Group-Conditioned Representation Modulation for Fair Skin Disease Diagnosis

Gelei Xu[1*], Haoran Fan[1*], and Yiyu Shi[1]

University of Notre Dame, Notre Dame, IN 46556, USA
{gxu4, hfan2, yshi4}@nd.edu

Abstract. Training models on heterogeneous medical data often results in performance disparities across demographic groups. In skin disease diagnosis, these disparities are particularly challenging since sensitive attributes can correlate with clinically relevant factors, making traditional suppression-based fairness interventions potentially harmful to diagnostic accuracy. This limitation motivates us to revisit this problem through a multi-objective optimization perspective, where each group defines a performance objective within a shared backbone. When aggregate performance is optimized under shared parameterization, the learned representations may under-express group-specific characteristics due to shared capacity allocation. To mitigate such effect, we propose a group-specific interference masking framework that enables conditional parameter utilization without altering the backbone architecture. The method identifies representation components that lack stable alignment with a given group and selectively modulates them without introducing additional trainable parameters. Experiment results demonstrate that the proposed approach consistently reduces inter-group performance gaps while maintaining competitive overall accuracy. Our code is available at link.

Keywords: Skin Diseases · Fairness · Masking.

1 Introduction

AI-based systems have achieved strong performance in skin disease diagnosis, yet persistent disparities remain across demographic groups such as skin type and sex [10,5]. These disparities raise concerns about reliability in clinical deployment and have motivated extensive research on fairness in medical imaging [25,20]. Existing interventions operate at different stages of the learning pipeline and primarily aim to mitigate the influence of sensitive attributes. Pre-processing methods adjust data distributions or learn attribute-decorrelated representations prior to model training [27,13,22]. In-processing methods incorporate adversarial objectives or regularization terms to reduce the encoding of group-correlated signals [6,15,28]. Post-processing methods modify decision thresholds or prediction outputs to enforce parity constraints across demographic groups [7,9,18].

* Equal contribution.

Many of these approaches assume that performance disparities arise from the presence of sensitive attribute information in representations or predictions, and that fairness can be improved by reducing their influence. This assumption is reasonable in applications such as recidivism risk assessment or income prediction, where sensitive attributes are not causally related to the target outcome [8,1]. In dermatology, however, attributes such as skin type are associated with clinically relevant factors including disease prevalence and lesion morphology. For example, skin type encodes diagnostic signals such as UV susceptibility [3,24], which is crucial for clinical decision-making. Attenuating such signals can therefore compromise predictive accuracy. Empirical evidence further suggests that fairness gains achieved under this strategy are often achieved not by improving disadvantaged groups but by uniformly degrading performance across all groups, which is undesirable in high-stakes clinical settings [7].

We therefore revisit fairness in skin disease diagnosis from a multi-objective optimization perspective. Under shared parameterization, joint training must satisfy multiple group-specific objectives with a single set of parameters, which can produce compromise representations that under-express group-specific characteristics. This view parallels insights from multi-task learning [29,17], where shared capacity must be allocated across competing objectives. Motivated by this perspective, we focus on strengthening group-conditioned expressiveness within a shared backbone. Specifically, we aim to enable each group to utilize shared parameters in ways that better reflect its statistical structure while reducing representational interference from uniform activation patterns, which can otherwise lead to suboptimal performance even when overall objectives are satisfied.

To operationalize this idea, we propose GIM, a **G**roup-specific **I**nterference **M**asking framework built on a shared backbone. The method introduces a binary mask for each group to modulate intermediate activations, enabling group-conditioned expressiveness without altering the backbone parameters. Since convolutional channels capture distinct feature patterns [12], they provide a natural granularity for conditional modulation [14]. For each group, we identify channels whose activations lack stable alignment with that group and selectively suppress them during fine-tuning. The entire procedure operates on a pretrained model with all backbone weights remain shared accress groups, and introduces no additional trainable parameters or meaningful inference overhead.

