Debiasing Computer Vision Models using Data Augmentation based Adversarial Techniques

Teerath Kumar[1,*,†], Abhishek Mandal[2,*,†], Susan Leavy[3], Suzanne Little[2], Alessandra Mileo[2] and Malika Bendechache[4]

[1]*CRT-AI, School of Computing, Dublin City University, Dublin, Ireland*

[2]*Insight SFI Research Center for Data Analytics & School of Computing, Dublin City University, Dublin, Ireland*

[3]*Insight SFI Research Center for Data Analytics & School of Information and Communication Studies, University College Dublin, Dublin, Ireland*

[4]*ADAPT & Lero Research Centres, School of Computer Science, University of Galway, Galway, Ireland*

Abstract

Deep learning models in computer vision, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have been found to exhibit significant biases related to factors such as gender and ethnicity. These biases often originate from inherent imbalances in training data predominantly sourced from the internet. In this study, we aim to address gender bias in computer vision models by curating a specialized dataset that highlights gender-related disparities. Additionally, we measure dataset diversity across six datasets (FFHQ, WIKI, IMDB, LFW, UTK Faces, diverse dataset), five professions (CEO, engineer, nurse, politician, and teacher) and different query retrieval tasks using the Image Similarity Score (ISS). To reduce learned gender biases and increase data diversity, we propose adversarial data augmentation techniques that specifically target facial regions within images. These techniques, named Partial Mix (PM), that partially mixes two gendered faces in a squared pattern, and Noise Addition (NA), that adds noise to the facial region, are designed to mitigate bias. Our experimental results demonstrate increased data diversity across the six datasets and professions, along with reduction in gender bias for CNN-based models. However, these adversarial techniques were less effective in reducing bias for Vision Transformers. This discrepancy highlights the unique challenges for bias mitigation posed by ViTs. Consistent with prior research, our findings indicate that ViTs learn from a broader set of visual cues compared to CNNs. This increased sensitivity makes ViTs more prone to amplifying biases, emphasizing the need for tailored bias mitigation strategies when deploying these models in real-world applications.

Keywords

Adversarial Debiasing, Data Augmentation, Data Diversity, Fairness, Gender Bias

1. Introduction

Social biases related to ethnicity [1], gender [2], geographical region and culture [3, 4, 5] is now well-documented problem in computer vision. These biases mainly originate in training data primarily sourced from the Internet and is propagated and amplified throughout the machine learning pipeline [2, 3]. Such issues can cause a multitude of problems when models are deployed in real-world applications, including variances in accuracy in facial recognition systems depending on gender and race [1] and the generation of stereotypical images related to gender [6]. Such biases can cause harm, foster discrimination, and stymie progress towards a more equitable and just society [3, 2].

Numerous strategies have been proposed to mitigate bias in computer vision models. These include the expansion of dataset diversity, as outlined in Kärkkäinen et al.'s work [2], as well as the deployment of adversarial debiasing techniques [7]. In the context of image data augmentation for debiasing, previous research is relatively scarce [7, 8, 9]. The aforementioned studies have primarily employed

3rd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE24), co-located with AIxIA 2024, November 25-28, 2024, Bolzano, Italy

[*] Corresponding author.

