Toward Fair Ultrasound Computing Tomography: Challenges, Solutions and Outlook

ABSTRACT

Medical image reconstruction plays a pivotal role in early cancer detection, which can significantly enhance both the quality and longevity of a patient's life through timely treatment. However, the extent to which current image reconstruction methods accurately represent all populations, and whether they underperform for certain groups, remains largely unexplored. In this work, we will examine the deep learning (DL)–based approach to image reconstruction and its associated fairness concerns. Initially, our experiments confirmed the unfairness's presence. Subsequently, by addressing the issue from two perspectives, we gained valuable insights, which deepened our understanding of the problem. To assess a model's fairness, it's crucial to evaluate it from various perspectives, as relying on a single metric can often yield misleading results.

KEYWORDS

USCT, Image Reconstruction, Fairness, InversionNet

ACM Reference Format:

. 2024. Toward Fair Ultrasound Computing Tomography: Challenges, Solutions and Outlook. In *Great Lakes Symposium on VLSI 2024 (GLSVLSI '24), June 12–14, 2024, Clearwater, FL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3649476.3660387

1 INTRODUCTION

Image reconstruction refers to the process of generating or restoring an image from a set of observations or measurements. This is a common challenge in fields such as medical imaging, astronomy, computer vision, and photography. The aim is to produce a high-quality image that is as close as possible to the original or intended image, despite limitations or imperfections in the data collection process. In medical imaging, for example, techniques like computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound computed tomography (USCT) rely heavily on image reconstruction algorithms to convert raw data into a visual form that can be interpreted by medical professionals. Breast cancer is the most common type of cancer among women worldwide. According to the latest cancer statistics, breast cancer accounts for 31% of female cancer cases and is the second leading cause of cancer death among women [1]. Ultrasound Computer Tomography (USCT) which uses the reflection, transmission, and attenuation of ultrasound signals to describe structural information has attracted interest as a non-invasive, radiation-free imaging method in breast cancer detection.

Previous works [13, 18, 20, 36] provided a variety of computational methods to complete quantitative reconstructions (SOS map) of a tissue's acoustic properties from USCT data. However, they all miss the fairness issue of whether the methods provided in the paper are consistently valid for all groups of people. Paper [7] also raises the fairness question, where the data used are MRI data. This paper demonstrates that it is essential to consider fairness in DL algorithms, particularly in terms of demographic characteristics. The model performance is evaluated using image reconstruction metrics and findings reveal statistically significant performance biases between the gender and age subgroups.

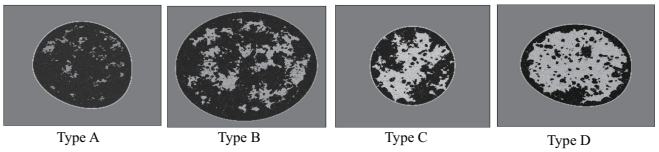


Figure 1: Four examples of the breast types: Type A (almost all fatty), Type B (scattered fibroglandular density), Type C (heterogeneous density), and Type D (extremely dense).

In our work, we delve into issues of unfairness by examining the USCT breast cancer dataset, where we identify performance disparities among different groups. To address these disparities, we classify the affected groups into 'unprivileged' and 'privileged' based on established fairness criteria. Our approach to mitigating these unfairness issues encompasses strategies within both data and training dimensions. From a data perspective, the most direct solution appears to be increasing the representation of the unprivileged group to achieve a balanced dataset. However, this proves to be impractical. Instead, we experiment with repeated sampling from smaller groups, aiming to enrich the model's exposure to these underrepresented categories. On the training side, we employ a controlled loss function technique, assigning weights greater than 1 to unprivileged group samples. This method effectively amplifies their impact during the loss calculation phase of training. To assess the effectiveness and fairness of our model, we apply various metrics. These evaluations help us determine the consistency of our interventions in promoting fairness within the model's performance.

In the rest of the work, section 2 reviews the related work and provides motivation; section 3 defines the problem, and presents two methods. Experiments are reported in section 4 and the conclusion is given in section 5.