Experiments on Fitzpatrick-17k and ISIC 2019 show that the proposed method reduces inter-group performance gaps while maintaining competitive overall accuracy. The performance gains are more substantial for groups that exhibit lower baseline accuracy under standard joint training. This pattern suggests that group-conditioned parameter utilization may benefit groups whose representations are less effectively expressed in a fully shared setting, offering a promising direction for improving fairness within fixed model capacity.

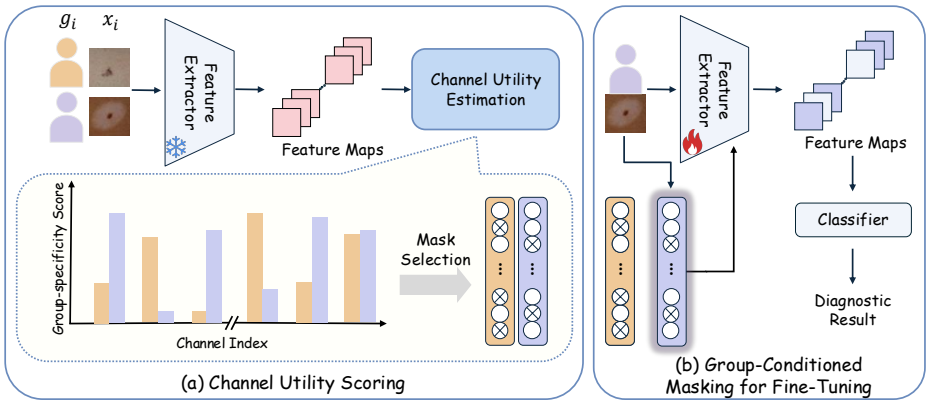

Fig. 1: Framework overview: (a) computing per-group channel utility scores and deriving binary masks from a frozen backbone, and (b) applying the masks during fine-tuning to modulate shared representations within a shared model.

2 Methodology

2.1 Problem Formulation and Framework Overview

We consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i, g_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ denotes an input image, $y_i \in \mathcal{Y}$ is the ground-truth class label, and $g_i \in \mathcal{G}$ represents the group attribute (e.g., skin type). A pretrained model $f_\theta : \mathcal{X} \to \mathcal{Y}$ with parameters θ is trained jointly on samples from all groups to produce predictions $\hat{y}_i = f_\theta(\mathbf{x}_i)$.

This setting implicitly requires a single set of parameters to satisfy multiple group-specific objectives. As a result, the learned representations may reflect compromises that do not fully capture group-dependent characteristics, leading to suboptimal performance across groups. We therefore aim to enhance group-wise performance within a shared backbone by enabling group-conditioned parameter utilization, so that each group can more effectively express its statistical structure under fixed model capacity. For clarity, we focus on the binary case where $g_i \in \{0, 1\}$, corresponding to two demographic groups. The formulation is extensible to multi-group settings.

The overview of our method is shown in Figure 1. First, given a pretrained backbone, we analyze intermediate representations to estimate group-conditional utility scores for ecach channels. These scores quantify the alignment consistency of each channel with respect to each group under the frozen model. Second, based on these scores, we construct group-specific masks that modulate channel activations during group-conditioned fine-tuning. The masks selectively suppress channels with low utility while preserving shared parameters across groups. This design preserves the original parameters and architecture while enabling lightweight group-specific adaptation within a single shared model. In the following section, we will dive into the implementation details of each stage.

2.2 Channel Utility Scoring

To enable group-conditioned parameter utilization within a shared backbone, we require a mechanism to identify representations that are consistently useful for a given group. Our objective is to quantify, for each group, the extent to which an intermediate representation unit contributes stable and group-consistent activation patterns. We refer to this quantity as group-conditional utility.