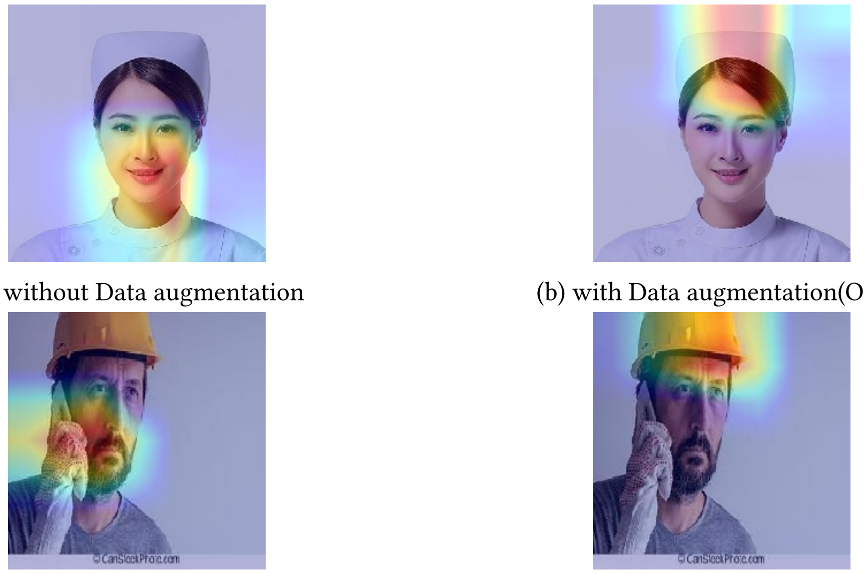
[†] These authors contributed equally.

teerath.menghwar2@mail.dcu.ie (T. Kumar); abhishek.mandal2@mail.dcu.ie (A. Mandal); susan.leavy@ucd.ie (S. Leavy); suzanne.little@dcu.ie (S. Little); alessandra.mileo@dcu.ie (A. Mileo); malika.bendechache@universityofgalway.ie (M. Bendechache)

data augmentation to address different facets of bias. Zhang et al. have explored data augmentation as a means to balance class representation [7], while Li et al. have focused on leveraging data augmentation for enhancing cross-bias generalization [9]. Smith et al. have also explored data augmentation within an evolutionary framework to combat gender and age bias [8]. Our research explores novel aspect of gender debiasing via data augmentation, particularly in the context of face recognition. Furthermore, our work contributes in the following ways:

- We address gender bias in computer vision models using data augmentation techniques with the help of face recognition and propose two novel data augmentation approaches: Partial Mixing (PM) and Uniform Noise Blur (NA).
- We measure and compare dataset diversity across six datasets (FFHQ, WIKI, IMDB, LFW and UTK Faces, Diverse Dataset), five professions and different query retrieval tasks, using two different variations of the Image Similarity Score (ISS) metric.
- Our approaches demonstrate that CNN-based models can effectively reduce gender bias, while supporting existing research that bias mitigation in Vision Transformers is more challenging.

The remainder of this paper is organized as follows: Section 2 briefly reviews existing related work, Section 3 explains the proposed methodology, Section 4 discusses the experimental setup, the insight and findings and finally Section 5 presents the conclusions.



(a) without Data augmentation

(b) with Data augmentation(Ours)

(c) without Data augmentation

(d) with Data augmentation(Ours)

Figure 1: Left and right column represent the trained model class activation map (CAM) without and with data augmentation, respectively. Without data augmentation trained models indicate gender bias – Nurse is female biased and Engineer is male biased. These CAMs are generated using Xception architecture [10] trained with and without data augmentation.

2. Related Work

2.1. Gender Bias

The issue of gender bias in computer vision models has received significant attention within the research community, with a multitude of proposed techniques for mitigating this bias. These approaches encompass various strategies, including the manipulation of learned representations [11], adjustments to the training dataset [12], and the application of adversarial methods [13]. It is important to note that a majority of these debiasing techniques have been tailored for Convolutional Neural Networks (CNNs). However, as the landscape of computer vision continues to evolve, Vision Transformers (ViTs) have gained prominence, often surpassing CNNs in numerous tasks, such as image classification [14, 15]. Mandal et al. observed that ViTs tend to exacerbate social biases to a greater extent when compared to CNNs [16].

