

Debiasing Cardiac Imaging with Controlled Latent Diffusion Models

Grzegorz Skorupko¹ (✉) [0000-0002-1059-6915], Richard Osuala¹ [0000-0003-1835-8564], Zuzanna Szafranowska¹ [0000-0003-4030-9726], Kaisar Kushibar¹ [0000-0001-7507-5208], Nay Aung^{2, 3} [0000-0001-5095-1611], Steffen E Petersen^{2, 3} [0000-0003-4622-5160], Karim Lekadir^{1, 4} [0000-0002-9456-1612], and Polyxeni Gkontra¹ [0000-0001-8828-6143]

¹ Barcelona Artificial Intelligence in Medicine Lab (BCN-AIM), Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Spain

grzegorz.skorupko@ub.edu

² William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, Charterhouse Square, London, UK

³ Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, West Smithfield, London, UK

⁴ Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Spain

Abstract. The progress in deep learning solutions for disease diagnosis and prognosis based on cardiac magnetic resonance imaging is hindered by highly imbalanced and biased training data. To address this issue, we propose a method to alleviate imbalances inherent in datasets through the generation of synthetic data based on sensitive attributes such as sex, age, body mass index, and health condition. We adopt ControlNet based on a denoising diffusion probabilistic model to condition on text assembled from patient metadata and cardiac geometry derived from segmentation masks using a large-cohort study, specifically, the UK Biobank. We assess our method by evaluating the realism of the generated images using established quantitative metrics. Furthermore, we conduct a downstream classification task aimed at debiasing a classifier by rectifying imbalances within underrepresented groups through synthetically generated samples. Our experiments demonstrate the effectiveness of the proposed approach in mitigating dataset imbalances, such as the scarcity of younger patients or individuals with normal BMI level suffering from heart failure. This work represents a major step towards the adoption of synthetic data for the development of fair and generalizable models for medical classification tasks. Notably, we conduct all our experiments using a single, consumer-level GPU to highlight the feasibility of our approach within resource-constrained environments. Our code is available at <https://github.com/faildeny/debiasing-cardiac-mri>.

Keywords: Deep Learning · Generative models · Fairness · Cardiac MRI

1 Introduction

Cardiovascular diseases remain the main cause of mortality worldwide, accounting for approximately one third of annual deaths globally [5]. Cardiovascular magnetic resonance (CMR) is currently the gold standard in evaluating the structure and function of the heart. However, its acquisition is expensive and the annotation process of multi-slice cine sequences requires a significant amount of time. Consequently, the amount of available training data is limited, hindering the adoption of deep learning based algorithms. Despite the efforts to automate CMR dataset collection, annotation and analysis, end-to-end models are still not common. Such solutions are more affected by the inherent biases in the training data especially when the data is scarce. For example, Puyol et al. [13] showed discrepancies in the performance of CMR segmentation models for subgroups based on sex and race. This finding was primarily attributed to the pronounced imbalance in the training dataset, which consisted mostly of individuals of white race. Such biases can significantly influence the decision-making process of classification models and were widely studied and addressed in various medical domains [7,12,20,17,8]. Advancements in generative deep learning models opened paths to previously unexplored approaches in tackling this crucial challenge in machine learning, namely, algorithmic bias. Some studies have proposed bias mitigation methods through different sampling strategies or modifications to model architecture and training procedures [18,21]. Nonetheless, in the medical domain, the adoption of generative models to mitigate biases through the use of synthetic data has received relatively little attention. Recent works based on GANs and Diffusion models focusing on dermatology, chest X-ray and histopathology domains, are among the very few examples in this direction [9,6]. Ktena et al. [6] proposed models conditioned on both diagnostic and sensitive attributes, such as sex, age, or skin tone, allowed to augment the unbalanced training dataset and successfully reduce the biases in classification tasks. However, to the best of our knowledge, none of the previous works focused on MRI modalities or cardiovascular domain, nor did they allow for conditioning image generation on shape information from segmentation masks or textual prompts.