We adopt convolutional channels as the unit of conditional modulation. Channel-level intervention provides a structured and computationally efficient granularity for activation control. Prior work has shown that individual channels capture distinct feature patterns and that structured channel modulation can meaningfully affect group-wise performance [14]. Compared to weight-level masking, channel-level modulation avoids excessive parameter fragmentation and preserves computational efficiency. Compared to layer-level control, it offers finer resolution in representation adjustment.

Let f_θ denote a pretrained convolutional network. Consider a target convolutional layer with C output channels. For input sample i, let $\mathbf{f}_{i,k} \in \mathbb{R}^{H\times W}$ denote the feature map of channel k. We reshape $\mathbf{f}_{i,k}$ into a vector and apply ℓ_2 normalization to obtain $\hat{\mathbf{f}}_{i,k} \in \mathbb{R}^{HW}$. Normalization ensures that subsequent similarity comparisons reflect directional consistency rather than activation magnitude.

For each group g, we summarize the typical activation pattern of channel k using a prototype defined as

$$\boldsymbol{\mu}_k^g = \mathbb{E}_{i\in\mathcal{D}_g}\left[\hat{\mathbf{f}}_{i,k}\right], \tag{1}$$

where $\mathcal{D}_g$ denotes samples from group g. The prototype provides a stable summary of group-wise channel behavior and reduces sensitivity to sample-level noise. In practice, $\boldsymbol{\mu}_k^g$ is computed as the empirical mean over $\mathcal{D}g$ using the frozen $f\theta$.

We then define a group-specificity score that evaluates the alignment consistency of channel k for group g

$$S_k^g = \mathbb{E}_{i\in\mathcal{D}_g}\left[\log \frac{\exp\left(\langle\hat{\mathbf{f}}_{i,k}, \boldsymbol{\mu}_k^g\rangle\right)}{\exp\left(\langle\hat{\mathbf{f}}_{i,k}, \boldsymbol{\mu}_k^g\rangle\right) + \exp\left(\langle\hat{\mathbf{f}}_{i,k}, \boldsymbol{\mu}_k^{1-g}\rangle\right)}\right], \tag{2}$$

where $\langle\cdot,\cdot\rangle$ denotes cosine similarity and $\boldsymbol{\mu}_k^{1-g}$ is the prototype of the other group. This score measures whether activations of channel k for samples in group g are consistently closer to the group-g prototype than to the competing prototype. Higher values indicate stable group-aligned behavior, while lower values indicate weaker or inconsistent alignment. The score is computed on the frozen pretrained model so that channel evaluation is independent of subsequent fine-tuning.

2.3 Group-Conditioned Masking for Fine-Tuning

The group-conditional utility scores provide a ranking of channels according to their alignment consistency within each group. Channels with low S_k^g exhibit

weak or unstable alignment for group g and are therefore less likely to contribute reliable group-specific signals. We use this ranking to modulate how shared capacity is utilized during inference and fine-tuning.

For each group g, we sort channels by S_k^g and construct a binary mask $\mathbf{m}^g \in \{0,1\}^C$ that deactivates a fraction $\lfloor r \cdot C \rfloor$ of channels with the lowest scores, where $r \in [0,1]$ denotes the mask ratio. This operation does not remove channels globally. Instead, it conditionally suppresses channels that exhibit limited utility for the current group while retaining their contribution for other groups. In this way, shared parameters are selectively allocated across groups without introducing additional model components.

Let $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ denote the feature tensor at the selected layer. For samples belonging to group g, the masked feature map is $\tilde{\mathbf{F}}^g = \mathbf{m}^g \odot \mathbf{F}$, where $\odot$ denotes channel-wise multiplication. The masks are binary and fixed once constructed.

During group-conditioned fine-tuning, the mask $\mathbf{m}^g$ is applied to both forward activations and backward gradients for samples in group g. This ensures that channels suppressed for a given group do not contribute to its representation or receive gradient updates from its samples. All backbone parameters remain shared across groups, and no additional trainable parameters are introduced. The intervention operates through conditional activation patterns within a single shared model. The mask ratio r controls the extent of conditional modulation. Larger values suppress more low-utility channels and increase group-specific capacity allocation, whereas excessive masking can restrict shared representational capacity and reduce overall accuracy. We select r using validation data and retain the configuration that maintains stable predictive performance across groups.