2.2. Data Augmentation

Data augmentation aims to increase data diversity so that deep learning models can be trained to improve the generalization ability [17, 18, 19]. At present only limited data augmentation work has focused on debiasing. Zhang et al. [7] explored machine learning fairness in image classification, addressing bias from imbalanced data and harnessing adversarial examples as data augmentation for data distribution balance. Li et al. [9], aim to improve cross-bias generalization using data augmentation. They introduce “safety” and “unbiasedness” constraints to address the influence of biased cues in training data without manual intervention. Smith et al. [8], tackles gender and age classification biases by leveraging data augmentation techniques. The authors introduce an innovative approach that optimizes data augmentation settings through an evolutionary process, effectively reducing bias and improving model generalization. Though these above research works explore and mitigate gender bias using different data augmentation techniques, in our work we introduce two novel adversarial data augmentation techniques to address gender bias. The effect of data augmentation on gender debiasing is illustrated in Figure 1

3. Methodology

In this section, we introduce an alternative methodology. Initially, we employ facial recognition on the input image using the well-established and highly efficient face recognition algorithm, Single-shot Detection (SSD) [20]. To perform this task, we utilize a pre-trained model [1] and detected faces using OpenCV [2]. Once the facial region has been successfully detected within the original image, x, we proceed to apply the newly proposed data augmentation techniques as follows:

1. **Partial Mixing (PM) :** In this approach, the facial regions x_m and x_f of male and female, respectively are taken. Each is divided into four equal parts, and a random selection of squares from both facial regions is mixed. A mask, M, is partitioned into four segments, each filled with either 0’s or 1’s to respectively include or exclude those squares. Subsequently, an element-wise multiplication is conducted between the mask, M, and the male facial region, x_m, and $1 - M$ and female facial region, x_f, then both are added, resulting in the generation of the augmented image, $\tilde{x_a}$, as illustrated in Equation 1. Finally, the augmented facial region $\tilde{x_a}$ is reinserted into the original image. The overall process is depicted in Figure 2.

$$\tilde{x_a} = M \odot x_m + (1 - M) \odot x_f \tag{1}$$

2. **Noise addition (NA):** In this strategy, we incorporate uniformly distributed noise, generated within the range of 0 to 1, as expressed in Equation 2. This randomly generated noise, denoted as n_r, is then added to the facial region, x_m or x_f. Consequently, an augmented facial region, $\tilde{x}_a$, is produced, as outlined in Equation 3.

$$n_r = uniform(0, 1) \tag{2}$$

$$\tilde{x}_a = x_f + n_r \tag{3}$$

Then $\tilde{x}_a$ is placed back to its position in the original images. The overall process is shown in Figure 3.

[1] https://raw.githubusercontent.com/opencv/opencv_3rdparty/dnn_samples_face_detector_20180205_fp16/res10_300x300_ssd_iter_140000_fp16.ca