To address this gap, we propose an open-source pipeline involving training of a resource-intensive Stable diffusion model [14] within a limited computational environment. More precisely, we adopt a latent diffusion model with combined

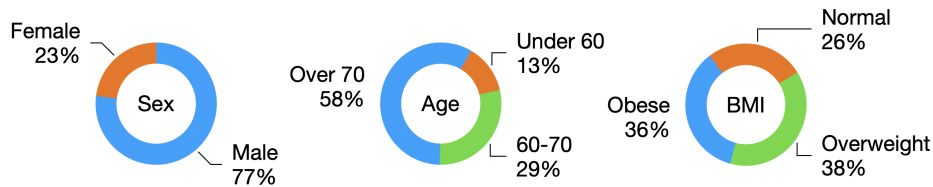


Fig. 1. Demographic statistics of patients diagnosed with heart failure from the UK Biobank imaging study.

text and image inputs to generate spatially consistent CMR frames to mitigate biases introduced by unbalanced training data in CMR-based deep learning models for disease diagnosis. This approach facilitates the generation of CMR data for underrepresented patient subgroups, considering factors such as sex, age, BMI, and heart conditions, alongside spatial-temporal features defined through segmentation masks from multiple cardiac phases. We evaluate the quality of the generated images using the well-established Fréchet Inception Distance score (FID) [3]. Furthermore, we assess the impact of these generated images on a heart failure classification model, analyzing the model’s fairness across diverse patient subgroups. Overall, the proposed approach serves as a targeted augmentation method in resource-limited environments. Despite the emphasis of this work on diagnosis, our approach holds potential for other medical tasks, including prognosis and segmentation.

2 Methodology

2.1 Dataset

For the purposes of this study we use the UK Biobank (UKBB), a large-scale research resource comprising over 500,000 participants. The UKBB collected demographic information, electronic health record (EHR), biomarkers and genomics from patients aged 40-69 years at the time of recruitment for the study between 2006 and 2010 across the United Kingdom.

In this work, we focus on a subset of patients that were part of the imaging study of the UKBB and underwent a CMR scan. In total, our dataset consists of 25480 multi-slice, short-axis cine CMRs with annotations for end-diastole (ED) and end-systole (ES) frames. The annotation masks provide labels for key cardiac structures, namely the right ventricle (RV) cavity, the left ventricle (LV) cavity and the myocardium. Based on International Classification of Diseases (ICD-10) codes from in-hospital patient data, we identified a subset of 270 patients diagnosed with heart failure at the time of the CMR acquisition. Fig. 1 provides the distribution of characteristics of the participants included in the study. In our analysis, we divided patients into groups by age: below 60, 60-70 and over 70 years old, by BMI: below 25 (underweight and normal), 25-30 (overweight) and over 30 (obese), and by sex.

Data pre-processing Due to the multidimensional nature of CMR samples (4D), we conduct several data preprocessing steps to adapt to the image format most commonly used in state-of-the-art classification models, i.e. 2D, 3-channel images, to be generated by the Stable diffusion model with ControlNet. Firstly, from each 4D CMR sample, we select the central slice of the volume. Additionally, to keep cardiac motion information, we stack the cine frames from ED and ES phases as color channels, effectively creating a 2D RGB image of the same shape as the data used for pretraining the generative model. To further extend the number of training images from each patient, we extract 3 central heart slices

instead of one. Moreover, we increase the number of samples by selecting the cine frames directly before and after the annotated ED and ES phases. This approach results in a nine-fold increase in the number of training images. It should be noted that we do not apply this augmentation to the validation or test sets, where we solely use one central slice with frames marked as ED and ES.