3 Experiments

3.1 Datasets and Metrics

Dataset. We evaluate our method on Fitzpatrick-17k [10] and ISIC 2019 [5], two widely used benchmarks for skin disease classification. Fitzpatrick-17k contains 16,577 images spanning 114 skin conditions and provides Fitzpatrick skin type annotations. Following prior work, we treat skin type as the sensitive attribute and group types 1–3 as light skin and types 4–6 as dark skin. ISIC 2019 includes 25,331 dermoscopic images across eight diagnostic categories. For this dataset, we consider patient sex as the sensitive attribute.

Metrics. We evaluate model performance using precision, recall, and macro-averaged F1. For fairness metrics, following [23], we report Equal Opportunity (EOpp) and Equalized Odds (EOdds). We also report FATE [14] to quantify the overall trade-off between accuracy and fairness. A higher FATE indicates a better trade-off for sensitive groups.

Implementation Details. Following the experimental protocol established in prior work [23,4,14], we adopt VGG-11 for Fitzpatrick-17k and ResNet-18 for ISIC 2019 as backbones to ensure direct comparability and adopt the the data-augmentation policy. All images are resized to 128×128 before training. In our

Table 1: Accuracy and fairness results on the Fitzpatrick-17k dataset. Bold fonts and underlined fonts denote the best and second-best performance, respectively.

		Accuracy			Fairness	
Method	Group	Precision	Recall	F1-score	Eopp ↓ / FATE ↑	Eodd ↓ / FATE ↑
VGG-11 [19]	Dark	0.563	**0.581**	<u>0.546</u>	0.361 / 0.0000	0.182 / 0.0000
	Light	0.482	0.495	0.473		
	Avg. ↑	0.523	0.538	0.510		
HSIC [15]	Dark	0.548	0.522	0.513	0.331 / 0.0645	0.166 / 0.0683
	Light	0.513	0.506	0.486		
	Avg. ↑	0.530	0.515	0.500		
DomainIndep [21]	Dark	0.559	0.540	0.512	0.323 / 0.1268	0.161 / 0.1370
	Light	**0.541**	**0.529**	<u>0.512</u>		
	Avg. ↑	**0.550**	0.534	<u>0.521</u>		
FairPrune [23]	Dark	<u>0.567</u>	0.519	0.507	0.330 / 0.0329	0.165 / 0.0405
	Light	0.496	0.477	0.459		
	Avg. ↑	0.531	0.498	0.483		
SCP [14]	Dark	**0.568**	<u>0.576</u>	**0.547**	<u>0.278</u> / <u>0.2495</u>	<u>0.139</u> / <u>0.2559</u>
	Light	0.499	0.504	0.492		
	Avg. ↑	0.533	<u>0.540</u>	0.520		
FaMI [16]	Dark	0.560	0.552	0.525	0.289 / 0.1857	0.135 / 0.2445
	Light	0.502	0.498	0.489		
	Avg. ↑	0.535	0.528	0.503		
GIM (Ours)	Dark	**0.568**	0.558	0.543	**0.253** / **0.3423**	**0.127** / **0.3453**
	Light	<u>0.531</u>	<u>0.527</u>	**0.513**		
	Avg. ↑	**0.550**	**0.543**	**0.532**		

framework, S_k^g is computed from the final convolutional layer of VGG-11 and the `layer4` output of ResNet-18, where higher-level semantic features are most directly represented.