[2] https://docs.opencv.org/3.4/d6/d0f/group__dnn.html

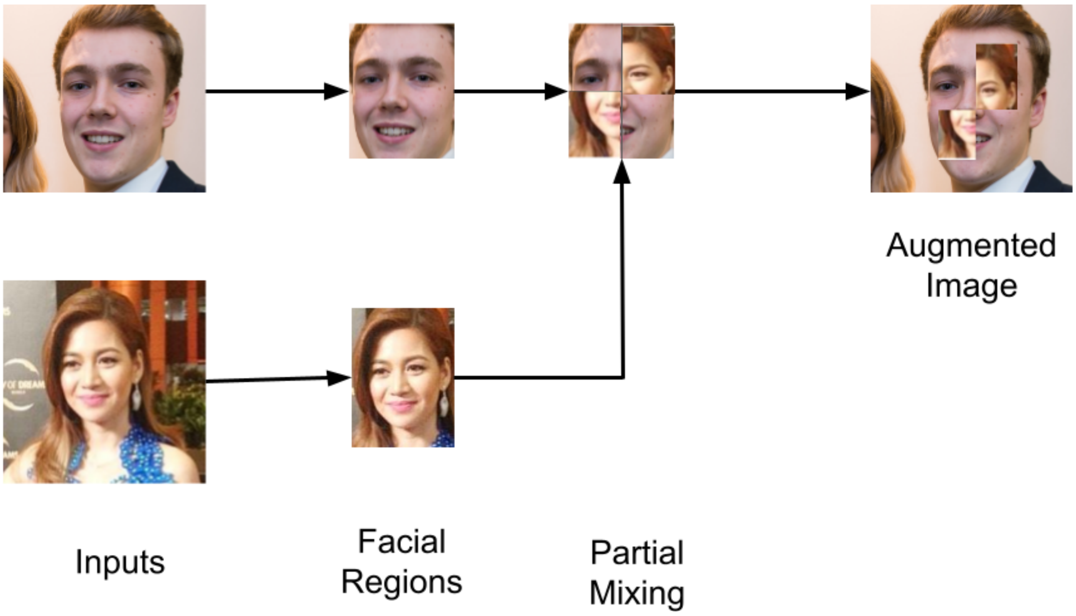


Figure 2: Partial Mixing Data Augmentation Process

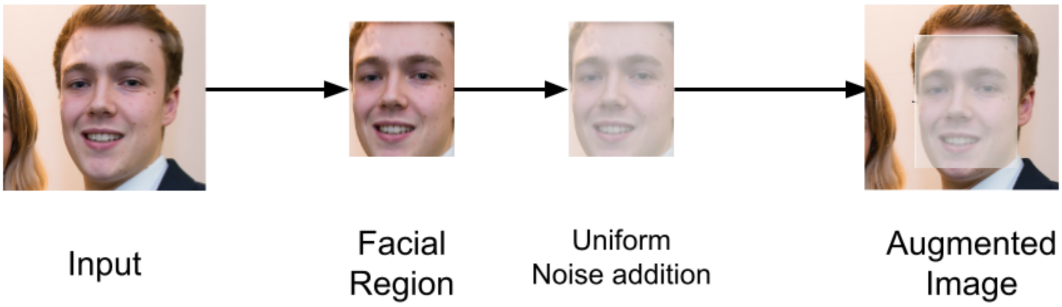


Figure 3: Noise Addition Data Augmentation Process

4. Results

4.1. Experimental setup

Image Similarity Score (ISS) is used to measure the diversity of a dataset. There are two variants of ISS – (i) ISS$_{Intra}$, which measures diversity in the dataset, (ii) ISS$_{cross}$, which measures diversity across the datasets – both introduced by Mandel [4] and both with a range of 0 to 2. To measure ISS$_{intra}$, we used six diverse datasets (FFHQ [21] , WIKI [22] , IMDB [22], LFW [23], UTK [24] and Diverse Dataset [4]),

five professions and a query retrieval task dataset, as described in Mandel [4]. The Image Similarity Score (ISS) measures how similar two images are based on features extracted by a pre-trained Convolutional Neural Network (CNN). In this study, we use VGG16, a 16-layer deep CNN trained on the ImageNet dataset. The feature extraction layers of VGG16 were employed to capture features from the images. To reduce the dimensionality of these extracted features, we applied Principal Component Analysis (PCA). For two images, I_1 and I_2, with corresponding feature vectors V_1 and V_2, the similarity between the images is calculated as shown in Equation 4 [4] and also algorithm for ISS_{intra} and ISS_{cross} are proposed by Mandal et al. [4]. A higher Image Similarity Score indicates greater dissimilarity between images.

Table 1

ISS$_{intra}$ of datasets for baseline results are from [4].

Dataset	Baseline	with PM	with NA
FFHQ [21]	0.9940	**1.04**	1.01
Diverse Dataset [4]	0.9895	**1.07**	1.02
WIKI [22]	0.9786	**1.11**	1.01
IMDB [22]	0.9661	**1.21**	1.00
LFW [23]	0.9536	**1.11**	1.03
UTK [24]	0.9418	**1.11**	1.02

$$\begin{gathered}\operatorname{sim}\left(I_1, I_2\right)=1-\frac{v_1 \cdot v_2}{\left\|v_1\right\|_2 \cdot\left\|v_2\right\|_2} \\ \operatorname{sim}\left(I_1, I_2\right) \in[0,2]\end{gathered} \tag{4}$$

