Application of Style Transfer Algorithm in Mitigating Gender and Ethnic Biases in Image Datasets

Ken Su(zhengya5@ualberta.ca)
Yuxuan Fu(fu8@ualberta.ca)

1 Abstract

Over the past decade, the success of machine learning models, particularly in image recognition and classification tasks, has been overshadowed by the unveiling of inherent gender and racial biases. These biases originate from training datasets that do not represent diverse populations. This proposal explores the application of style transfer algorithms as a new approach to address and mitigate these biases, with a special focus on gender and ethnic attributes. Utilizing the Generative Adversarial Networks (GAN) driven style transfer proposed by Georgopoulos et al.[5], our goal is to create a more balanced dataset by transferring gender and ethnic features between different images, thereby enhancing demographic representativeness in training datasets. Our approach builds on previous efforts to reduce bias and transfer gender features in image datasets. This research has the potential to make significant contributions to the fairness of artificial intelligence, providing a more equitable foundation for machine learning applications across various fields.

2 Introduction

The rapid advancements in Machine Learning (ML) and Deep Learning (DL) have driven significant improvements in image recognition and classification tasks. However, as the use of these technologies proliferates, inherent biases, especially regarding gender and ethnic representativeness, have come under scrutiny. The root of such biases primarily lies in the training datasets, which often lack diverse demographic representativeness. Numerous studies have highlighted the impact of these biases on the downstream tasks ML models aim to accomplish[2].

Reducing bias in image datasets is crucial for ensuring fairness and inclusivity in artificial intelligence applications. Various techniques have been proposed in the literature to address gender and ethnic biases in image datasets. Among them, this study proposes utilizing image and feature-level semantic preservation enhanced novel bias mitigation techniques as promising solutions[3]. Moreover, exploring gender bias mitigation methods in classification tasks can provide a comprehensive understanding of bias impacts on downstream tasks[3].

Style transfer is an emerging technique that offers a unique approach to tackle these challenges. This method involves transferring style features from one image to another while preserving content, thereby generating new images that eliminate unfairness attributes, as illustrated in Figure 1. Georgopoulos et al. proposed a GAN-based style transfer method capable of creating additional images reflecting

various attributes such as race, potentially mitigating demographic biases[5]. Similarly, efforts to assemble new facial image databases together with inclusive gender and ethnic balanced databases have taken a step towards addressing biases[6].

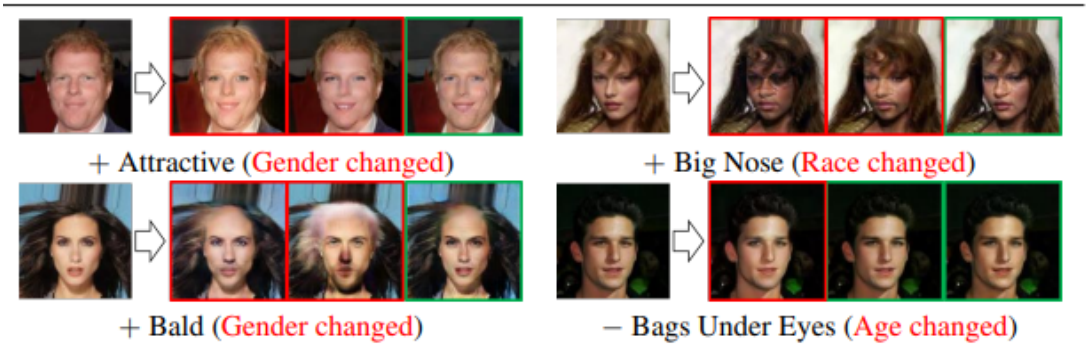

Figure1 : Feature transfer through deep learning algorithms[16]

Our proposed research aims to build upon these foundations by employing style transfer algorithms to generate more balanced and demographically representative datasets. By transferring gender and ethnic features between different images to create a fairer training dataset, it is possible to significantly reduce gender and ethnic biases in ML models. Furthermore, exploring style-based multi-task learning to mitigate demographic biases in facial datasets provides a compelling precedent for our research[4].

Through extensive experiments and comparative analysis with existing bias mitigation techniques, our goal is to reduce the impact of gender and ethnic biases in image datasets on ML models through style transfer algorithms. Our research not only contributes to the existing body of knowledge in reducing biases but also has broader implications for promoting fairness and inclusivity in artificial intelligence technologies.