Baselines. We compare our framework against various bias mitigation baselines. Vanilla is the model trained directly on the ResNet-18 or VGG-11 backbone. HSIC [15] penalized statistical dependence between learned features and the sensitive variable. DomainIndep learns group-specific classifiers with shared parameter [21]. FairPrune [23] improves group fairness by pruning parameters whose importance differs across sensitive groups to reduce inter-group performance gaps. SCP [14] performs channel pruning to remove disparity-inducing channels. FaMI [16] minimizing mutual information between learned representations and sensitive attributes through a gradient-based regularization. We refer to our method as **GIM** (**G**roup-specific **I**nterference **M**asking).

Results on Fitzpatrick-17k Dataset. Table 1 presents accuracy and fairness metrics on the Fitzpatrick-17k dataset. The baseline VGG-11 exhibits substantial disparity between the dark and light groups, with Eopp of 0.361 and Eodd of 0.182. Existing fairness methods reduce these gaps to varying extents, although reductions in disparity are generally accompanied by reduction in average F1-score. DomainIndep achieves strong overall accuracy and competitive fairness performance by incorporating group-specific modules. SCP, another competitive baseline, also applies channel modulation to improve fairness. However, as its

Table 2: Accuracy and fairness results on the ISIC 2019 dataset. Bold fonts and underlined fonts denote the best and second-best performance, respectively.

		Accuracy			Fairness	
Method	Group	Precision	Recall	F1-score	Eopp ↓ / FATE ↑	Eodd ↓ / FATE ↑
ResNet-18 [11]	Female	<u>0.802</u>	<u>0.717</u>	**0.754**	0.031 / 0.0000	0.019 / 0.0000
	Male	0.752	0.711	0.729		
	Avg. ↑	<u>0.777</u>	0.714	<u>0.742</u>		
HSIC [15]	Female	0.744	0.660	0.696	0.042 / -0.4114	0.020 / -0.1092
	Male	0.718	0.697	0.705		
	Avg. ↑	0.731	0.679	0.700		
DomainIndep [21]	Female	0.729	**0.747**	0.734	0.086 / -1.8065	0.042 / -1.2428
	Male	0.725	0.694	0.702		
	Avg. ↑	0.727	**0.721**	0.718		
FairPrune [23]	Female	0.776	0.711	0.734	0.026 / 0.1411	0.014 / 0.2429
	Male	0.721	**0.725**	0.720		
	Avg. ↑	0.748	<u>0.718</u>	0.727		
SCP [14]	Female	0.787	0.701	0.736	<u>0.015</u> / <u>0.5080</u>	**0.006** / **0.6761**
	Male	**0.765**	<u>0.712</u>	<u>0.735</u>		
	Avg. ↑	0.776	0.707	0.736		
FaMI [16]	Female	0.779	0.705	0.734	0.024 / 0.2042	0.013 / 0.2942
	Male	0.745	0.708	0.718		
	Avg. ↑	0.758	0.706	0.726		
GIM (Ours)	Female	**0.803**	0.713	<u>0.752</u>	**0.013** / **0.5833**	<u>0.009</u> / <u>0.5290</u>
	Male	<u>0.764</u>	<u>0.712</u>	**0.736**		
	Avg. ↑	**0.783**	0.713	**0.744**		

approach is based on suppressing sensitive attribute information, this design may contribute to its relatively lower accuracy compared with our method. GIM achieves the lowest Eopp and Eodd while also obtaining the highest average Precision, Recall, and F1-score. Compared with the second-best baseline SCP, GIM reduces FATE(Eopp) by 37.2% and FATE(Eodd) by 34.9%. These findings support the effectiveness of conditional representation modulation within a shared backbone for improving group-wise performance balance.

Results on ISIC 2019 Dataset. As shown in table 2, fairness-oriented methods in this settings often lead to a decrease in overall accuracy, and the vanilla ResNet18 remains among the top-performing models in terms of accuracy, indicating the difficulty of improving fairness without affecting predictive performance. Our method achieves the lowest Eopp and the second-lowest Eodd among all compared approaches, while also obtaining the highest average Precision and F1-score. In addition, it improves performance for the male group while maintaining comparable performance for the female group. These results suggest that conditional representation modulation can preserve stable predictive performance while further enhancing group-wise balance in a shared backbone.