2.2 Conditioned image generation

An overview of the proposed pipeline for generating synthetic CMR images based on textual information and cardiac geometry information is provided in Fig. 2. All experiments for generating synthetic CMR data are conducted using the ControlNet introduced by Zhang et al. [22]. This model expands the capabilities of a standard Stable diffusion model [14], by enabling stable fine-tuning with text and image input modalities. Its core concept is based on creating a copy of the original pretrained model and appending the spatial input only to its cloned branch. The outputs of this part of the network are connected with *zero convolution* layers to the original Stable diffusion architecture. Usage of *zero convolution* allows to mitigate the negative impact of noise during the first stages of training and to preserve the states of the backbone of the trainable copy. At the same time, weights of the original pretrained branch remain locked, to preserve its generative capabilities learned from previously seen billions of images. With this approach, pretrained latent diffusion models can be adapted to different imaging domains without the need of costly training from scratch, which is not always feasible.

Diffusion model training We conduct all experiments on a single Nvidia 3080Ti GPU with 16GB of memory. To train the diffusion model, we adopt the implementation provided by [22]. To fully leverage the advantages of the pretrained model, we upscale the training samples to 512x512 pixels to match

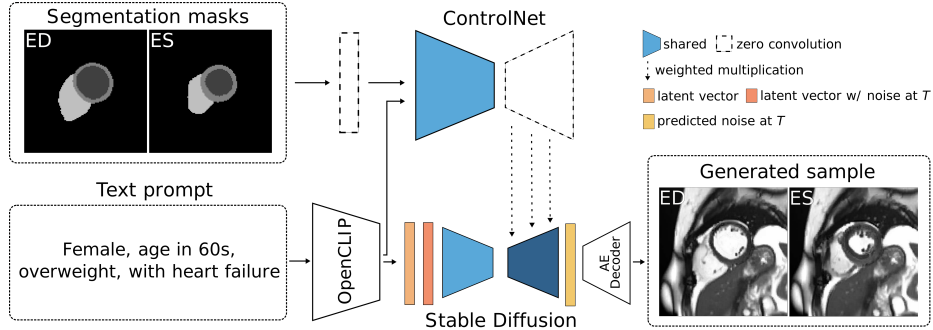


Fig. 2. Overview of the proposed pipeline for generating synthetic CMR data conditioned on textual information and cardiac geometry derived from segmentation masks.

the final pretraining resolution of a Stable diffusion 2.1-base model [14]. In the training setup, we use the pretrained image AutoEncoder network and the Open-Clip [4] text encoder pretrained on the LAION-5B [15] dataset. During the training phase, we exclusively fine-tune the ControlNet branch of the model. In this setup, it is possible to train the model with batch size of 1 with 2 gradient accumulations. We train the model with a learning rate of $1e-5$ for 5 epochs, which takes approximately 3 days in our setup. All the code is based on PyTorch framework [11].

Debiased dataset generation To address biases resulting from underrepresentation of certain groups, we use weighted random sampling on our initial dataset. Patients are grouped based on sex, age, BMI and diagnosis, which creates 36 groups in total. For example, female, overweight patients younger than 60 years that are healthy belong to the same subset. Based on each group’s population we calculate the sampling weights that are inversely proportional to their size. We subsequently generate synthetic images based on existing prompts and masks for underrepresented groups. This way, we ensure that imaging inputs are coherent with patient’s characteristics and do not contribute to additional noise. Experiments done on synthetic datasets with less constrained approach (e.g. mask selection based only on diagnosed condition) showed significantly worse performance on downstream tasks. We hypothesise that this may be attributed to non obvious differences in segmentation masks of patients among different subgroups, however this claim has not been further investigated.

2.3 Downstream classification model training

For the downstream classification task, due to the relatively small dataset size, we use a well established ResNet-18 model [2] with weights pretrained on the ImageNet dataset [1]. All training samples are scaled to the native pretraining resolution of 224×224 pixels. We train each model with batch size of 64 for 10 epochs with a starting learning rate $1e-4$, and reduction by 2 on plateau for 3 epochs. We apply standard augmentations including e.g. random flipping and Gaussian noise. The model weights are saved after each epoch and best checkpoint is selected based on balanced accuracy on the validation set. We keep 20% of the dataset for test, and from the training part we exclude 20% as validation. During training, we perform undersampling, as increasing the proportion of healthy subjects did not improve the performance and mainly increased the training time. However, during evaluation we keep the original prevalence, which is around 1%. We run each experiment 8 times on the same data split, as we are limited to the same split due to relatively long diffusion model training and inference.