3 Brief Summary of Existing Work

Style Transfer, Traditional GAN, and Diffusion Models represent key methods in feature transfer techniques, aiding in alleviating biases in image datasets. Style transfer helps in transferring style elements between images while preserving the content. Traditional GAN generates new data resembling a given dataset, aiding demographic representation. Although diffusion models are typically used for network data, they can be extended to image data, modeling feature transfer as a diffusion process across defined networks.

3.1 Style Transfer:

Style Transfer is a technique in computer vision and graphics which generates a new image by merging the content of one image with the style of another, while retaining the original image's content. The process initially provides a style image, transforming another image to that style, and preserving the original image's content as much as possible. An advancement in this field is the utilization of deep neural networks to

separate and recombine image content and style, aiming to transfer any visual style to a content image[10]. Moreover, it has been proven that well-trained convolutional neural networks with sufficiently labeled data are very effective in handling style transfer problems, viewing it as distribution alignment issues[11]. A pioneering work in this field by Gatys et al. introduced a neural algorithm for artistic style[16], defining two distance metrics, one for content and another for style, then optimizing the image to minimize these distances. This technique displays promise in generating demographically diverse images by transferring style features related to gender and ethnicity across different images, aiding in mitigating biases in datasets.

3.2 Traditional GAN (Generative Adversarial Networks):

Proposed by Goodfellow et al., GAN constitutes a class of machine learning frameworks where two networks (generator and discriminator) are trained concurrently through adversarial training. The generator creates new data instances, while the discriminator evaluates them against a real dataset. Over time, the generator learns to produce more realistic data. In the context of bias mitigation, GAN can be used to generate images representing underrepresented population groups, thereby enhancing dataset diversity and reducing inherent biases. Georgopoulos showcased a GAN-based style transfer method for creating additional images reflecting various attributes such as race, potentially mitigating demographic biases[1]. The use of GAN in image applications has been widely explored, with guidelines and surveys discussing their applications in image processing and sequential data processing tasks.

3.3 Diffusion Models:

Traditionally used for network data, diffusion models can be innovatively extended to image datasets. A paper developed a unified framework for image-to-image translation based on conditional diffusion models, evaluating this framework on tasks like coloring, restoration, uncropping, and JPEG recovery[13]. Latent Diffusion Models (LDM) represent another advancement, achieving new state-of-the-art levels in image restoration and competitive performance across various tasks including unconditional image generation, semantic scene synthesis, and super-resolution[14]. In this context, feature transfer can be modeled as a diffusion process across image networks, where nodes represent images and edges represent similarity or relationships between images. By modeling the propagation of gender and ethnic features as a diffusion process, it's conceivable to generate new images or modify existing images to better represent diverse demographics. Although this is a more abstract extension of diffusion models, it represents a novel method in feature transfer and bias mitigation in image datasets.

4 How you Plan To Implement Your Ideas

The domain of fairness in machine learning models generally has three areas: preprocessing (data processing), in-processing (model training), and post-processing, representing three different angles to address or mitigate "unfairness." In this work, we will focus only on preprocessing, operating on images to be trained through style transfer algorithms, reducing the impact of gender and ethnic features on model fairness.

Initially, let the algorithm learn two independent latent spaces with the objective of achieving information separation of protected attributes and target attributes. These two latent spaces should have different objectives: one for mapping the target attributes and the other for retaining the information of protected attributes. This design allows the model to prevent unnecessary transformation of protected attributes while editing target attributes. For example, when editing a protected attribute (gender), even using a style transfer algorithm to transfer male features to female features, it should be able to protect other protected attributes (such as race) from being affected. At the same time, the features to be transferred, in addition to the unique features of males and females, should also extract other features like bold, big eyes, thick lips, and blond hair.