We curated a visual dataset with ten classes: *CEO, Engineer, Baseball, Rugby, Snowboarding, Nurse, School Teacher, Hairdryer, Shopping, and Dollhouse.* The first five categories are generally (in a social or stereotypical sense) male-dominated and the last five female are female-dominated [3]. Selenium was used to query the Google Search API by creating fresh environments without tracking cookies. We created 4 datasets: (1) a biased training dataset with the first five classes being over-represented with images of men and the last five being over-represented with images of women in a ratio of 4:1, two data-augmented versions of the biased dataset using (2) Partial Mix (PM) and (3) Noise addition (NA) and (4) a manually gender-balanced dataset to generate a reference with unbiased accuracy. It is important to note, dataset size was increased after performing augmentation. Each training dataset contained at least 7,500 images. Eight model architectures were chosen to give appropriate coverage over CNNs and ViTs referring to current high performing and popular architectures: four CNNs (Inception v3, Xception, ResNet 150, and VGG16) and four ViTs (B/16, B/32, L/16, and L/32). With the initial layers frozen, we fine-tuned five models for each architecture, a total of 40 models for each dataset and tested their accuracy on a manually gender-balanced test dataset. The models were all pre-trained on the ImageNet dataset.

4.2. Findings and discussions

4.2.1. Intra-dataset Image Similarity (ISS$_{intra}$) Evaluation:

The results in Table 1 demonstrate the effectiveness of both of our methods, “with PM” and “with NA”, in improving the Intra-dataset Image Similarity Score (ISS$intra$) across multiple datasets. Our PM approach consistently outperforms the baseline for all datasets, achieving the highest ISS$intra$ values representing greater diversity in the results. Specifically, the PM method achieves notable improvements on the FFHQ, Diverse Dataset, WIKI, IMDB, LFW, and UTK datasets, with the highest ISS$_{intra}$ observed for the IMDB dataset at 1.21. In contrast, the NA approach, while still improving upon the baseline, yields slightly lower scores compared to the PM method but consistently surpasses the baseline. This demonstrates that both methods contribute to improved dataset diversity, with the PM approach being more effective overall.

4.2.2. Query-based ISS$_{intra}$ Analysis for Various Language-Location Pairs:

Table 2 provides a detailed analysis of the ISS$_{intra}$ values across different queries and language-location pairs, further comparing our PM and NA methods to the baseline. Across nearly all queries, both approaches show improvements over the baseline, with the “with NA” method often slightly outperforming “with PM” in specific regions and queries.

For the CEO query, the NA method demonstrates the highest improvements, particularly in the Arabic-West Asia & North Africa region, achieving an ISS$intra$ of 0.9923, significantly surpassing

Table 2

Image Similarity Score across all possible queries. Baseline results are from [4].