2.4 Evaluation metrics

Classification performance To evaluate the performance of the classifiers on the downstream heart failure diagnosis task, given the highly imbalanced nature

of our data, we opt for balanced accuracy as our primary metric, defined as:

$$\text{Balanced Accuracy} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (1)$$

where TP, FN, TN and FP stand for true positives, false negatives, true negatives, and false positives, respectively. The balanced accuracy metric, derived from the arithmetic mean of sensitivity and specificity, enables us to compare performance across subgroups, despite differences in disease prevalence within each population.

3 Results

3.1 Synthetic data evaluation

FID comparison between synthetic subpopulations Table 1 provides the FID scores for CMR images generated for the different subgroups of interest based on sex, age, BMI and health condition. More precisely, we assess the FID values between real CMR data from different subgroups eg. male versus female, as well as between real and synthetic CMR within same subgroups. Overall, the FID scores of the synthetic samples are in an acceptable range in line with findings in the literature [10]. There is a visible trend of FID being smaller (i.e. better) for groups of the same population - both for real and synthetic data. Interestingly, the FID of real samples with heart failure is higher than in case of samples of healthy patients. The synthetic data does not seem to follow this trend, which suggests that the generative model tends to produce heart failure samples with less non-typical, outlying visual features. This conclusion is in line with findings presented previously by [9] in case of GANs. Nonetheless, it is worth noting that, as shown by [19], FID scores do not necessarily represent the utility of synthetic data on downstream tasks.

Qualitative analysis Sample images in Fig. 3 show the generative model’s capability to learn the relation between visual BMI indicators with textual information introduced through the prompt used to create each sample. An increase of pericardial adipose tissue (PAT), marked with red arrows, can be observed in the images with the progression of BMI level. As noted in [16] PAT characteristics are a significant factor in discrimination of HF patients.

CMR images generated with different random seeds for the same input data (see Fig. 4) showcase the model’s ability to create a diverse set of samples. This feature enhances the augmentation potential of the proposed approach, especially for highly underrepresented groups where simple oversampling methods are not effective due to model overfitting.

3.2 Downstream task: Heart failure classification

Table 2 presents the balanced accuracy achieved in the CMR-based diagnosis for different subgroups using four distinct sets of training data. It should be noted

Table 1. FID scores comparison for images generated for different subpopulations. HF stands for heart failure. Due to a limited amount of real data, 500 samples from each group were randomly selected for the score calculation.

	FID [3] ↓		
Sex	Male - Male	Male - Female	Female - Female
Real - Real (lower bound)	30.39 (0.34)	44.20 (0.88)	30.37 (0.17)
Real - Synthetic	82.01 (1.95)	81.93 (1.03)	79.56 (2.82)
Age	50s - 50s	50s - 70s	70s - 70s
Real - Real (lower bound)	29.86 (0.21)	33.47 (0.53)	32.02 (0.12)
Real - Synthetic	76.31 (0.99)	85.68 (2.22)	84.51 (1.28)
BMI	Normal - Normal	Normal - Obese	Obese - Obese
Real - Real (lower bound)	29.14 (0.27)	41.98 (0.34)	33.25 (0.33)
Real - Synthetic	78.11 (1.09)	88.52 (2.23)	83.51 (1.45)
Condition	Healthy - Healthy	Healthy - HF	HF - HF
Real - Real (lower bound)	30.51 (0.09)	40.43 (0.37)	48.90 (0.19)
Real - Synthetic	83.55 (0.91)	95.80 (2.16)	79.86 (0.72)

that there are 3168 real samples in the training set. The additional, synthetic images are always balanced among all subgroups and classes, i.e. there is an equal number of patients with all combinations of sex, age and BMI groups, as well as diagnosis (healthy or HF).