2 RELATED WORK AND MOTIVATION

Fairness in machine learning addresses efforts to mitigate bias in automated decision-making processes using machine learning models. Decisions are deemed unfair if they rely on discriminatory attributes such as gender, ethnicity, sexual orientation, and disability. Since 2016, research into fairness in natural language processing [4, 6, 16] has surged, and more recently, studies on fairness in ML-based computer vision [3, 29, 33, 35, 37] have emerged. In general, all the solutions proposed in the above research works can be divided into three categories: (1) Preprocessing, such as data augmentation [10, 22, 33], data clean [21], and model optimization [23]; (2) In-processing, like adversarial training [24, 31, 32], weighted loss function [8, 15], fair batch [19] and fair training strategies

[29, 34]; and (3) Post-processing, such as model fusing [12, 14, 38] and thresholding [17, 27].

The growing emphasis on fairness in both NLP and CV has also sparked questions about its role in image reconstruction. Work [7] provides insights into fairness in DL-based image reconstruction and aims to improve equity in medical AI applications. They identify existing bias in performance between gender and age groups using the publicly available dataset and discuss the factors that may impact the fairness of the algorithm, including inherent characteristics and spurious correlations. Paper [5] indicates that women with dense breasts have a higher risk of developing breast cancer at the same time, due to a reduction in the contrast between the tumor and surrounding breast tissue, mammography exhibits low sensitivity and specificity in dense breast tissue. This characteristic makes it more difficult to detect high-density breast tumors, while low-density ones are easier to detect.

2.1 Fairness definitions and rules

Among the various definitions of fairness, since we study fairness for different subgroups, we consider only group fairness in our analysis. Group fairness, often discussed within the context of machine learning, statistics, and social policy, aims to ensure equitable treatment and outcomes across different groups defined by certain attributes like race, gender, age, or any other demographic factors. Statistical Parity Difference (SPD) [2] is a fairness metric used to measure the level of group fairness in decision-making processes, particularly in models and algorithms. It quantifies the difference in the probability of positive outcomes between two groups, aiming to identify biases that might favor one group over another. SPD is closely related to the concept of demographic parity, one of the foundational notions in the study of algorithmic fairness. This measure must be equal to 0 to be fair.

$$SPD = P\left(\hat{Y} = 1|A = unprivileged\right) - P\left(\hat{Y} = 1|A = privileged\right) \tag{1}$$

where $\hat{Y}$ are the model predictions and A is the group of the sensitive attribute.

Disparate Impact (DI) [2] is a concept primarily used in the legal and human resources fields, particularly in the United States, to address indirect discrimination in employment, housing, lending, and other areas. Unlike disparate treatment, which involves intentional discrimination, disparate impact refers to policies, practices, or criteria that appear neutral but disproportionately affect members of a protected class more than others, without a justified business necessity.

$$DI = P\left(\hat{Y} = 1|A = unprivileged\right) / P\left(\hat{Y} = 1|A = privileged\right) \tag{2}$$

where $\hat{Y}$ are the model predictions and A is the group of the sensitive attribute.

2.2 Breast types and fairness issue in USCT

In Fig. 1 we plotted the four examples of the breast based on the four different levels of breast density (percentage of fibroglandular tissue) defined according to the American College of Radiology's (ACR) Breast Imaging Reporting and Data System (BI-RADS) [9]: (A) almost all fatty breasts, (B) breasts with scattered fibroglandular density, (C) breasts with heterogeneous density, and (D) extremely dense breasts.

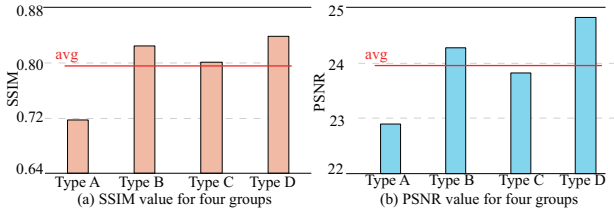


Figure 2: Two important matrics SSIM and PSNR for different groups: the red line indicates the average for the whole dataset, it is obvious that type A is always below the average under the two matrics which will be considered as the unprivileged group. Others are automatically categorized into privileged groups

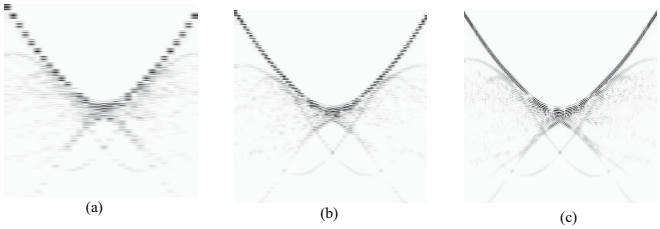


Figure 3: Increase the density of the input waveform to show the model performance SSIM and PNSR.