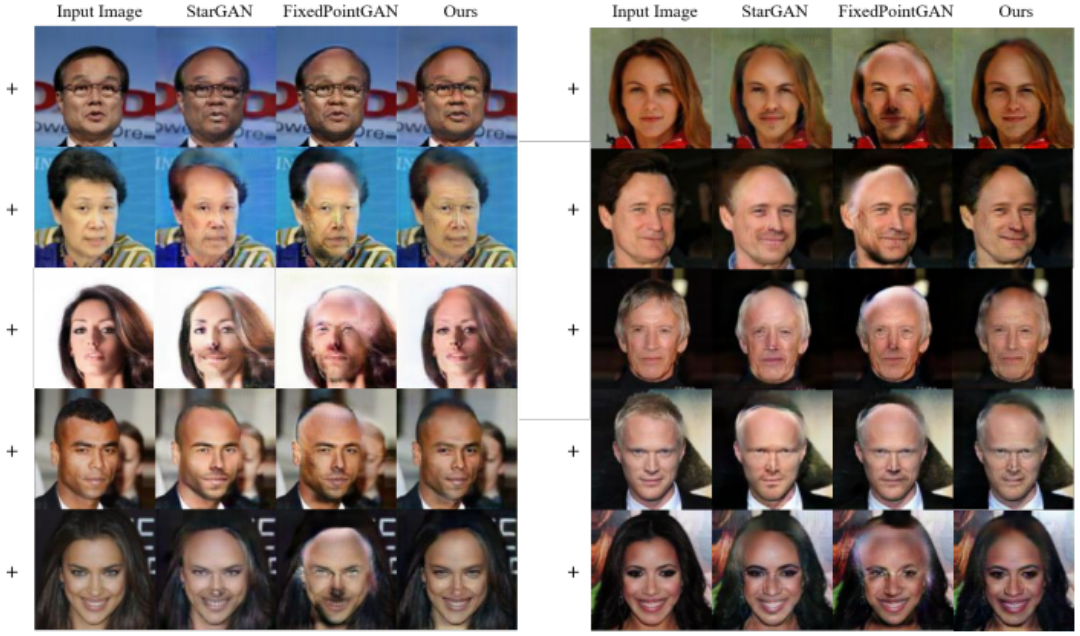

Figure2: Transfer effect of bold attribute

To evaluate the algorithm's effect on fairness, various types of experiments should be conducted on datasets, like using the CeleA dataset, employing style transfer, traditional GAN, diffusion models methods to compare the sizes of fairness indicators.

5 Timeline

Now to Oct 22, 2022: Conducting a literature review, reading an article each day, and preparing to finalize the submission three days before the deadline.

Oct 23, 2023 to Nov 5, 2023: Focusing on a 2D-to-2D style transfer project.

Nov 5, 2023 to Nov 9, 2023: Reading papers related to AI model training data bias mitigation.

Nov 10, 2023 to Nov 20, 2023: Concentrating on improving the implementation of the style transfer project.

Nov 20, 2023 to Nov 30, 2023: Checking code correctness and debugging.

Dec 1, 2023 to Dec 5, 2023: Creating a presentation in PowerPoint and rehearsing.

Dec 5, 2023 to Dec 18, 2023: Double checking the implementation and finishing up the final report.

6 Description of 5 Labs

First lab: try to successfully execute and understand the code from our reference article. When that goal has been accomplished, we will discuss with Guangfang how best to continue in this direction while exchanging initial concepts.

Second lab: we will begin the implementation phase for 2D Style Transfer Mitigation Data Bias.

Third Lab: For this lab, our main aim will be addressing any difficulties experienced while working on 2D Style Transfer Mitigation Data Bias. Guangfang will offer guidance while our classmates collaborate towards finding solutions collaboratively.

Fourth Lab: Our fourth laboratory session will consist of engaging in dialogues with both Guangfang and fellow classmates regarding related research areas, specifically regarding 3D texture stylization as a means of mitigating data bias during 2D style transfer work.

Fifth Lab: For our fifth lab session, we will come together as a group and address practical issues regarding 2D Style Transfer Mitigation Data Bias implementation.

7 Literature Review

7.1 Research Field and Objectives

7.1.1 Research Field:

My research in machine learning centers on mitigating inherent gender and racial biases found in image recognition and classification tasks. By reviewing relevant literature, I learned that biases stemmed largely from an absence of training datasets representing diverse populations. Faced with these difficulties, I decided to investigate style transfer algorithms as an innovative means of correcting and mitigating biases related to gender and racial attributes. Georgopoulos et al. have proposed using Generative Adversarial Networks (GAN) driven style transfer as part of this research to create more balanced datasets by sharing gender and race features among various images thereby expanding demographic representation within training datasets. This research draws closely on my earlier considerations and builds upon earlier efforts at

eliminating bias and transferring gender features in image datasets to advance artificial intelligence with greater fairness, which in my view provides a fairer foundation for multidisciplinary machine learning applications.

