z', y)\})$, which typically disregards demographic identity attributes, thereby missing critical fairness assessment. To incorporate fairness, we first compute a performance discrepancy Δ, defined as the aggregate deviation of each demographic group's metric from the overall performance, expressed as:

$$\Delta = \sum_{A \in \mathcal{A}} |\mathcal{M}(\{(z', y)\}) - \mathcal{M}(\{(z', a, y) | a = A\})|, \tag{3}$$

where Δ approaches zero when performance equity across groups is achieved, reflecting minimal disparity. The ESP metric can then be formulated as follows:

$$\text{ESP} = \frac{\mathcal{M}(\{(z', y)\})}{1 + \Delta}, \tag{4}$$

which ensures a balanced evaluation of model accuracy and fairness. This metric, ESP, aligns with $\mathcal{M}$ when Δ is minimized, indicating equitable performance across demographics. Conversely, an increased Δ denotes significant disparities between different demographic groups, lowering the ESP score with a larger penalty to indicate that the models obtain less fairness. This unified metric facilitates a comprehensive assessment of deep learning models, emphasizing not only their accuracy (as measured by segmentation and classification metrics such as Dice, IoU and AUC) but also their fairness across different demographic groups.

Cup-Rim Segmentation Results under Domain Shifts

Baselines: We first utilized TransUNet as the baseline model to perform training on the source domain. After the model training phase, we directly evaluated the model's performance on the target domain without any domain adaptation or generalization strategies. The results of this baseline model on source and target domains are detailed in the first four rows of Table 1, Tables 2, and Table 3. It can be observed that a significant decrease in segmentation accuracy for both the optic cup and rim when applying a model trained on the source domain directly to the target domain. This performance decline highlights the critical issue of domain shift in medical imaging segmentation tasks, underscoring the challenge of transferring learned models between differing imaging domains without appropriate adaptation or generalization strategies.

Domain Adaptation for Segmentation: To assess the performance of existing domain adaptation methods, we selected three state-of-the-art models as baseline methods for domain adaptation: **PixMatch** [44], **CBST** [92], and **DAFormer** [22]. Given that DAFormer has demonstrated superior performance over the other two methods in previous tasks, and considering its applicability to both domain adaptation and domain generalization tasks, we have integrated our proposed fair identity attention mechanism into DAFormer, denoted as **DAFormer-FIA** to evaluate the effectiveness of our model. Table 1, Tables 2, and Table 3 present the domain adaptation results of these four models across three demographic attributes: gender, race, and ethnicity. It is noteworthy that our DAFormer-FIA achieves improvements in the segmentation of both the cup and rim across all attributes compared to the baseline DAFormer. Specifically, our model improved the ES-Dice for cup segmentation from 0.781 to 0.802 and the ES-Dice for rim segmentation from 0.344 to 0.528 in the gender attribute. For race attribute, it increased the ES-Dice for cup segmentation from 0.785 to 0.796, and for rim segmentation, from 0.326 to 0.521. In terms of ethnicity attribute, the

Table 3: Optic Cup and Rim segmentation performance on the **FairDomain-Segmentation** dataset using different Domain Adaptation (DA) and Domain Generalization (DG) methods with **ethnicity** as the demographic attribute.