In Fig. 2 we have experimented with the data, and through our experiments we have found that the model performs differently on each of the subgroups. For images in category A, the image quality after reconstruction is lower in both SSIM and PSNR, and we categorize such categories as a unprivileged. The unprivileged category is not defined by numerical scarcity but by poor modeling results.

3 METHOD

In this section we will introduce three methods to optimize the fairness, the first one is the density increment which needs a higher density equipment in the data capture. The second one is the two-stage training which splits the input into two groups privileged and unprivileged, then does the upsampling for the unprivileged group. The third one is the fairness-weighted loss, like the previous method, it also splits the input data, and it will have an extra ratio for the unprivileged group.

3.1 Increase the density of the input waveform

Since the USCT waveform data will involve hardware devices like receivers and signal sources during the acquisition process when the number of signal sources and receivers is bigger, the data density of the acquired samples is higher, which corresponds to a better overall result after image reconstruction. If fewer signal sources

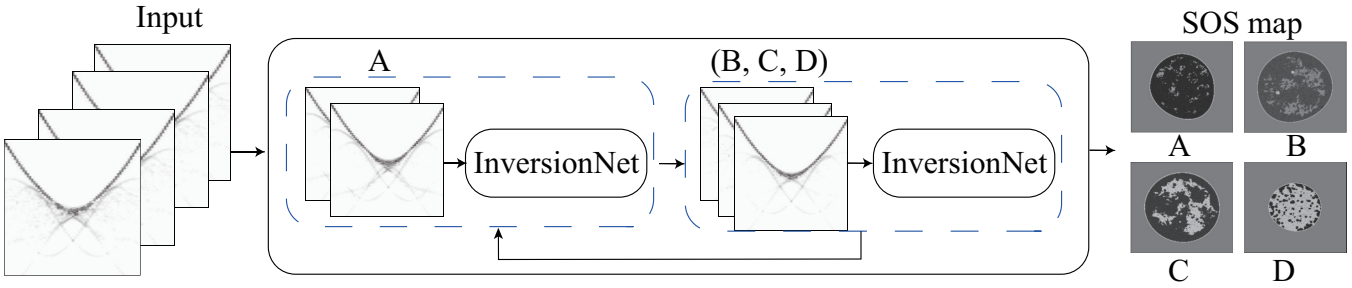


Figure 4: Two-stage training framework: for the input data, four types are mixed but during the training, in the first epoch, type A will be trained then in the second epoch, types B, C, and D will be trained, and finally go back to the A

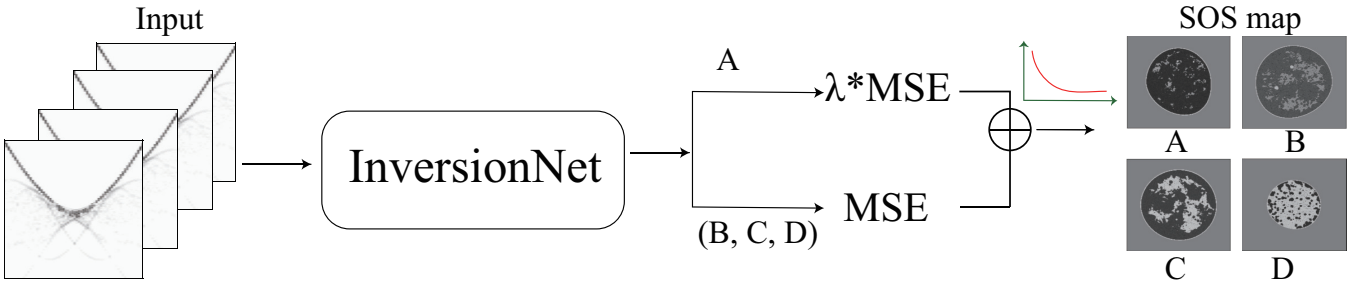


Figure 5: Fairness weighted loss: weighted loss will be applied during the training process, it means the unprivileged type A will gain an extra ratio when calculating the loss, this will magnify the importance of type in the overall dataset

