Debiasing Classifiers Using Latent Space Manipulation

by

Elif Sema Balcıoğlu

Berkay Döner

Merve Rabia Barın

Submitted to the Department of Computer
Engineering in partial fulfillment of
the requirements for the degree of
Bachelor of Science

Undergraduate Program in Computer Engineering

Boğaziçi University

Spring 2022

Debiasing Classifiers Using Latent Space Manipulation

APPROVED BY:

	Prof. Tunga Güngör
Pınar Yanardağ	.....
	(Project Supervisor)

DATE OF APPROVAL: 13.06.2022

ACKNOWLEDGEMENTS

This publication has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TUBITAK (Project No:118c321). We also acknowledge the support of NVIDIA Corporation through the donation of the TITAN X GPU and GCP research credits from Google.

ABSTRACT

Debiasing Classifiers Using Latent Space Manipulation

As the applications of machine learning models in the real world increase by the day, some research efforts have focused on revealing and mitigating biases in these models that are learned from imbalanced data. In this work, we propose a framework that aims to explain the factors behind a classifier's decisions and make it insensitive to changes in the biased factors, focusing on deep learning classifiers for human faces. We first exploit the disentanglement property of StyleGAN2's [1] StyleSpace to identify potential biases of a classifier. Then we compare two possible debiasing methods. For the first method, we manipulate the style channels that control the biased attributes to produce a balanced dataset on which we train a fairer classifier. Second method uses ensemble method on test time. We manipulate the style channels that control the biased attributes and generate several images. We then take the average of the predictions on these images as the final result. We perform several experiments and show results both quantitatively and qualitatively.

ÖZET

Saklı Uzay Manipülasyonları ile Sınıflandırı Modellerin Yanlılığının Azaltılması

Makine öğrenmesi modellerinin gerçek dünyadaki uygulamaları her geçen gün arttıkça, bazı araştırmalar, dengesiz verilerden öğrenilen bu modellerdeki yanlılıkları ve önyargıları ortaya çıkarmaya ve azaltmaya odaklanmıştır. Bu çalışmada, insan yüzleri için kullanılan derin öğrenme sınıflandırıcılarına odaklanarak, bir sınıflandırıcının kararlarının arkasındaki faktörleri açıklamayı ve bu sınıflandırıcıları yanlı faktörlerdeki değişikliklere karşı duyarsız hale getirmeyi amaçlayan bir metot öneriyoruz. İlk olarak sınıflandırıcının potansiyel yanlılıklarını belirlemek için StyleGAN2'nin saklı uzayının ayrık yapısından yararlanıyoruz. Daha sonra iki yanlılık azaltma yöntemini karşılaştırıyoruz. İlk yöntemde, yanlı öznitelikleri kontrol eden stil kanallarını değiştirerek üzerinde daha adil bir sınıflandırıcı eğittiğimiz dengeli bir veri kümesi üretiyoruz. İkinci yöntem ise, test zamanında topluluk yöntemini kullanıyor. Yanlı öznitelikleri kontrol eden stil kanallarını manipüle ederek birkaç yeni imge üretiyoruz. Daha sonra nihai sonuç olarak bu imgeler üzerindeki ortalama sınıflandırıcı tahminini alıyoruz. Metotlarımızı kullanarak birçok deney yapıyoruz ve sonuçları hem nicel hem de nitel olarak gösteriyoruz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

CelebA	Large-scale CelebFaces Attributes (CelebA) Dataset
FFHQ	Flickr-Faces-HQ Dataset (FFHQ)
GAN	Generative Adversarial Networks

1. INTRODUCTION AND MOTIVATION

Since machine learning models are data-driven, models trained on real data will likely learn the biases of the data. Given that most popular datasets have unbalanced data (such as CelebA, as pointed out in [2] and displayed in 3.3) or come from an unknown distribution (FFHQ [3]) across important attribute classes such as gender and skin color, we cannot rely on the decisions of models trained on such datasets for critical real-world applications as these models are not experienced enough to make correct decisions about underrepresented features.