7.1.2 Research Objectives:

My research centers around improving the fairness of machine learning models during preprocessing; specifically during image manipulation through style transfer algorithms to minimize gender and race effects on model fairness. Step one involves instructing an algorithm to learn two separate latent spaces to effectively separate information regarding protected attributes from target attributes. These two latent spaces serve two different functions; one for mapping target attributes while retaining information of protected attributes; this design strives to avoid unintended transformations of protected attributes while editing target attributes. Example: when editing protected attributes like gender, even when employing style transfer to change male features to female features using male features as the target attribute, other protected attributes (such as race) must remain unaffected. My research seeks to develop a robust preprocessing framework, contributing significantly towards my overarching goal of increasing fairness and justice within machine learning models. Through careful design, the features to be transferred should include both distinguishable male and female characteristics as well as features such as baldness, big eyes, thick lips and blonde hair - features which could prove particularly important when translating characters onto images for computerized translation systems.

7.2 Related Work

This section reviews key literature relevant to this research, outlining key findings and contributions. A particular emphasis will be given to applications of machine learning, image recognition and style transfer technology in combatting biases and increasing fairness.

7.2.1 Matched Sample Selection with GANs for Mitigating Attribute Confounding

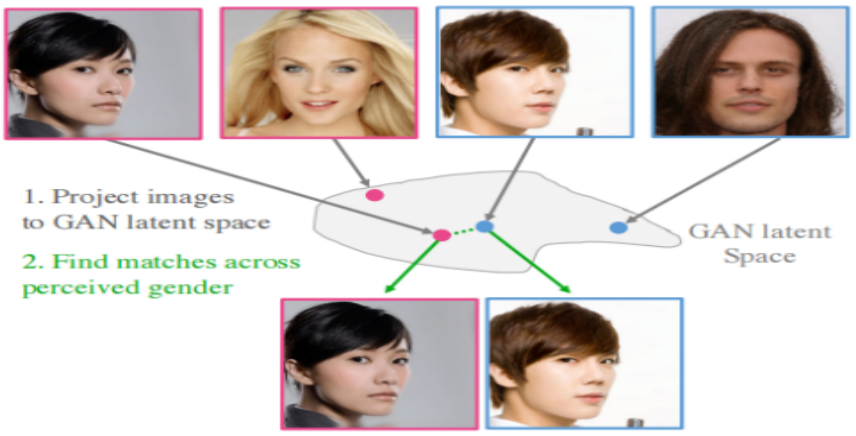

Figure 1. Finding matched samples by GAN latent-space projection.[17]

Authors: Chandan Singh, Guha Balakrishnan, Pietro Perona (2021)

Summary: The use of Generative Adversarial Networks (GANs) to select matched samples and reduce attribute confounding—a crucial first step in decreasing bias within datasets—is examined in this research by Chandan Singh and his colleagues. This paper focuses mostly on bias problems that frequently occur in machine learning datasets. The authors suggest a cutting-edge method that makes use of GAN technology to produce images that closely resemble the original dataset. To extract meaningful feature pairs from the source photos, they present a culling approach. They also present a matching approach to improve attribute matching accuracy. The writers optimize the generator's and discriminator's loss functions to further improve matching accuracy. This research presents new methods for matching sample selection and attribute confounding mitigation that can be used to produce more equal and less biased datasets for machine learning problems.

Key Innovations:

1. Introduction of GAN technology for generating images akin to the original images.
2. Culling strategy for extracting useful feature pairs from original images.
3. Matching strategy for default attributes to enhance matching accuracy.
4. Optimization of the loss functions of the generator and discriminator to further improve matching accuracy.