and receivers are used, then the density of the acquired data is low and the image reconstruction is more difficult. By this property, we try to use a larger number of signal sources and receivers in our experiments, trying to solve the fairness problem by increasing the density of the collected data. Fig. 3 provides examples that illustrate clear disruptions in the signal waveform in Fig. 3 (a), indicating discontinuity. However, using higher-density equipment enhances our observation of the waveform, revealing a reduction in these disruptions and an improvement in continuity in Fig. 3 (c).

3.2 Two-stage training

In our preliminary experiments, we observed significant variability in the model's performance across the four categories. This variability is likely attributable to inconsistencies within the datasets for each category. To address this, we aim to enhance the consistency and alignment of our final training datasets by selectively repeating some samples during the training process. In Fig. 4, a two-stage training strategy is designed. Four types of data will be used as the input but when for the training, the four types of data will be split into privileged and unprivileged and then trained in order, in cycles of 2, with the first training targeting class A, the second training targeting class B, C, D. After the completion of such training for a complete cycle, the training will start again from A cycle. Due to the different amounts of data in each group, the group with less data will repeat the training for a sample.

Given that Category A underperforms and is thus considered an unprivileged group, while Categories B, C, and D form the privileged, we initially mix all four types of training data. The first step involves separating these categories. Training begins by focusing on Category A to update the model, followed by sequential training with Categories B, C, and D, each updating the model's parameters in turn, forming a cycle. This cycle repeats from unprivileged to privileged, with model weights updated periodically until convergence is reached. Due to varying class sizes, classes with fewer samples may undergo repeated sampling, using the largest class size as a reference. Upon completing the training, we use the test dataset to generate the corresponding Speed of Sound (SOS) plots from the waveforms.

3.3 Fairness-weighted loss

In Fig. 5, we did not repeatedly sample the data during training, but used a weighted loss function. It will help us to magnify the role of unprivileged groups.

The concept of fairness-weighted loss in machine learning is designed to address biases and inequalities in the training process, aiming to ensure that the model performs equitably across different groups or classes, especially when data representation is imbalanced. This approach adjusts the loss function based on the fairness criterion, assigning different weights to the loss for different groups to mitigate bias. For example, if a certain group is underrepresented or consistently misclassified, the loss associated with errors for that group might be increased to incentivize the model to learn better representations for that group. This helps in achieving a more balanced performance across all groups, promoting fairness in model predictions. As in the first method, the dataset is still divided into a unprivileged group and a privileged group, and the unprivileged

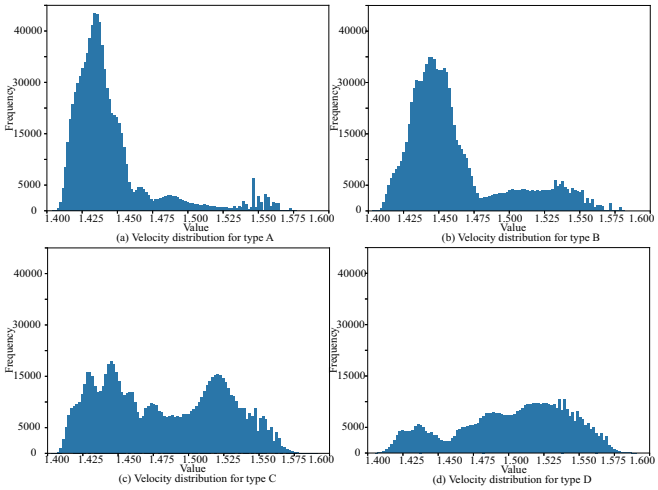


Figure 6: The summary of each type's SOS map

group will receive an additional expansion factor λ when calculating the loss function as a way of expanding their influence on the model parameter updates. The coefficient is computed as follows:

$$\lambda = \frac{N_{privileged}}{N_{unprivileged}} \tag{3}$$

$N_{privileged}$ represents the number of samples in the privileged group. $N_{unprivileged}$ represents the number of samples in the unprivileged group. λ is always bigger than 1.