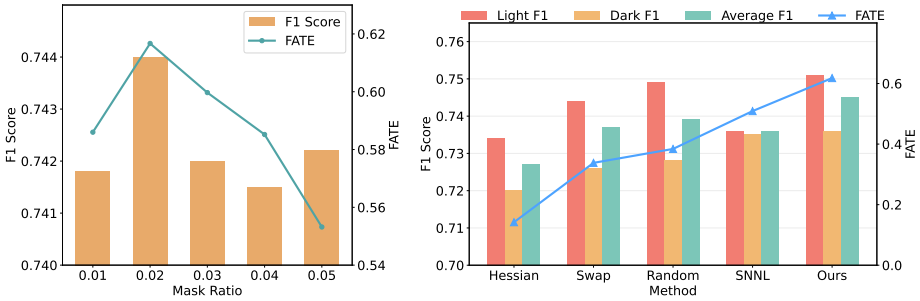

Fig. 2: Ablation studies on ISIC 2019: (a) varying the mask ratio r; (b) comparing mask selection strategies.

3.2 Ablation Study

We conduct ablation studies on the ISIC 2019 dataset to examine the effect of the mask ratio and the mask selection strategy. We report the average macro F1-score for accuracy and the FATE score computed with Eopp.

Effect of Mask Ratio We vary the mask ratio r from 1% to 5%, as shown in Figure 2 (a). Results show that compare to vanilla situation, introducing a small degree of masking improves both F1-score and FATE. The best balance is achieved at $r = 2\%$, where the model attains the highest F1-score together with the peak FATE value. Further increasing r leads to performance saturation and slight degradation at $r = 4\%$, indicating that excessive masking begins to suppress channels that contribute to predictive performance. These results reflect a trade-off in which larger values of r increase group-specific capacity allocation by suppressing more low-utility channels, while excessive masking restricts shared representational capacity and degrades overall accuracy.

Effect of Mask Selection Strategy We further evaluate different mask selection strategies, as shown in Figure 2(b). In addition to the proposed approach, we compare with a Hessian-based strategy [23], random masking, swap masking, which applies the mask learned for one group to samples from the other group, and SNNL [14], which selects channels based on representation similarity. Compared with non-group-specific strategies, our method achieves a more favorable trade-off between F1-score and FATE across all evaluated approaches. This performance may be attributed to the effectiveness and robustness of the proposed channel utility score. Furthermore, the swap setting results in clear degradation in both F1-score and FATE, and performs worse than random masking. This observation suggests that the learned masks encode group-dependent information rather than reflecting arbitrary channel selection.

4 Discussion and Conclusion

Although GIM shows potential to improve the accuracy-fairness trade-off, several aspects warrant further investigation. First, the mask ratio affects the balance between group-conditioned specialization and shared capacity; future work can explore adaptive mask-selection strategies. Second, our experiments focus on binary group settings; extending GIM to multi-group and intersectional settings through subgroup discovery or compositional group modeling would improve its clinical applicability [2]. Finally, GIM currently modulates late convolutional channels in CNN backbones. Future work should examine how group-conditioned modulation can be defined for non-convolutional architectures, including transformer-based foundation models and agentic AI systems [26].

To conclude, this paper identifies internal representational conflict at the channel level as a primary source of suboptimal performance and disparity in models trained on heterogeneous data. Our Group-specific Channel Masking framework was proposed to directly address this issue, by identifying and masking the specific, detrimental channels responsible for inter-group conflict. Experimental validation on two skin disease datasets demonstrates that our framework significantly improves accuracy and reduces performance disparities.

Disclosure of Interests. The authors declare no competing interests.