7.2.2 FairFaceGan

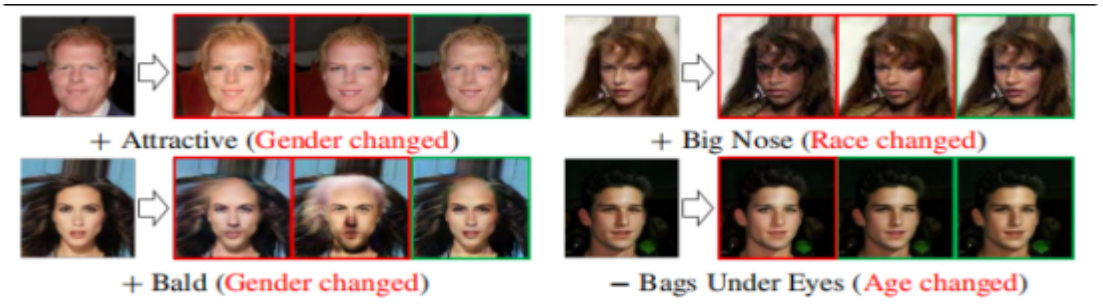

Figure 2. Image translation results on CelebA dataset [16]

Authors: Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, Hyeran Byun (2020)

Summary: Sunhee Hwang and colleagues provide in their article entitled FairFaceGan an innovative strategy to ensuring fairness within facial recognition systems, through style transfer technology and other measures. The authors propose the FairFaceGAN model as an attempt to address bias issues associated with image-to-image translation models. It features components like Protected Attribute Classifier (PAC) and Fair Attribute Retention Network (FAR), among others. Key to this approach is Fair Feature Retention Loss (FPR), an approach aiming to better preserve protected characteristics such as gender and race in image translation processes. In addition, innovative techniques for achieving fairness in facial

recognition and translation, style transfer technology can eliminate bias; and style transfer technology may even eliminate bias completely. This article describes these advances further and details innovative techniques designed to achieve fairness during these processes.

Key Innovations:

1. Proposal of FairFaceGAN model to address bias issues in image-to-image translation models.
2. Introduction of Protected Attribute Classifier (PAC) and Fair Attribute Retention Network (FAR) modules.
3. Adoption of Fair Feature Retention Loss (FPR) loss function to better retain protected attributes.

7.2.3 Arbitrary Style Transfer with Deep Feature Reshuffle

Figure 3. Comparison with previous neural style transfer methods.[10]

Authors: Shuyang Gu, Congliang Chen, Jing Liao, Lu Yuan (2018)

Summary: Shuyang Gu and his team explored in their paper entitled, "Arbitrary Style Transfer with Deep Feature Reshuffle", the concept of arbitrary style transfer using deep feature reshuffling. The authors present innovative concepts such as multi-resolution processing that enables networks to avoid local optima at different input resolutions and global and local style losses that enable their learning at both scales simultaneously. Image segmentation techniques are utilized as well in order to enhance feature capture from input images and this paper offers key insight into style transfer algorithms which may prove instrumental in mitigating biases associated with image recognition processes.

Key Innovations:

1. Introduction of multi-resolution processing concept, enabling the network to better avoid local optima at different input resolutions.
2. Introduction of global and local style losses, allowing the network to learn features on a global scale while also better learning details locally.
3. Introduction of adaptive layer normalization technology, enabling the network to maintain similar performance across different layers and resolutions.
4. Introduction of image segmentation technology, enabling the network to better capture feature information in input images.

7.2.4 Mitigating Demographic Bias in Facial Datasets

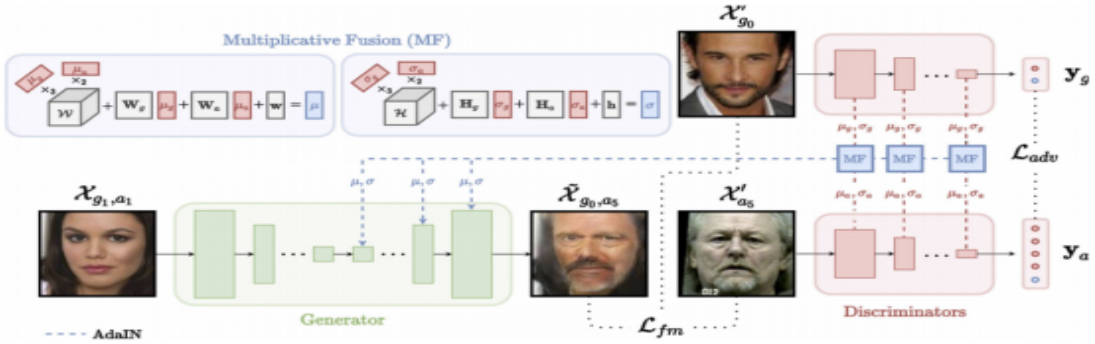

Figure 4.Overview of the proposed method for multi-attribute transfer[2]