Query	Language Location Pair	ISS$_{Intra}$		
		Baseline	with PM	with NA
CEO	Arabic-West Asia & North Africa	0.8990	0.9901	**0.9923**
	English-North America	0.9690	**0.9724**	0.9711
	English-West Europe	0.9295	0.9558	**0.9571**
	Hindi-South Asia	0.9978	0.9993	**0.9998**
	Indonesian-SE Asia	0.9837	0.9926	**0.9931**
	Mandarin-East Asia	0.9895	0.9974	**0.9986**
	Russian-East Europe	0.9597	0.9962	**0.9977**
	Spanish-Latin America	0.9747	0.9904	**0.9933**
	Swahili-Sub Saharan Africa	0.9771	0.9917	**0.9939**
Engineer	Arabic-West Asia & North Africa	0.9864	0.9946	**0.9967**
	English-North America	0.9883	**1.0021**	1.0014
	English-West Europe	1.0009	**1.0017**	1.0009
	Hindi-South Asia	1.0031	1.0025	**1.0039**
	Indonesian-SE Asia	0.9872	0.9885	**0.9899**
	Mandarin-East Asia	0.9911	0.9921	**0.9935**
	Russian-East Europe	1.0072	**1.0082**	1.0076
	Spanish-Latin America	0.9850	0.9967	**0.9981**
	Swahili-Sub Saharan Africa	0.9837	0.9952	**0.9961**
Nurse	Arabic-West Asia & North Africa	1.0026	1.0029	**1.0032**
	English-North America	0.9716	0.9921	**0.9936**
	English-West Europe	0.9956	0.9974	**0.9985**
	Hindi-South Asia	0.9845	0.9959	**0.9973**
	Indonesian-SE Asia	0.9759	0.9925	**0.9941**
	Mandarin-East Asia	0.9890	0.9972	**0.9985**
	Russian-East Europe	0.9980	0.9972	**0.9983**
	Spanish-Latin America	1.0006	**1.0011**	1.0007
	Swahili-Sub Saharan Africa	0.9585	0.9937	**0.9955**
Politician	Arabic-West Asia & North Africa	0.9773	0.9942	**0.9955**
	English-North America	0.9959	**0.9984**	0.9976
	English-West Europe	0.9794	0.9954	**0.9967**
	Hindi-South Asia	0.9799	0.9929	**0.9941**
	Indonesian-SE Asia	0.9723	0.9916	**0.9924**
	Mandarin-East Asia	0.9763	0.9948	**0.9961**
	Russian-East Europe	0.9384	**0.9987**	0.9982
	Spanish-Latin America	0.9885	**0.9941**	0.9935
	Swahili-Sub Saharan Africa	0.9436	**0.9978**	0.9971
School Teacher	Arabic-West Asia & North Africa	1.0143	1.0149	**1.0155**
	English-North America	0.9977	0.9976	**0.9981**
	English-West Europe	0.9401	0.9987	**0.9998**
	Hindi-South Asia	1.0000	**1.0006**	1.0003
	Indonesian-SE Asia	0.9860	**1.0015**	1.0012
	Mandarin-East Asia	1.0086	**1.0092**	1.0087
	Russian-East Europe	0.9762	0.9948	**0.9955**
	Spanish-Latin America	0.9659	0.9975	**0.9982**
	Swahili-Sub Saharan Africa	0.9859	**1.0030**	1.0024

Table 3
Overall Image Similarity Score for Professions. Baseline results are from [4].

	ISS_{intra}			ISS_{cross}		
Query	Baseline	with PM	with NA	Baseline	with PM	with NA
CEO	0.9644	**0.9873**	0.9862	0.9846	0.9956	**0.9960**
Engineer	0.9925	0.9980	**0.999**	0.9939	0.9972	**0.9980**
Nurse	0.9862	**0.9967**	0.9931	0.9900	0.9961	**0.9965**
Politician	0.9724	**0.9953**	0.9930	0.9836	**0.9964**	0.9952
School Teacher	0.9860	**1.0020**	1.0010	0.9904	**0.9977**	0.9931
Mean Value	0.9803	**0.9958**	0.9944	0.9885	**0.9966**	0.9957

the baseline (0.8990). A similar trend is observed for the Engineer query, where the NA method outperforms PM in most regions, especially for Russian-East Europe and Spanish-Latin America, where NA achieves ISS$intra$ values of 1.0076 and 0.9981, respectively. For the Nurse query, the NA method again consistently outperforms both the baseline and PM, with a remarkable improvement in the Swahili-Sub Saharan Africa region, where the ISS_{intra} increases from 0.9585 (baseline) to 0.9955. The Politician query also shows substantial gains with both approaches, particularly in Russian-East Europe, where the NA method reaches an ISS_{intra} of 0.9982, an increase over the baseline of 0.9383. Finally, for the School Teacher query, both methods show increased performance in nearly all regions, with the NA method showing slightly higher ISS_{intra} scores, particularly in Arabic-West Asia & North Africa (1.0155) and Spanish-Latin America (0.9982).