		Method	Overall ES-Dice↑	Overall Dice↑	Overall ES-IoU↑	Overall IoU↑	Hispanic Dice↑	Non-Hispanic Dice↑	Hispanic IoU↑	Non-Hispanic IoU↑
	Cup	Baseline (Source)	0.871	0.888	0.784	0.808	0.887	0.906	0.807	0.839
	Cup	Baseline (Target)	0.684	0.700	0.551	0.555	0.699	0.722	0.563	0.587
	Rim	Baseline (Source)	0.845	0.861	0.743	0.762	0.860	0.879	0.761	0.786
	Rim	Baseline (Target)	0.489	0.495	0.340	0.342	0.495	0.506	0.342	0.348
DA	Cup	PixMatch [44]	0.740	0.775	0.628	0.660	0.763	0.811	0.649	0.699
DA	Cup	CBST [44]	0.768	0.797	0.655	0.686	0.794	0.833	0.683	0.730
DA	Cup	DAFormer [22]	0.773	0.796	0.654	0.680	0.794	0.824	0.677	0.717
DA	Cup	**DAFormer-FIA**	**0.790**	**0.810**	**0.674**	**0.700**	**0.808**	**0.834**	**0.697**	**0.734**
DA	Rim	PixMatch [44]	0.643	0.673	0.499	0.523	0.673	0.719	0.513	0.562
DA	Rim	CBST [92]	**0.675**	**0.696**	**0.528**	**0.544**	**0.693**	**0.723**	**0.542**	**0.573**
DA	Rim	DAFormer [22]	0.405	0.418	0.264	0.270	0.416	0.448	0.269	0.295
DA	Rim	**DAFormer-FIA**	0.516	0.531	0.359	0.369	0.529	0.558	0.368	0.395
DG	Cup	GIN+IPA [51]	0.714	0.750	0.557	0.594	0.710	0.760	0.540	0.607
DG	Cup	AADG [41]	0.705	0.715	0.594	0.598	0.715	0.700	0.597	0.605
DG	Cup	DAFormer [22]	0.746	0.746	0.630	0.640	0.746	0.746	0.640	0.656
DG	Cup	**DAFormer-FIA**	**0.785**	**0.810**	**0.672**	**0.701**	**0.809**	**0.842**	**0.699**	**0.743**
DG	Rim	GIN+IPA [51]	0.536	0.555	0.391	0.402	0.537	0.572	0.398	0.426
DG	Rim	AADG [41]	0.603	0.621	0.453	0.466	0.620	0.649	0.465	0.494
DG	Rim	DAFormer [22]	0.630	0.641	0.491	0.499	0.642	0.624	0.499	0.485
DG	Rim	**DAFormer-FIA**	**0.676**	**0.689**	**0.531**	**0.540**	**0.688**	**0.708**	**0.540**	**0.557**

ES-Dice for cup segmentation was enhanced from 0.773 to 0.790, and the ES-Dice for rim segmentation was elevated from 0.405 to 0.516. Concurrently, for cup segmentation, our model achieved the highest ES-Dice scores of 0.796 for gender, 0.796 for race, and 0.790 for ethnicity. Similarly, it secured the best ES-IOU scores of 0.692 for gender, 0.682 for race, and 0.674 for ethnicity. Regarding specific demographic attributes, DAFormer-FIA increased the Dice for rim segmentation in the male category within gender from 0.344 to 0.533, marking an improvement of approximately 54.94%. In the white category within race, it improved from 0.329 to 0.530, a 61.09% increase. Lastly, in the Hispanic category within ethnicity, the Dice was enhanced from 0.416 to 0.529, translating to a 27.16% uplift. These substantial percentage improvements underscore the effectiveness of our model across diverse demographic attributes. This consistent enhancement across various demographic attributes highlights the effectiveness of integrating fair identity attention into domain adaptation tasks, particularly in augmenting the segmentation accuracy of critical ocular structures.

Domain Generalization for Segmentation: To validate the performance of existing domain generalization methods, we selected three state-of-the-art models as baselines for domain generalization: GIN+IPA [51], AADG [41], and DAFormer [22]. Similarly, we integrated our proposed fair identity attention mechanism into DAFormer to assess the effectiveness of our model. Tables 2, Table 1, and Table 3 outline the domain generalization results of these four models across three demographic attributes: race, gender, and ethnicity. It can be found from these tables that our DAFormer-FIA not only enhances the segmentation of both the cup and rim across all three attributes over the standard DAFormer but also achieves the highest segmentation performance for both the cup and rim across

Table 4: Classification performance on the **FairDomain-Classification** dataset using different Domain Adaptation (DA) and Domain Generalization (DG) methods with **gender**, **race**, and **ethnicity** as the demographic attribute.

	Method	Gender Overall ES-AUC↑	Gender Overall AUC↑	Male Group AUC↑	Female Group AUC↑	Race Overall ES-AUC↑	Race Overall AUC↑	Asian Group AUC↑	Black Group AUC↑	White Group AUC↑	Ethnicity Overall ES-AUC↑	Ethnicity Overall AUC↑	Non-Hispanic Group AUC↑	Hispanic Group AUC↑
	Baseline (Source)	0.795	0.803	0.806	0.796	0.730	0.803	0.788	0.730	0.815	0.744	0.803	0.806	0.727
	Baseline (Target)	0.536	0.537	0.536	0.538	0.508	0.537	0.555	0.511	0.524	0.517	0.537	0.539	0.501
DA	CGDM [15]	0.611	0.618	0.613	0.625	0.556	0.618	0.662	0.604	**0.671**	0.579	0.618	0.620	0.554
DA	CDTrans [81]	0.631	0.633	0.634	0.631	0.603	0.633	0.658	0.614	0.628	0.600	0.633	0.635	0.579
DA	**CDTrans-FIA**	**0.633**	**0.635**	**0.636**	**0.633**	**0.606**	**0.636**	**0.663**	**0.619**	0.631	**0.607**	**0.636**	**0.638**	**0.589**
DG	GroupDro [60]	0.513	0.530	0.538	0.506	0.481	0.530	0.578	0.512	0.495	0.501	0.530	0.532	0.476
DG	IRM [3]	0.667	0.672	0.671	**0.678**	0.666	0.672	0.668	0.677	0.672	0.601	0.672	0.677	0.558
DG	G2DM [1]	0.610	0.616	0.618	0.609	0.593	0.616	0.606	0.611	0.592	0.607	0.616	0.616	**0.602**
DG	**IRM-FIA**	**0.670**	**0.676**	**0.678**	0.670	**0.671**	**0.696**	**0.679**	**0.693**	**0.680**	**0.614**	**0.674**	**0.680**	0.580