4 EXPERIMENT

4.1 Dataset and Implementation

To evaluate the method, we adopt anatomically accurate numerical breast phantoms (NBPs) [25], which are constructed by using the Virtual Imaging Clinical Trial for Regulatory Evaluation (VICTRE) toolset [26] for USCT virtual imaging studies. In particular, the generated NBPs are stratified based on the four different levels of breast density (percentage of fibroglandular tissue) defined according to the American College of Radiology's (ACR) Breast Imaging Reporting and Data System (BI-RADS) [9]. The breast size and percentage of fibroglandular tissue of each NBPs are randomly chosen based on the the physiological range for each breast density type. Each NBP corresponds to a speed of sound (SOS) map, where the values were assigned randomly within realistic ranges, varying spatially. The training dataset contains 1,230 NBPs while there are 164 NBPs in the testing set, each type has 41 samples.

To highlight the distinctions among categories, we have organized the Speed of Sound (SOS) summary for each class into Fig. 6. During analysis, we excluded segments where speed equals 1.5, considering these as non-essential, akin to a picture's background. Class A typically exhibits lower sound speed values, with peaks occurring at lower ranges. Conversely, in subsequent classes, the peaks of wave speeds shift rightward, with increasingly higher peaks emerging.

We develop our approach using PyTorch, leveraging InversionNet [30] as the backbone for APS-FWI. For the initial learning rates, we apply 1e-4. The batch size is set to 16. To regularize the model,

Table 1: Increase the density to see how the model performance changed

Density	Group	SSIM	PSNR	SPD (SSIM)	DI (SSIM)	SPD (PSNR)	DI (PSNR)
0.5	Unpri.	0.7173	22.9051	-0.1048	0.8725	-1.4052	0.9422
	Pri.	0.8221	24.3103				
0.9	Unpri.	0.7268 ✓	23.4702 ✓	-0.1021 ✓	0.8768 ✓	-1.1648 ✓	0.9527 ✓
	Pri.	0.8289 ✓	24.6350 ✓				
1	Unpri.	0.7315 ✓	23.3933 ✓	-0.0981 ✓	0.8818 ✓	-1.2521 ✓	0.9492 ✓
	Pri.	0.8296 ✓	24.6454 ✓				

Table 2: Comparisons between two-stage training and fairness weighted loss for density=0.9.

Method	Group	SSIM	PSNR	SPD (SSIM)	DI (SSIM)	SPD (PSNR)	DI (PSNR)
Baseline	Unpri.	0.7268	23.4702	-0.1021	0.8768	-1.648	0.9527
	Pri.	0.8289	24.6350				
Two-stage	Unpri.	0.5361 ×	18.2316 ×	-0.0273 ✓	0.9515 ✓	-0.6641 ✓	1.0378 ✓
	Pri.	0.5634 ×	17.5675 ×				
Fairness Weighted loss	Unpri.	0.7300 ✓	23.5294 ✓	-0.0990 ✓	0.8806 ✓	-1.1022 ✓	0.9552 ✓
	Pri.	0.8290 ✓	24.6316 ✓				

we use a consistent weight decay of 1e-4 across both phases. The AdamW optimizer is utilized for training, continuing until model convergence is achieved.

we calculated the peak signal-to-noise ratio (PSNR) [11] which is the approximate estimation of human perception of reconstruction quality and also structural similarity index measure (SSIM) [28] that reflects the structure of objects in the scene, from the perspective of image composition on the entire test dataset.

4.2 Main Results

In Table 1, we present comparative results between the model performance of different input waveform densities. key observations emerge from this comparison. ✓ indicates that the experimental results are better than the baseline and × indicates that the experimental results are worse. We conducted three basic experiments based on varying numbers of sources and receivers, observing that increasing either parameter—which implied the use of devices with higher sampling density for waveform capture to produce final SOS maps—marginally enhanced the model's imaging performance from the perspective of SSIM and PSNR. This improvement was evident in both SSIM and PSNR metrics, for both unprivileged and privileged groups. From a fairness standpoint, it's true that bigger devices fair better, but this approach presents a significant challenge, as the cost of the equipment tends to increase over time. We aim to avoid escalating hardware costs as a solution to this issue.