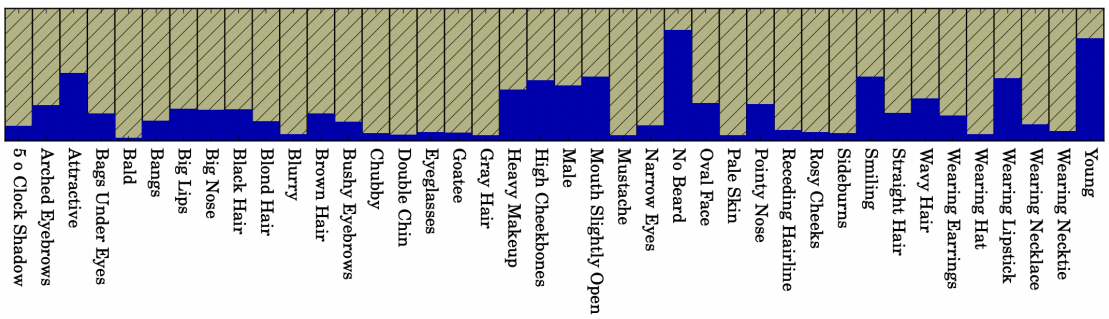


Figure 1.1: **Label Distribution of CelebA** [2] As can be seen, some categories are highly imbalanced, which may cause biases on trained classifiers.

In an interesting study on the reliability of facial recognition technologies, Gender In Shades [4] measured the accuracy of gender classification products developed by global companies (*Microsoft, IBM, etc.*) on a balanced dataset and compared the error rates between four different intersectional groups (*lighter male, darker male, lighter female, darker female*). While the classification accuracy of males is higher than that of females, darker females are the most frequently misclassified group among the four groups. The results show that gender classifiers have a skin tone bias. Apart from the performance issues, these sensitive attributes should not affect the prediction of the classifier at all to avoid negative results against certain individuals. These protected attributes are identified by experts to prevent discrimination. For example, the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA) identified several protected attributes such as race, gender, and religion [5]. All of these analyses indicate that a debiasing framework should be able to both discover the factors that influence

the decision-making process and eliminate the bias caused by the biased attribute by ensuring that the classifier's prediction is not affected by the changes in the biased attribute.

In this paper, we focus on biases introduced by correlations in real world data. If some features are frequently co-occurring in the training data, classifiers may not bother to learn these attributes separately, and their decision may be influenced by the presence of the entangled feature. For example, if most of the people wearing eyeglasses in a dataset are old, an age classifier may predict a higher age for a young person with eyeglasses.

To this end, we propose a framework that detects a biased feature for a pre-trained classifier based on a correlation in the dataset and trains a new classifier with a balanced dataset generated by StyleGAN2 to de-bias the classifier for that feature. The generated dataset is balanced according to the biased attribute and contains image pairs, one of which is the negated version of the other with respect to the biased attribute.

Our pipeline could be summarized as follows:

- We find attributes affecting the classifier's decision.
- We detect a biased attribute among the identified attributes using human interpretation.
- We create a balanced dataset with respect to the biased attribute by manipulating the corresponding style channel.
- We debias a classifier with the new synthetic dataset.
- We improve the decision of the classifier by ensembling the predictions on the original and the manipulated images on test time.

2. STATE OF THE ART

GANs: Generative Adversarial Networks (GANs) are deep learning frameworks that include two models, the generative and the discriminative model [6]. While the generative model tries to understand the distribution of the training data and generate images that are similar to the training dataset, the discriminative model focuses on identifying whether the input image is from the training dataset or generated by the generative model. The main goal of GANs is to generate images that are indistinguishable from the dataset.

StyleGAN [7] and StyleGAN2 [1] are two popular GAN models capable of generating high-quality images based on the distribution of their training dataset. The former improves on the basic GAN architecture by first mapping the random noise input to an intermediate latent space $\mathcal{W}$ and then using instance normalization by affine transformation of $\mathcal{W}$. The latter addresses several artifacts on the images generated by StyleGAN and eliminates them by replacing the instance normalization with modulation layers on the StyleGAN architecture. In this work, we used the pre-trained StyleGAN2 model due to its success.

Latent Space Manipulation: A line of research has focused on discovering attribute-specific directions in the latent spaces of GANs to manipulate the semantics of the generated image. Some supervised methods [8–11] use classifiers trained on specific attributes to discover corresponding directions. [12] identifies attribute-specific control channels using a small number of positive samples. Alternative methods [13–17] find meaningful latent directions using unsupervised approaches.