References

1. Asuncion, A., Newman, D., et al.: Uci machine learning repository (2007)
2. Bissoto, A., Hoang, T.D., Flühmann, T., Sun, S., Baumgartner, C.F., Koch, L.M.: Subgroup performance analysis in hidden stratifications. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 594–603. Springer (2025)
3. Caini, S., Gandini, S., Sera, F., Raimondi, S., Fargnoli, M.C., Boniol, M., Armstrong, B.K.: Meta-analysis of risk factors for cutaneous melanoma according to anatomical site and clinico-pathological variant. European journal of cancer **45**(17), 3054–3063 (2009)
4. Chiu, C.H., Chung, H.W., Chen, Y.J., Shi, Y., Ho, T.Y.: Toward fairness through fair multi-exit framework for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 97–107. Springer (2023)
5. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
6. Deng, W., Zhong, Y., Dou, Q., Li, X.: On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In: International Conference on Information Processing in Medical Imaging. pp. 158–169. Springer (2023)
7. Duan, Y., Xu, G., Shi, Y., Lemmon, M.: The cost of local and global fairness in federated learning. In: International Conference on Artificial Intelligence and Statistics. pp. 4186–4194. PMLR (2025)

8. Flores, A.W., Bechtel, K., Lowenkamp, C.T.: False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation **80**, 38 (2016)
9. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 329–338 (2019)
10. 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)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hesse, R., Fischer, J., Schaub-Meyer, S., Roth, S.: Disentangling polysemantic channels in convolutional neural networks. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4799–4803 (2025)
13. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and information systems **33**(1), 1–33 (2012)
14. Kong, Q., Chiu, C.H., Zeng, D., Chen, Y.J., Ho, T.Y., Hu, J., Shi, Y.: Achieving fairness through channel pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2024)
15. 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)
16. Sadri, A.R., DeSilvio, T., Viswanath, S.E.: Mutual information regularization for fairness-aware deep imaging representations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 399–409. Springer (2025)
17. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. Advances in neural information processing systems **31** (2018)
18. Shen, X., Szeto, J., Li, M., Huang, H., Arbel, T.: Exposing and mitigating calibration biases and demographic unfairness in mllm few-shot in-context learning for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 226–236. Springer (2025)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Wang, P., Tong, L., Wu, J., Liu, J., Liu, Z.: Fair-moe: Medical fairness-oriented mixture of experts in vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 186–196. Springer (2025)
21. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8919–8928 (2020)
22. Wang, Z., Yang, S., Chen, W., Zhang, Z., Wang, M., Zhou, F., Tian, Y., Wang, M., Zhao, Y., Zheng, Y., et al.: Fairness-aware vcdr-controlled generation for glaucoma diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 256–266. Springer (2025)

23. Wu, Y., Zeng, D., Xu, X., Shi, Y., Hu, J.: Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 743–753. Springer (2022)
24. Xu, G., Duan, Y., Liu, Z., Li, X., Jiang, M., Lemmon, M., Jin, W., Shi, Y.: Incorporating rather than eliminating: Achieving fairness for skin disease diagnosis through group-specific expert. arXiv preprint arXiv:2506.17787 (2025)
25. Xu, G., Duan, Y., Xia, J., Chiu, C.H., Lemmon, M., Jin, W., Shi, Y.: Rethinking fairness in medical imaging: Maximizing group-specific performance with application to skin disease diagnosis. Medical Image Analysis p. 103950 (2026)
26. Xu, G., Li, X., Chen, Y., Duan, Y., Wu, S., Yu, H., Chiu, C.H., Ni, J., Tang, N., Li, T.J.J., et al.: A comprehensive survey of ai agents in healthcare. Journal of Biomedical Informatics p. 105045 (2026)
27. Xu, Z., Zhao, S., Quan, Q., Yao, Q., Zhou, S.K.: Fairadabn: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 307–317. Springer (2023)
28. Yi, C., Xiong, Z., Qi, Q., Wei, X., Bathla, G., Lin, C.L., Mortazavi, B.J., Yang, T.: Adfair-clip: Adversarial fair contrastive language-image pre-training for chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2025)
29. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)