Using solely synthetic data during training results in the worst performance. Nonetheless, combining synthetic data and real data to achieve more balanced training datasets yields higher performance for underrepresented subgroups. Furthermore, the highest improvement, 5% in terms of balanced accuracy, is observed in diagnosing heart failure among individuals under the age of 60, and among individuals with normal BMI, which are the least represented groups within our dataset, as illustrated in Fig. 1. Furthermore, adding balanced synthetic data reduced the accuracy gap between groups with higher and lower disease prevalence, e.g. for males and females by 37%.

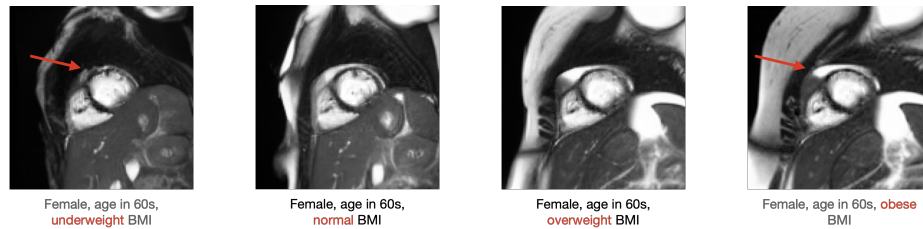


Fig. 3. Effect of altering sensitive features on the generated images using the prompt. Specifically in this example, the sensitive attribute BMI was modified between generation runs to observe its effect on the generated CMR scans.



Fig. 4. Variability in the generated CMR images achieved by using different seeds for prompt: *Male, age in 70s, normal BMI, with heart failure*. Note that the same segmentation mask (left) was used to obtain all samples. For simplicity, we only display ED frame from each synthetic sample.

Table 2. Balanced accuracy comparison per group and per model, trained on: real data only, real data and 2000 or 4000 synthetic images and synthetic data only.

	Prevalence	Real	Real + 2000	Real + 4000	Synthetic
Overall	1.08%	0.740 $\pm$.025	0.748 $\pm$.025	0.752$\pm$.019	0.683 $\pm$.014
Female	0.56%	0.649 $\pm$.019	0.655 $\pm$.045	0.664$\pm$.027	0.612 $\pm$.046
Male	1.56%	0.733$\pm$.030	0.722 $\pm$.025	0.717 $\pm$.024	0.637 $\pm$.023
Normal BMI	0.60%	0.668 $\pm$.026	0.704 $\pm$.045	0.716$\pm$.031	0.630 $\pm$.039
Overweight	0.96%	0.764$\pm$.029	0.745 $\pm$.037	0.742 $\pm$.043	0.694 $\pm$.045
Obese	2.43%	0.765$\pm$.033	0.747 $\pm$.027	0.750 $\pm$.032	0.689 $\pm$.042
Age < 60	0.37%	0.770 $\pm$.056	0.817$\pm$.063	0.812 $\pm$.054	0.762 $\pm$.042
Age 60-70	1.06%	0.737$\pm$.022	0.733 $\pm$.025	0.736 $\pm$.020	0.674 $\pm$.033
Age > 70	1.93%	0.721 $\pm$.036	0.721 $\pm$.037	0.730$\pm$.033	0.649 $\pm$.020

4 Conclusion

In this work, we explore the potential of generative latent diffusion models in tackling biases in CMR datasets. We demonstrate how textual inputs (sex, age, BMI, heart condition) and imaging inputs, such as segmentation mask encompassing cardiac shape information, can be combined and used to achieve highly customized, versatile control over synthetic data generation process.

Overall, we observed a consistent improvement in performance among underrepresented groups, upon the incorporation of synthetic, balanced data during training. This highlights the potential of targeted data augmentation as a promising avenue for improving fairness in cardiac imaging studies, paving the path towards the integration of deep generative models into clinical practice.