Fairness & Bias in Classifiers: Several works have focused on uncovering the decision-making processes of classifiers. Explaining In Style [18] identifies the key attributes that influence a classifier's decision. It embeds classifier-specific attributes (such as *retinal fundus* for retinal disease classifier) into the StyleSpace of StyleGAN2 by adding a classifier loss and conditioning the model on classifier outcomes during the

training process. In the final step, the model extracts the most effective style channels for the classifier results. The attributes controlled by these channels reveal what the classifiers learn and allow us to explain biases in these models. Another work [19] proposes the total variation loss and orthogonalization penalty to discover unknown biases of a classifier with minimal human effort. [20] augments GAN generated debiased data with real-world data, and then uses this data to train a fairer classification model than a classifier trained on only real-world data. It trains classifiers for the target attributes (classified) and the protected attributes (biased) and finds a hyperplane that separates the latent space (z-space) of GAN for these two attributes. Then, it randomly samples a z-vector and a pair vector is found that has the same score on the target attribute's classifier but is negated with respect to the protected attribute's score. The work of [21] is another attempt to debias a classifier. They train an initial classifier by amplifying its biases and debias another classifier with a loss that is reweighted with the score of amplified model. Although our baseline methodology is similar to that of [20], we manipulate the images using the *Style Space* [12], which has been shown to be the most disentangled space and is more promising compared to other latent spaces, since only the desired attribute is changed when the images are manipulated. Furthermore, we propose an approach to detect the biases of the classifiers with minimal human effort compared to [20] where biases are not systematically identified.

Ensemble Methods: Ensemble methods are frameworks that combine predictions from multiple base models to improve accuracy. Several ensemble methods, including the Random Forest algorithm [22] and XGBoost [23], have been demonstrated to be effective on regression and classification tasks. [24] proposes a method for increasing the accuracy of classifiers in the image domain by combining the scores of natural variations of the original image by perturbing its latent code in StyleGAN2 domain. Rather than using different models, this method "ensembles" the scores on different views of the same image. Authors created manipulations of the original image in their experiments by mixing the coarse and fine layers of the latent codes, which is limited due to the entanglement of the layers. In our experiments, we follow the same approach by calculating the scores on the views of the original image; however, we use the recently discovered *Style Space* [12] to carry out disentangled and accurate

manipulations. Furthermore, we manipulate the original image towards the biased attribute; thus, we aim to reduce the existing bias as well as increase the accuracy by using ensembling.

3. METHODS

3.0.1. Background

Generator $\mathcal{G}$ of StyleGAN2 structure functions as a mapping function $\mathcal{G} : \mathcal{Z} \to \mathcal{X}$ from input latent space $\mathcal{Z}$, to target image domain $\mathcal{X}$. The underlying structure include numerous latent spaces, $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{W}+$ and $\mathcal{S}$. The latent code $\mathbf{z} \in \mathcal{Z}$ is sampled from a prior distribution $p(\mathbf{z})$, generally chosen to be Gaussian. From $\mathbf{Z}$, a mapping-network $f : \mathcal{Z} \to \mathcal{W}$ that consists of 8 fully connected layers, produces the intermediate latent space $\mathcal{W}$. $\mathcal{W}+$ space is a version of $\mathcal{W}$ where different intermediate latent vectors are fed to each layer of the network. Each $\mathbf{w} \in \mathcal{W}$ is mapped into channel-wise style parameters s using learned affine transformations. The space spanned by style parameters s are denoted as *Style Space* [12] $\mathcal{S}$.

In this paper, we use the style space $\mathcal{S}$ to perform manipulations, as $\mathcal{S}$ space has been shown to be the most disentangled, complete and informative space [12] compared to $\mathcal{W}$ and $\mathcal{W}+$.

3.0.2. Training a Biased Classifier

First, we attempt to train a classification model on real images that learns the biases of this real-world training dataset. This *real* model constitutes our baseline model, which we examine according to its biases to debias it. For our classification task, we use MobileNetV2 [25], an efficient CNN-based architecture designed for mobile applications and comparable to the state-of-the-art models. Its architecture includes 19 residual bottleneck layers after a fully convolutional layer with 32 filters. We train this architecture with a labeled dataset balanced for the classifier's target attribute. The training setting is described in detail in Section 4.0.1.