Authors: Markos Georgopoulos, James Oldfield (2021)

Summary: Markos Georgopoulos and James Oldfield's research paper entitled, "Mitigating Demographic Bias in Facial Datasets", falls perfectly into line with our objectives for mitigating demographic bias within facial recognition datasets. Key innovation of this paper lies in its introduction of an unprecedented style transfer GAN framework capable of simultaneously moving multiple demographic attributes at the same time, by setting different target image sets, this framework can produce unique images for every attribute class while still protecting those unaffected by editing. This paper centers around developing an impartial dataset which accurately reflects demographic diversity. Such an innovative solution provides a promising way of combatting biases inherent to machine learning applications while at the same time increasing fairness for facial recognition systems and image-related tasks.

Key Innovations:The main innovation of this paper is the invention of a novel style transfer GAN framework, which is capable of transferring multiple demographic attributes simultaneously. By adjusting different target image sets, our framework can generate different images for each attribute class.

7.2.5 How to Boost Face Recognition with StyleGAN

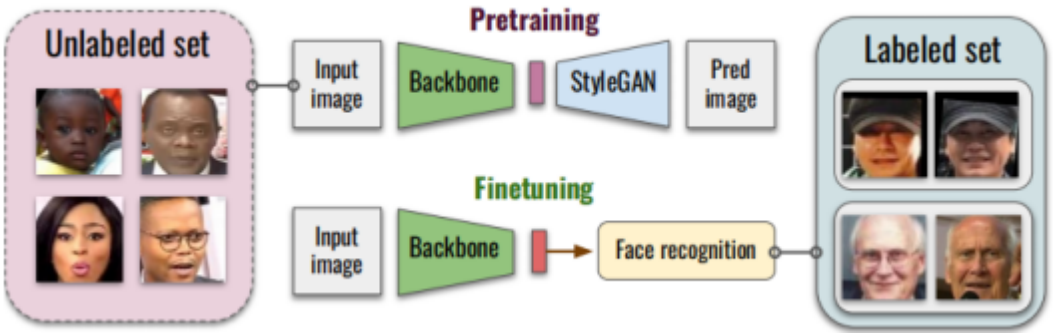

Figure 5. Our method is aimed at boosting the performance of face recognition.[15]

Authors: Artem Sevastopolsky, Yury Malkov, Nikita Durasov (2023)

Summary: StyleGAN2 can aid system improvements for facial recognition systems.

Furthermore, the authors present an innovative technique for improving image classification models using StyleGAN2-ADA to further bolster performance. Training techniques typically use self-supervised learning techniques to bolster a model's robustness, leading to enhanced stability and performance. New data augment techniques are presented here in order to further boost model robustness. Also included is an innovative data augmentation method to strengthen model robustness. We thoroughly investigated GAN-based techniques as machine learning model enhancers to boost accuracy and fairness - specifically how these can reduce bias during image classification, increase face recognition model accuracy and thus meet our larger aim of combatting bias through machine learning applications.

Key Innovations:

1. New method leveraging StyleGAN2-ADA for enhanced image classification model performance.
2. New data augmentation technique for improved model robustness.
3. Self-supervised learning techniques during training for enhanced model performance and robustness.

7.2.6 Synthesis

Each paper reviewed offers unique approaches and methodologies for mitigating biases and increasing fairness within image recognition and classification tasks using machine learning, particularly within machine learning image classification tasks. By carefully studying and comparing these works, our goal is to gain a solid understanding of existing solutions, challenges and future directions within this field; all key to accomplishing our overarching research goals.

7.3 Similarities and Differences

7.3.1 Similarities

1. Research Goals: All papers aim at leveraging image processing and machine learning techniques to mitigate biases and enhance fairness, specifically in facial recognition and image classification tasks.

2. Methodological Applications: The majority utilize Generative Adversarial Networks (GAN) and style transfer technology to achieve their goals, demonstrating the potential of these techniques in addressing image biases and promoting model fairness.

3. Dataset Utilization: Common image datasets like ImageNet and CelebA are frequently employed for validation and testing to gauge the effectiveness and performance of the proposed methods.