Moreover, all of our models, including the computationally-expensive multi-conditional diffusion models, were trained on a consumer-level single GPU. We show that this type of data augmentation could be achievable even in modest, non-commercial settings, contrary to common belief.

Acknowledgments. This work was conducted using the UK Biobank resource under access application 2964. Moreover, it received funding from the European Union’s Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project). KK holds the Juan de la Cierva fellowship with a reference number FJC2021-047659-I. NA acknowledges support from MRC Clinician Scientist Fellowship (MR/X020924/1) SEP declares consultancy for Circle Cardiovascular Imaging, Inc., Calgary, Canada.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640. NIPS’17, Curran Associates Inc., Red Hook, NY, USA (2017)
4. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). <https://doi.org/10.5281/zenodo.5143773>, if you use this software, please cite it as below.
5. Jagannathan, R., Patel, S.A., Ali, M.K., Narayan, K.M.V.: Global Updates on Cardiovascular Disease Mortality Trends and Attribution of Traditional Risk Factors. *Current Diabetes Reports* **19**(7), 44 (Jun 2019). <https://doi.org/10.1007/s11892-019-1161-2>
6. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Karthikesalingam, A., Cemgil, T., Goyal, S.: Generative models improve fairness of medical classifiers under distribution shifts (2023)
7. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* **117**(23), 12592–12594 (2020). <https://doi.org/10.1073/pnas.1919012117>
8. Luo, L., Xu, D., Chen, H., Wong, T.T., Heng, P.A.: Pseudo bias-balanced learning for debiased chest x-ray classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. pp. 621–631. Springer Nature Switzerland, Cham (2022)
9. Mikołajczyk, A., Majchrowska, S., Limeros, S.C.: The (de)biasing effect of GAN-based augmentation methods on skin lesion images (Jun 2022). https://doi.org/10.1007/978-3-031-16452-1_42
10. Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., et al.: medigan: a python library of pre-trained generative models for medical image synthesis. *Journal of Medical Imaging* **10**(6), 061403 (2023)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc. (2019)
12. Petersen, E., Feragen, A., da Costa Zemsch, M.L., Henriksen, A., Wiese Christensen, O.E., Ganz, M.: Feature robustness and sex differences in medical imaging: A case study in mri-based alzheimer’s disease detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. pp. 88–98. Springer Nature Switzerland, Cham (2022)

13. Puyol-Antón, E., Ruijsink, B., Mariscal Harana, J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., Chowienczyk, P., King, A.P.: Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. *Frontiers in Cardiovascular Medicine* **9** (2022). <https://doi.org/10.3389/fcvm.2022.859310>
14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10684–10695 (June 2022)
15. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022)
16. Szabo, L., Salih, A., Pujadas, E.R., Bard, A., McCracken, C., Ardissino, M., Antoniadis, C., Vago, H., Maurovich-Horvat, P., Merkely, B., Neubauer, S., Lekadir, K., Petersen, S.E., Raisi-Estabragh, Z.: Radiomics of pericardial fat: a new frontier in heart failure discrimination and prediction. *European Radiology* (Nov 2023). <https://doi.org/10.1007/s00330-023-10311-0>
17. Wachinger, C., Rieckmann, A., Pölsterl, S.: Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis* **67**, 101879 (2021). <https://doi.org/10.1016/j.media.2020.101879>
18. 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: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8916–8925 (2020). <https://doi.org/10.1109/CVPR42600.2020.00894>
19. Xing, X., Felder, F., Nan, Y., Papanastasiou, G., Simon, W., Yang, G.: You don't have to be perfect to be amazing: Unveil the utility of synthetic images. *arXiv preprint arXiv:2305.18337* (2023)
20. Zare, S., Nguyen, H.V.: Removal of confounders via invariant risk minimization for medical diagnosis. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. pp. 578–587. Springer Nature Switzerland, Cham (2022)
21. 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*. p. 335–340. AIES '18, Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3278721.3278779>
22. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)