all three attributes. Specifically, our model improved the ES-Dice score for cup segmentation from 0.787 to 0.816 and for rim segmentation from 0.659 to 0.704 in the gender attribute. Within the race attribute, there were increments in the ES-Dice scores for cup and rim segmentations, from 0.769 to 0.787 and 0.616 to 0.636, respectively. Regarding ethnicity, the model's cup segmentation ES-Dice improved from 0.746 to 0.785, and rim segmentation from 0.630 to 0.676. Regarding specific demographic attributes, DAFormer-FIA increased the Dice score for rim segmentation in the male category within gender from 0.663 to 0.704. In the black category within race, the IoU improved from 0.329 to 0.530. In the Non-Hispanic category within ethnicity, the Dice score was enhanced from 0.746 to 0.842. This indicates the significant impact of incorporating fair identity attention into domain generalization tasks, demonstrating substantial improvements in segmentation accuracy across diverse demographic attributes.

5.2 Glaucoma Classification Results under Domain Shifts

Baselines: As shown in Table 4, the basic overall AUC of using en face fundus images (i.e. source domain) to predict glaucoma is 0.803, while the ES-AUCs for gender, race, and ethnicity are 0.795, 0.730, and 0.744, respectively. There are significant AUC performance disparities for racial subgroups with Blacks are 0.085 lower than Whites and 0.058 lower than Asians. Similarly, Hispanics achieve significantly lower AUC than Non-Hispanics. However, if using slo fundus to predict glaucoma with the pretrained source-domain model, both overall AUC and group AUCs have dramatic drops for gender, racial, and ethnic groups. As shown in Fig. 2, the group disparities generally become worse while transforming the source domain to the target domain for both segmentation and classification tasks. This suggests that source and target domain fundus images present significantly different semantic distributions. It is necessary to adopt specific learning module to minimize such a domain shift and meanwhile reduce group performance disparities.

Domain Adaptation for Classification: In parallel with our approach for segmentation, we explored two leading domain adaptation (DA) methods for classification tasks: **CGDM** [15] and **CDTrans** [81], selecting the latter, **CDTrans**, for enhancement with our novel fair identity attention module, yielding **CDTrans-FIA**. This integration aims to bolster fairness across demographics under DA.

Performances across gender, race, and ethnicity for these models are showcased in Table 4. Our proposed CDTrans-FIA stands out by securing substantial improvements across all demographic attributes. It records the highest ES-Dice scores 0.633 for gender, 0.606 for race, and 0.607 for ethnicity, demonstrating the efficacy of our proposed fair identity attention in domain adaptation, especially for enhancing glaucoma classification algorithmic fairness. Additionally, the proposed method (Table 4) significantly uplifts accuracy for minority groups; it boosts performance for blacks by 0.005 to 0.016 AUC scores, and for Hispanics by 0.01 to 0.035 AUC scores, outperforming other DA approaches. Moreover, our model exhibits notable advancements in both accuracy and fairness in the target domain, surpassing the baseline with significant improvements of around 0.1 AUC scores in overall AUC and ES-AUC.