7.3.2 Differences

1. Specific Methods and Techniques: Some papers focus on balancing datasets and mitigating bias through image matching and style transfer (e.g., "Matched Sample Selection with GANs for Mitigatively Attribute Confounding", or "Mitigating Demographic Biases with Style Transfer") while other delve into optimizing facial recognition model performance through pre-trained models with unlabeled data (e.g. "How to Boost Face Recognition With StyleGAN").

2. Experimental Design and Results: Divergent experimental designs and result evaluations are found, for instance, "Arbitrary Style Transfer with Deep Feature Reshuffle" primarily examines style transfer performance, while "FairFaceGan" emphasizes on fairness augmentation.

3. Application Domains: Some papers focus on image editing and style transfer techniques like Arbitrary Style Transfer with Deep Feature Reshuffle" and FairFaceGan" while others explore facial recognition and dataset bias reduction such as How to Increase Face Recognition With StyleGAN or Mitigating Demographic Biases with Style Transfer".

7.4 Conclusion

These papers present valuable research that sheds light on the power and utility of Generative Adversarial Networks (GANs) and style transfer technologies as effective strategies to minimize bias within image datasets. These contributions represent a crucial advance toward combatting bias within machine learning applications. These papers showcase an emerging convergence towards creating more balanced datasets. Achieving this aim could pave the way to developing machine learning models which not only deliver equitable predictions but are more accurate as well.

These papers have had far-reaching implications beyond just academia; their influence can be felt across technologies and societies alike. I found these papers inspiring as they provided me with new understanding regarding potential technical solutions for bias-related challenges such as model design/development/evaluation/utilization issues affecting machine learning applications. From these studies came new tools with which I can contribute meaningfully towards discussions surrounding bias mitigation within machine learning research endeavors.

These papers have inspired an immense curiosity for exploring uncharted territories of emerging methods and cutting-edge technologies, especially machine learning environments characterized by equity and justice - these research contributions instilled within me the resolve and commitment necessary to explore and pioneer new avenues for advancement towards fairness, inclusivity and reliability in machine learning in its future form.

8 Reference

[1]Krishnan, A., & Rattani, A. (2023). A novel approach for bias mitigation of gender classification algorithms using consistency regularization. Image and Vision Computing, v.11 art. no. 104793.
[2]Markos Georgopoulos, James Oldfield.International . Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer .IEEE.2021-15.
[3]Didan Deng, Zhaokang Chen, Bertram E. Shi. Multitask Emotion Recognition with Incomplete Labels.IEEV.2020-10.
[4] Xinke Shen, Xianggen Liu, Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition.IEEE.2021-09.
[5]Wenying Wu, Pavlos Protopapas, Gender Classification and Bias Mitigation in Facial Images.2020-13
[6]Tony Sun, Andrew Gaut. Mitigating Gender Bias in Natural Language Processing: Literature Review.ACK.2019-06.
[7]Artūrs Stafanovičs, Toms Bergmanis. Mitigating Gender Bias in Machine Translation with Target Gender Annotations.EMNLP.2020-10
[8] Yijun Li, Chen Fang. Universal Style Transfer via Feature Transforms.NLPS.2017-05.
[9] Fujun Luan, Sylvain Paris,.Deep Photo Style Transfer. Computer Vision and Pattern Recognition.2017-04
[10] Shuyang Gu, Congliang Chen. Arbitrary Style Transfer with Deep Feature Reshuffle. Computer Vision and Pattern Recognition.2018-05
[11] Zhou Yin, Wei-Shi Zheng.Adversarial Attribute-Image Person Re-identification. Computer Vision and Pattern Recognition.2018-07.
[12] Leon A. Gatys.Controlling Perceptual Factors in Neural Style Transfer.CVPR.2017-05.
[13] Leon A. Gatys. Image Style Transfer Using Convolutional Neural Networks.CVPR.CVPR.2016-05
[14] Robin Rombach, Andreas Blattmann .High-Resolution Image Synthesis with Latent Diffusion Models
[15]Artem Sevastopolsky, Yury Malkov.How to Boost Face Recognition with StyleGAN?. ICCV 2023-6.
[16]Sunhee Hwang, Sungho Park, Dohyung Kim.FairFaceGAN: Fairness-aware Facial Image-to-Image Translation.BMVC 2020-11.
[17]Chandan Singh, Guha Balakrishnan, Pietro Perona.Matched sample selection with GANs for mitigating attribute confounding.Computer Vision and Pattern Recognition 2021-05.