In subsequent experiments, we will choose different densities and observe the effectiveness of the other two methods.

In Table 2, When analyzing a specific device configuration, two-stage training appears to enhance fairness, albeit at a significant expense. This approach detrimentally affects the model's performance across both privileged and unprivileged groups, yet it also facilitates improvements in the fairness metrics of the final model for these groups. This approach was primarily rejected because it compromised the model's imaging capabilities for both groups. As we have previously described, we will use two different indicators of fairness, SPD and DI. SPD is closely related to the concept

of demographic parity, but DI addresses indirect discrimination. Upon evaluating the model's performance across different groups, fairness-weighted loss does not intentionally sacrifice individual group performance to enhance overall fairness. Some groups even show a performance improvement. From the second setting density =0.9, the sources are 16 and the receivers are 128, the SSIM of the baseline is 0.7268 but it increased to 0.7300 for the unprivileged group. For the privileged group, the value increased from 0.8289 to 0.8290. The fairness of the model is also improved in terms of the metrics of fairness.

In Table 3, diverse indicators yield conflicting outcomes, demonstrating that no single measurement can fully capture the scenario. For instance, examining only the SSIM results in the setting density=0.5, along with the corresponding SPD and DI, might lead one to conclude that the weighted fairness loss function falls short in enhancing the model's fairness. However, a closer look at the PSNR results and their corresponding SPD and DI indicates an improvement in fairness. This discrepancy underscores the limitation of relying on a solitary metric for evaluation. In practice, a comprehensive assessment requires multiple measures to accurately evaluate each approach. All of these observations show us the existence of fairness problems and the results of different metrics.

Table 3: Comparisons between two-stage training and fairness weighted loss for density= 0.5.

Method	Group	SSIM	PSNR	SPD (SSIM)	DI (SSIM)	SPD (PSNR)	DI (PSNR)
Baseline	Unpri.	0.7173	22.9051	-1.048	0.8725	-1.4052	0.9422
	Pri.	0.8221	24.3103				
Two-stage	Unpri.	0.5321 ×	17.5921 ×	-0.0327 ✓	0.9421 ✓	0.2952 ✓	1.0171 ✓
	Pri.	0.5648 ×	17.2969 ×				
Fairness Weighted loss	Unpri.	0.7050 ×	23.4330 ✓	-0.1182 ×	0.8564 ×	-1.2841 ✓	0.9480 ✓
	Pri.	0.8232 ✓	24.7171 ✓				

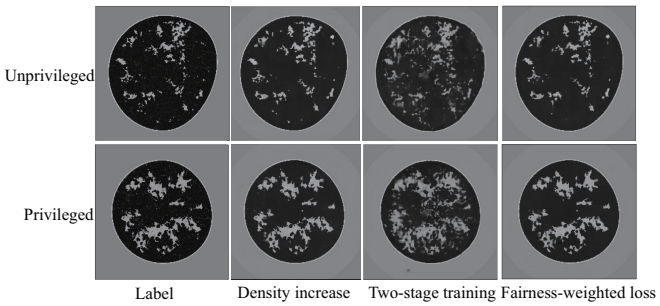


Figure 7: Visualization results for SOS map of one sample from the unprivileged group.

4.3 Visualization Results

Visualization results enable us to evaluate the effectiveness of each method. Figure 7 presents a sample from each group for analysis, highlighting the method's differences. The results are derived from the settings (32,128).

In Fig. 7 visualization reveals that the baseline and fairness-weighted loss produce results closely resembling each other, capturing the general outline but missing some details compared to the label image. In contrast, two-stage training diverges significantly, displaying rough edges, blurring, and excessive detail, including elements not present in the original. This discrepancy is likely due to its samples being oversampled. The same situation happened in the privileged group. In the method of two-stage training, the image now contains a significantly larger amount of information that was not originally present, but unlike before, the edges are relatively smooth. With this approach two-stage to model training, the model is trained to be so detail-oriented that it constructs a lot of information that is not there. This means that if you want to train by increasing the amount of data in a unprivileged group be sure to be careful about the number of repetitions of sample.