Domain Generalization for Classification: Similarly, we selected three state-of-the-art methods including GroupDro [60], IRM [2], and G2DM [15] for the DA-based classification. Given that IRM generally achieves the best overall AUC performances for all three demographic attributes, we select IRM as the backbone architecture to incorporate our proposed fair identity attention module, termed as IRM-FIA. According to the results presented in Table 4, IRM-FIA surpasses both GroupDro and G2DM in overall AUC and ES-AUC scores, proving its superior capability in facilitating equitable glaucoma classification across diverse identity groups. In addition, IRM-FIA is generally superior to IRM in both overall AUC and ES-AUC for all three attributes, with the improvement most prominent for race. For instance, as detailed in Table 4, there was an increase of 0.024 in the overall AUC, with the AUC for Asians, Blacks, and Whites each showing gains of over 0.01 AUC score. This underscores the effectiveness of our fair identity attention module in not only boosting overall classification accuracy but also minimizing disparities among different subgroups.

6 Conclusion

To summarize, this paper focuses on addressing fairness in AI, especially in medical AI, which is essential for equitable healthcare. The issue of fairness in domain transfer, due to the fact that clinics may use varied imaging technologies, remains largely unexplored. Our work introduces FairDomain, a comprehensive study on algorithmic fairness in domain transfer tasks including domain adaptation and generalization for both medical segmentation and classification. We propose a novel plug-and-play fair identity attention module that enhances fairness by learning feature relevance through the attention mechanism according to demographic attributes in domain transfer tasks. We also create the first fairness-centric dataset with two paired imaging modalities for the same patient cohort to exclude the confounding impact of demographic distribution variation on model fairness to allow precise assessment of the impact of domain shift on model fairness. Our fair identity attention model can improve existing domain adaptation and generalization methods with better model performance accounted for fairness.

References

1. Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T.H., Mitliagkas, I.: Generalizing to unseen domains via distribution matching. arXiv preprint arXiv:1911.00804 (2019)
2. Araslanov, N., Roth, S.: Self-supervised augmentation consistency for adapting semantic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 13216–13225 (2021)
3. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
4. Asuncion, A., Newman, D.: Uci machine learning repository (2007)
5. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: Metareg: Towards domain generalization using meta-regularization. In: NeurIPS (2018)
6. Beutel, A., Chen, J., Zhao, Z., Chi, E.H.: Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075 (2017)
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
8. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Wang, Y.C.F., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1992–2001 (2017)
9. Chen, Y., Liu, F., Wang, H., Wang, C., Liu, Y., Tian, Y., Carneiro, G.: Bomd: bag of multi-label descriptors for noisy chest x-ray classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21284–21295 (2023)
10. Chen, Y., Liu, Y., Wang, C., Elliott, M., Kwok, C.F., Peña-Solorzano, C., Tian, Y., Liu, F., Frazer, H., McCarthy, D.J., et al.: Braixdet: Learning to detect malignant breast lesion with incomplete annotations. Medical Image Analysis **96**, 103192 (2024)
11. Chen, Y., Wang, H., Wang, C., Tian, Y., Liu, F., Liu, Y., Elliott, M., McCarthy, D.J., Frazer, H., Carneiro, G.: Multi-view local co-occurrence and global consistency learning improve mammogram classification generalisation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–13. Springer (2022)
12. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7892–7901 (2018)
13. Choi, S., et al.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: CVPR (2020)
14. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science advances **4**(1), eaao5580 (2018)
15. Du, Z., Li, J., Su, H., Zhu, L., Lu, K.: Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3937–3946 (2021)
16. Everett, M.J., Oakley, J.D.: Automated analysis of the optic nerve head: measurements, methods and representations (Aug 11 2015), uS Patent 9,101,293
17. Ganin, Y., et al.: Domain-adversarial training of neural networks. In: Journal of Machine Learning Research (2016)

18. Ghifary, M., et al.: Domain generalization for object recognition with multi-task autoencoders. ICCV (2015)
19. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1820–1828 (2021)
20. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
21. Hoffman, J., et al.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. In: ArXiv (2016)
22. Hoyer, L., Dai, D., Van Gool, L.: Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9924–9935 (2022)
23. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
24. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
25. Kim, M.P., Ghorbani, A., Zou, J.: Multiaccuracy: Black-box post-processing for fairness in classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 247–254 (2019)
26. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Scientific Data **9**(1), 291 (2022)
27. Kuzilek, J., Hlosta, M., Zdrahal, Z.: Open university learning analytics dataset. Scientific data **4**(1), 1–8 (2017)
28. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: Meta-learning for domain generalization. In: AAAI (2018)
29. Li, H., Wang, Y., Wan, R., Wang, S., Li, T.Q., Kot, A.: Domain generalization for medical imaging classification with linear-dependency regularization. Advances in neural information processing systems **33**, 3118–3129 (2020)
30. Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D.D., Chen, M.: Medical image classification with convolutional neural network. In: 2014 13th international conference on control automation robotics & vision (ICARCV). pp. 844–848. IEEE (2014)
31. Li, Y., et al.: Deep domain generalization via conditional invariant adversarial networks. ECCV (2018)
32. Liu, F., Chen, Y., Tian, Y., Liu, Y., Wang, C., Belagiannis, V., Carneiro, G.: Nvum: Non-volatile unbiased memory for robust medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 544–553. Springer (2022)
33. Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., Carneiro, G.: Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20706 (2022)