4.2.3. Overall ISS Performance: Intra-dataset and Cross-dataset Comparison:

As presented in Table 3, both approaches, PM and NA, consistently outperform the baseline in both ISS$intra$ and ISS$cross$ evaluations. For the CEO query, the cross-dataset score (ISS_{cross}) for the NA method is slightly higher (0.9960) compared to PM (0.9956), showing a marginal improvement over the baseline. A similar pattern is observed for the Engineer and Politician queries, where the NA method again shows higher cross-dataset performance. For Nurse and School Teacher, the PM method performs slightly better in ISS$intra$, but the NA method maintains higher cross-dataset scores. This is particularly evident for the School Teacher query, where the NA method scores 1.001 in ISS$intra$ and 0.9931 in ISS_{cross}, outperforming both the baseline and PM.

When averaged across all queries, the PM method achieves the highest mean ISS$intra$ score (0.9958), while the NA method follows closely with 0.9944, both outperforming the baseline (0.9803). Similarly, for ISS$cross$, PM leads with 0.9966, followed closely by NA (0.9957), both again surpassing the baseline value of 0.9885. These results emphasize the effectiveness of our methods, particularly the PM approach, in increasing dataset diversity both within and across datasets.

4.2.4. Bias reduction in CNNs and ViTs:

As shown in Table 4, CNN models demonstrated improvements in accuracy with the application of the Uniform Noise Blur technique, and in two cases (Inception V3 and Xception) with Partial Mix, though none surpassed the performance of manually debiased data. Inception V3 and Xception showed the most consistent gains, while VGG16 saw only minor improvement. In contrast, Vision Transformers (ViTs) showed no improvements with either augmentation method, indicating that these techniques were ineffective in reducing bias for ViTs. This highlights the need for more tailored approaches for bias mitigation in ViTs, which likely depend on cues beyond facial features.

Table 4
Accuracy of all models on the gender-balanced test dataset. Accuracies higher than the biased dataset are in bold.

Model Type	Model	Biased Accuracy	Partial Mix	Uniform Noise Blur	Unbiased Accuracy
CNN	Inception V3	0.72	**0.73**	**0.74**	0.79
	ResNet 152	0.76	0.76	**0.77**	0.85
	VGG 16	0.57	0.56	**0.58**	0.66
	Xception	0.74	**0.75**	**0.75**	0.79
ViT	ViT B/16	0.55	0.52	0.55	0.57
	ViT B/32	0.50	0.50	0.49	0.57
	ViT L/16	0.39	0.37	0.39	0.40
	ViT L/32	0.56	0.54	0.54	0.60

5. Conclusion

Our study demonstrates that adversarial data augmentation techniques, Partial Mix (PM) and Noise Addition (NA), significantly enhance dataset diversity and reduce gender bias, particularly in CNN models. This is reflected in improvements across both Intra-dataset and Cross-dataset Image Similarity Scores (ISS), indicating a more visually diverse and balanced representation of gender-related features. CNN models trained with these augmented datasets show a noticeable reduction in bias, supporting the effectiveness of targeted facial-region augmentation. However, Vision Transformers (ViTs) do not exhibit the same reduction in gender bias. Despite the use of the same augmentation techniques, ViTs continue to amplify biases, likely due to their ability to learn from broader visual cues, such as clothing and objects, beyond just facial features. This heightened sensitivity makes them more resistant to bias mitigation through facial-focused augmentations, leading to less improvement in diversity and fairness compared to CNNs. In summary, while adversarial data augmentation enhances diversity and mitigates bias in CNNs, it is less effective for ViTs, which require more comprehensive strategies that account for the broader context in which biases are learned. Future work should focus on developing bias mitigation methods that target a wider range of visual signals, particularly for ViTs, to ensure equitable and fair representation in computer vision models.

6. Acknowledgments

This research was supported by Science Foundation Ireland under grant numbers 18/CRT/6223 (SFI Centre for Research Training in Artificial intelligence), SFI/12/RC/2289/P_2 (Insight SFI Research Centre for Data Analytics), 13/RC/2094/P_2 (Lero SFI Centre for Software) at Dublin City University and 13/RC/2106/P_2 (ADAPT SFI Research Centre for AI-Driven Digital Content Technology). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference On Fairness, Accountability And Transparency, 2018, pp. 77–91.