In Fig. 8, we display a velocity profile, illustrating the velocity's continuous variation across different positions. Fig. 8 reveals that the two-stage method, marked by the green line, exhibits sudden velocity changes, underscoring our previous observation in Fig. 7 where the two-stage method introduces excessive detail not present originally. While fairness loss and baseline methods can simulate velocity transitions, they fail to accurately capture the details. They adequately match large mutations but falter with minor variations. Additionally, whereas the original data demonstrates a significant drop, the reconstructed data fails to replicate this descent accurately.

5 CONCLUSION

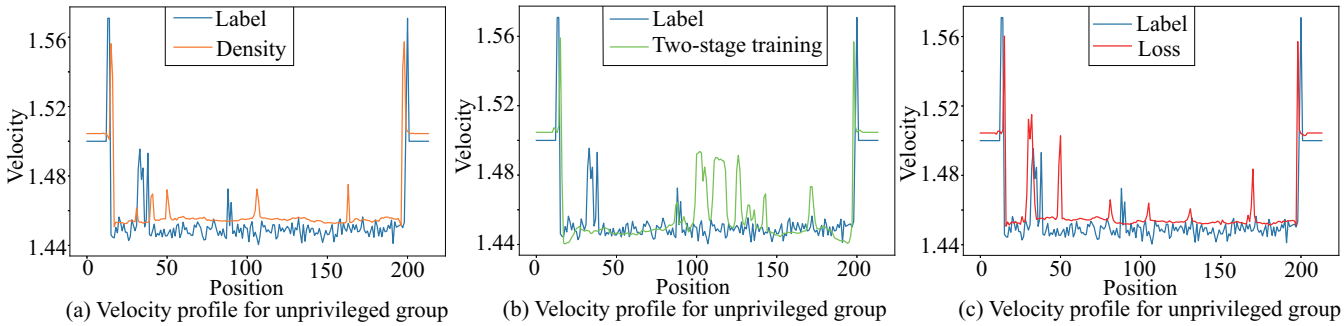
In this paper, we address the issue of fairness in USCT image reconstruction, highlighting how certain categories are consistently overlooked, complicating the reconstruction process. Our exploration of potential solutions involved two distinct methods Two-stage and fairness-weighted loss. Initially, we discovered that repeated use of data can lead the model to overemphasize minor details, detrimentally affecting the outcomes. Conversely, employing a loss function yielded inconclusive results, with its effectiveness varying under different conditions. These experiments underscore the necessity of abandoning a one-size-fits-all approach in favor of a more nuanced perspective that considers multiple factors when tackling image reconstruction biases.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the National Institutes of Health (NIH) (Award No. 1R01EB033387-01).

REFERENCES

[1] [n. d.]. breast cancer. https://www.breastcancer.org/facts-statistics/.
[2] [n. d.]. metrics. https://www.mathworks.com/help/risk/explore-fairness-metrics-for-credit-scoring-model.html/.
[3] Cynthia L Bennett and Os Keyes. 2020. What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing* 125 (2020), 1–1.
[4] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).
[5] Norman F Boyd, Helen Guo, Lisa J Martin, Limei Sun, Jennifer Stone, Eve Fishell, Roberta A Jong, Greg Hislop, Anna Chiarelli, Salomon Minkin, et al. 2007. Mammographic density and the risk and detection of breast cancer. *New England journal of medicine* 356, 3 (2007), 227–236.
[6] Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts.*



(a) Velocity profile for unprivileged group of the density increasement

(b) Velocity profile for unprivileged group of the two-stage training

(c) Velocity profile for unprivileged group of the weighted loss

Figure 8: Velocity profile for unprivileged groups.

[7] Yuning Du, Yuyang Xue, Rohan Dharmakumar, and Sotirios A Tsaftaris. 2023. Unveiling Fairness Biases in Deep Learning-Based Brain MRI Reconstruction. In *Workshop on Clinical Image-Based Procedures*. Springer, 102–111.