34. Liu, Y., Tian, Y., Wang, C., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Translation consistent semi-supervised segmentation for 3d medical images. arXiv preprint arXiv:2203.14523 (2022)
35. Lohia, P.K., Ramamurthy, K.N., Bhide, M., Saha, D., Varshney, K.R., Puri, R.: Bias mitigation post-processing for individual and group fairness. In: Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp). pp. 2847–2851. IEEE (2019)
36. Luo, Y., Khan, M.O., Tian, Y., Shi, M., Dou, Z., Elze, T., Fang, Y., Wang, M.: Fairvision: Equitable deep learning for eye disease screening via fair identity scaling (2024), https://arxiv.org/abs/2310.02492
37. Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., et al.: Fairclip: Harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12289–12301 (2024)
38. Luo, Y., Shi, M., Tian, Y., Elze, T., Wang, M.: Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20471–20482 (2023)
39. Luo, Y., Tian, Y., Shi, M., Elze, T., Wang, M.: Harvard glaucoma fairness: A retinal nerve disease dataset for fairness learning and fair identity normalization. arXiv preprint arXiv:2306.09264 (2023)
40. Luo, Y., Tian, Y., Shi, M., Pasquale, L.R., Shen, L.Q., Zebardast, N., Elze, T., Wang, M.: Harvard glaucoma fairness: A retinal nerve disease dataset for fairness learning and fair identity normalization. IEEE Transactions on Medical Imaging **43**(7), 2623–2633 (2024). https://doi.org/10.1109/TMI.2024.3377552
41. Lyu, J., Zhang, Y., Huang, Y., Lin, L., Cheng, P., Tang, X.: Aadg: automatic augmentation for domain generalization on retinal image segmentation. IEEE Transactions on Medical Imaging **41**(12), 3699–3711 (2022)
42. Matsuura, T., Harada, T.: Domain generalization using a mixture of multiple latent domains. In: AAAI (2020)
43. Mei, K., Zhu, C., Zuo, W.: Instance adaptive self-training for unsupervised domain adaptation. In: European Conference on Computer Vision. pp. 616–632. Springer (2020)
44. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12435–12445 (2021)
45. Miao, W.: Did the results of promotion exams have a disparate impact on minorities? using statistical evidence in ricci v. destefano. Journal of Statistics Education **18**(3) (2010)
46. Mitsch, C., Holzer, S., Wassermann, L., Resch, H., Urach, S., Kiss, B., Hommer, A., Vass, C., Schmidt-Erfurth, U.: Comparison of spectralis and cirrus spectral domain optical coherence tomography for the objective morphometric assessment of the neuroretinal rim width. Graefe's Archive for Clinical and Experimental Ophthalmology **257**, 1265–1275 (2019)
47. Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., Ourselin, S.: Global image registration using a symmetric block-matching approach. Journal of medical imaging **1**(2), 024003–024003 (2014)
48. Motiian, S., et al.: Unified deep supervised domain adaptation and generalization. In: ICCV (2017)