[2] K. Kärkkäinen, J. Joo, Fairface: Face attribute dataset for balanced race, gender, and age, ArXiv Preprint ArXiv:1908.04913 (2019).

[3] A. Wang, A. Liu, R. Zhang, A. Kleiman, L. Kim, D. Zhao, I. Shirai, A. Narayanan, O. Russakovsky, Revise: A tool for measuring and mitigating bias in visual datasets, International Journal Of Computer Vision 130 (2022) 1790–1810.

[4] A. Mandal, S. Leavy, S. Little, Dataset diversity: Measuring and mitigating geographical bias in image search and retrieval, in: Proceedings Of The 1st International Workshop On Trustworthy AI For Multimedia Computing, 2021, pp. 19–25.

[5] A. Mandal, S. Little, S. Leavy, Gender bias in multimodal models: A transnational feminist approach considering geographical region and culture, ArXiv Preprint ArXiv:2309.04997 (2023).

[6] A. Mandal, S. Leavy, S. Little, Measuring bias in multimodal models: Multimodal composite association score, in: International Workshop On Algorithmic Bias In Search And Recommendation, 2023, pp. 17–30.

[7] Y. Zhang, J. Sang, Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing, in: Proceedings Of The 28th ACM International Conference On Multimedia, 2020, pp. 4346–4354.

[8] P. Smith, K. Ricanek, Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing, in: Proceedings Of The IEEE/CVF Winter Conference On Applications Of Computer Vision Workshops, 2020, pp. 90–97.

[9] L. Li, F. Tang, J. Cao, X. Li, D. Wang, Bias oriented unbiased data augmentation for cross-bias representation learning, Multimedia Systems 29 (2023) 725–738.

[10] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition, 2017, pp. 1251–1258.

[11] Z. Wang, K. Qinami, I. Karakozis, K. Genova, P. Nair, K. Hata, O. Russakovsky, Towards fairness in visual recognition: Effective strategies for bias mitigation, in: Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition, 2020, pp. 8919–8928.

[12] C. Schumann, S. Ricco, U. Prabhu, V. Ferrari, C. Pantofaru, A step toward more inclusive people annotations for fairness, in: Proceedings Of The 2021 AAAI/ACM Conference On AI, Ethics, And Society, 2021, pp. 916–925.

[13] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, V. Ordonez, 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, 2019, pp. 5310–5319.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, Others, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv Preprint ArXiv:2010.11929 (2020).

[15] S. Khan, M. Naseer, M. Hayat, S. Zamir, F. Khan, M. Shah, Transformers in vision: A survey, ACM Computing Surveys (CSUR) 54 (2022) 1–41.

[16] A. Mandal, S. Leavy, S. Little, Biased attention: Do vision transformers amplify gender bias more than convolutional neural networks?, ArXiv Preprint ArXiv:2309.08760 (2023).

[17] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data 6 (2019) 1–48.

[18] T. Kumar, M. Turab, K. Raj, A. Mileo, R. Brennan, M. Bendechache, Advanced data augmentation approaches: A comprehensive survey and future directions, ArXiv Preprint ArXiv:2301.02830 (2023).

[19] T. Kumar, A. Mileo, R. Brennan, M. Bendechache, Rsmda: Random slices mixing data augmentation, Applied Sciences 13 (2023) 1711.

[20] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. Berg, Dssd: Deconvolutional single shot detector, ArXiv Preprint ArXiv:1701.06659 (2017).

[21] T. Karras, S. Laine, T. Aila, Nvlabs/ffhq-dataset, 2019. URL: https://github.com/NVlabs/ffhq-dataset.

[22] R. Rothe, R. Timofte, L. Van Gool, Dex: Deep expectation of apparent age from a single image, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 10–15.

[23] E. Learned-Miller, G. Huang, A. RoyChowdhury, H. Li, G. Hua, Labeled faces in the wild: A survey, in: Advances in Face Detection and Facial Image Analysis, 2016, pp. 189–248.

[24] Z. Zhang, Y. Song, H. Qi, Age progression/regression by conditional adversarial autoencoder, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.