[8] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2017. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613* (2017).

[9] Carl D'Orsi, L Bassett, S Feig, et al. 2018. Breast imaging reporting and data system (BI-RADS). *Breast imaging atlas, 4th edn. American College of Radiology, Reston* (2018).

[10] Weituo Hao, Mostafa El-Khamy, Jungwon Lee, Jianyi Zhang, Kevin J Liang, Changyou Chen, and Lawrence Carin Duke. 2021. Towards fair federated learning with zero-shot data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3310–3319.

[11] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*. IEEE, 2366–2369.

[12] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 2018. Seven-point checklist and skin lesion classification using multi-task multimodal neural nets. *IEEE journal of biomedical and health informatics* 23, 2 (2018), 538–546.

[13] Panagiotis Kouloumtzios, Tomasz Rymarczyk, and Manuchehr Soleimani. 2021. A triple-modality ultrasound computed tomography based on full-waveform data for industrial processes. *IEEE Sensors Journal* 21, 18 (2021), 20896–20909.

[14] Hongming Li and Yong Fan. 2019. Early prediction of Alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 368–371.

[15] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10052–10061.

[16] Pingchuan Ma, Shuai Wang, and Jin Liu. 2020. Metamorphic Testing and Certified Mitigation of Fairness Violations in NLP Models.. In *IJCAI*. 458–465.

[17] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*. PMLR, 107–118.

[18] Thomas Robins, Jorge Camacho, Oscar Calderon Agudo, Joaquin L Herraiz, and Lluís Guasch. 2021. Deep-learning-driven full-waveform inversion for ultrasound breast imaging. *Sensors* 21, 13 (2021), 4570.

[19] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2020. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696* (2020).

[20] Nicole V Ruiter, Michael Zapf, Torsten Hopp, Robin Dapp, Ernst Kretzek, Matthias Birk, Benedikt Kohout, and Hartmut Gemmeke. 2012. 3D ultrasound computer tomography of the breast: A new era? *European Journal of Radiology* 81 (2012), S133–S134.

[21] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2019. Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. *arXiv preprint arXiv:1911.12587* (2019).

[22] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 358–364.

[23] Yi Sheng, Junhuan Yang, Yawen Wu, Kevin Mao, Yiyu Shi, Jingtong Hu, Weiwen Jiang, and Lei Yang. 2022. The larger the fairer? small neural networks can achieve fairness for edge devices. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*. 163–168.

[24] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[25] Ge Wang, Jong Chul Ye, and Bruno De Man. 2020. Deep learning for tomographic image reconstruction. *Nature machine intelligence* 2, 12 (2020), 737–748.

[26] Kun Wang, Thomas Matthews, Fatima Anis, Cuiping Li, Neb Duric, and Mark A Anastasio. 2015. Waveform inversion with source encoding for breast sound speed reconstruction in ultrasound computed tomography. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* 62, 3 (2015), 475–493.

[27] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5310–5319.

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[29] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8919–8928.

[30] Yue Wu and Youzuo Lin. 2019. InversionNet: An efficient and accurate data-driven full waveform inversion. *IEEE Transactions on Computational Imaging* 6 (2019), 419–433.

[31] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 570–575.

[32] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*. PMLR, 11492–11501.

[33] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 506–523.

[34] Junhuan Yang, Yi Sheng, Sizhe Zhang, Ruixuan Wang, Kenneth Foreman, Mikell Paige, Xun Jiao, Weiwen Jiang, and Lei Yang. 2022. Automated architecture search for brain-inspired hyperdimensional computing. *arXiv preprint arXiv:2202.05827* (2022).

[35] Junhuan Yang, Yi Sheng, Yuzhou Zhang, Weiwen Jiang, and Lei Yang. 2023. On-device unsupervised image segmentation. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.

[36] Junhuan Yang, Hanchen Wang, Yi Sheng, Youzuo Lin, and Lei Yang. 2024. A Physics-guided Generative AI Toolkit for Geophysical Monitoring. *arXiv preprint arXiv:2401.03131* (2024).

[37] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision* 129 (2021), 3069–3087.

[38] Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, et al. 2021. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665* (2021).