49. Mukherjee, D., Petersen, F., Yurochkin, M., Sun, Y.: Domain adaptation meets individual fairness. and they get along. Advances in Neural Information Processing Systems **35**, 28902–28913 (2022)
50. Obermeyer, Z., Nissan, R., Stern, M., Eaneff, S., Bembeneck, E.J., Mullainathan, S.: Algorithmic bias playbook. Center for Applied AI at Chicago Booth (2021)
51. Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging **42**(4), 1095–1106 (2022)
52. Pan, X., et al.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV (2018)
53. Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10398 (2022)
54. Pham, T.H., Zhang, X., Zhang, P.: Fairness and accuracy under domain generalization. arXiv preprint arXiv:2301.13323 (2023)
55. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12556–12565 (2020)
56. Quadrianto, N., Sharmanska, V., Thomas, O.: Discovering fair representations in the data domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8227–8236 (2019)
57. Ramaswamy, V.V., Kim, S.S., Russakovsky, O.: Fair attribute classification through latent space de-biasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9301–9310 (2021)
58. Roh, Y., Lee, K., Whang, S., Suh, C.: Fr-train: A mutual information-based approach to fair and robust training. In: International Conference on Machine Learning. pp. 8147–8157. PMLR (2020)
59. Ruggles, S., McCaa, R., Sobek, M., Cleveland, L.: The ipums collaboration: integrating and disseminating the world's population microdata. Journal of demographic economics **81**(2), 203–216 (2015)
60. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
61. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. pp. 746–761. Springer (2020)
62. Shi, M., Afzal, M.M., Huang, H., Wen, C., Luo, Y., Khan, M.O., Tian, Y., Kim, L., Elze, T., Fang, Y., et al.: Equitable deep learning for diabetic retinopathy detection using multi-dimensional retinal imaging with fair adaptive scaling: a retrospective study. medRxiv pp. 2024–04 (2024)
63. Shi, M., Sun, J.A., Lokhande, A., Tian, Y., Luo, Y., Elze, T., Shen, L.Q., Wang, M.: Artifact correction in retinal nerve fiber layer thickness maps using deep learning and its clinical utility in glaucoma. Translational Vision Science & Technology **12**(11), 12–12 (2023)
64. Shi, M., Tian, Y., Luo, Y., Elze, T., Wang, M.: Rnflt2vec: Artifact-corrected representation learning for retinal nerve fiber layer thickness maps. Medical Image Analysis **94**, 103110 (2024)
65. Tian, Y., Liu, F., Pang, G., Chen, Y., Liu, Y., Verjans, J.W., Singh, R., Carneiro, G.: Self-supervised pseudo multi-class pre-training for unsupervised anomaly

detection and segmentation in medical images. Medical Image Analysis **90**, 102930 (2023)
66. Tian, Y., Pang, G., Liu, F., Chen, Y., Shin, S.H., Verjans, J.W., Singh, R., Carneiro, G.: Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 128–140. Springer (2021)
67. Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., Carneiro, G.: Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2022)
68. Tian, Y., Pang, G., Liu, Y., Wang, C., Chen, Y., Liu, F., Singh, R., Verjans, J.W., Wang, M., Carneiro, G.: Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder. In: International Workshop on Machine Learning in Medical Imaging. pp. 11–21. Springer (2023)
69. Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., Wang, M.: Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In: The Twelfth International Conference on Learning Representations (2024)
70. Truong, F., Others, L.: Fredom: Fairness domain adaptation approach to semantic scene understanding. In: CVPR (2023)
71. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 7472–7481 (2018)
72. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
73. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2018)
74. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Da-dad: Depth-aware domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3681–3690 (2019)
75. Wang, C., Liu, Y., Chen, Y., Liu, F., Tian, Y., McCarthy, D., Frazer, H., Carneiro, G.: Learning support and trivial prototypes for interpretable image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2062–2072 (2023)
76. Wang, J., et al.: Learning from extrinsic and intrinsic supervisions for domain generalization. In: ECCV (2020)
77. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET Image Processing **16**(5), 1243–1267 (2022)
78. Wang, Z., Dong, X., Xue, H., Zhang, Z., Chiu, W., Wei, T., Ren, K.: Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10379–10388 (2022)

79. Wightman, L.F.: Lsac national longitudinal bar passage study. lsac research report series. (1998)
80. Xu, Q., et al.: A fourier-based framework for domain generalization. In: CVPR (2021)
81. Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 (2021)
82. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: CVPR (2020)
83. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial intelligence and statistics. pp. 962–970. PMLR (2017)
84. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340 (2018)
85. Zhang, Y., Sang, J.: Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4346–4354 (2020)
86. Zhang, Y., Liu, W., Zhan, Y., Zhang, W., Zhang, W., Duchesne, X., Tao, D.: Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10276–10285 (2021)
87. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5810–5818 (2017)
88. Zhao, H., Zhang, S., Wu, G., Moura, J.M., Costeira, J.P., Gordon, G.J.: Adversarial multiple source domain adaptation. Advances in neural information processing systems **31** (2018)
89. Zhao, X., et al.: Test-time fourier style calibration for domain generalization (2021)
90. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)
91. Zong, Y., Yang, Y., Hospedales, T.: Medfair: Benchmarking fairness for medical imaging. arXiv preprint arXiv:2210.01725 (2022)
92. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV). pp. 289–305 (2018)
93. Zou, Y., et al.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European conference on computer vision (ECCV). pp. 289–305 (2018)