3.0.3. Identifying Style Channels

Our first goal is to find attributes that influence a classifier's decision, such as. *hair color* for an *age classifier*. [18] identifies style channels controlling these attributes after a classifier incorporated training process for StyleGAN2. Our goal is to capture these channels using a more lightweight method. In [12], the authors use a method to find relevant style channels for an attribute that requires only 20-30 positive samples for that attribute. Inspired by their method, we select k samples that are positively scored by the classifier and obtain channels that control the classifier's score for its target attribute. Instead of finding channels for a pre-determined attribute as in [12], we hope to find classifier-specific channels by integrating the classifier into the positive sample selection part.

The intuition behind the method proposed by [12] for identifying style channels is to compare the style vectors of the positive samples with the population statistics, and select the channels that deviate the most from the population statistics as the most relevant channels for our target attribute. The whole method can be summarized as follows:

- Randomly generate M = 10,000 style vectors and calculate the population mean and standard deviation for each style channel over these vectors.
- Generate k = 300 classifier positive examples.
- Normalize the style vectors of the positive examples using the population mean and standard deviation.
- Calculate mean and std over the normalized style vectors.
- Select the best 20 channels whose mean/std ratio is high because they are more relevant to the target attribute.

We do not select these channels directly, but first test their effects on the classification results by manipulating them in positive and negative directions by the magnitude of a scalar λ. Manipulating a style channel in the positive direction simply means adding λ to the value of the channel, and subtracting λ means proceeding in

the negative direction. If manipulating a channel increases or decreases the classifier's score for an image by a certain threshold in one direction or the other, we argue that this channel plays a role in the classifier's decision making process.

3.0.4. Bias Identification

After extracting the most effective channels, human subjects must interpret what those channels control on the generated image. In this step, a human can evaluate these attributes to see if it makes sense for these attributes to affect the classification result or if they were selected as the top channels due to a faulty correlation in the training data. For example, it makes sense to select a channel that controls dark/white hair color for an age classifier, since white hair color is mostly associated with older people. However, a channel that controls the attribute of wearing glasses should not be selected, because people at any age can wear glasses, thus, this should be interpreted as a *biased attribute*.

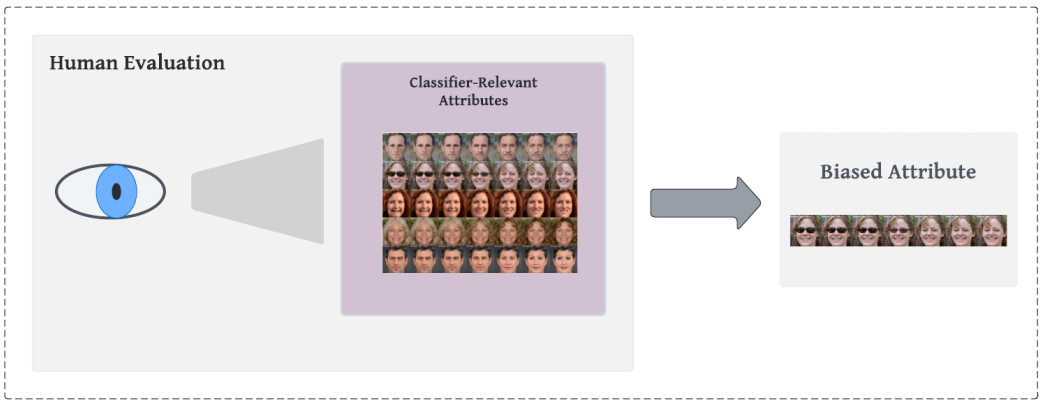


Figure 3.1: **Bias Identification Pipeline** Human subject evaluates classifier-related attributes identified in Section 3.0.3 and decides on biased attributes.

3.0.5. Synthetic Debiased Dataset

First, we sample 10k images from a labeled dataset and extract the target and biased attribute labels of the images. Then we map these images to the style space using an encoder network [26] and save their style codes. We manipulate images using the style channel identified in the Section 3.0.3 that have a negative biased attribute

Figure 3.2: Top 9 style channels identified by our bias identification method for age classifier. Rows correspond to different channels with associated semantic attributes, if identified any, and columns correspond to different manipulation strengths. Classifier scores for each image are displayed at the upper left corner.

label in the positive direction to add a biased attribute to them, and vice versa for positive samples of biased attribute. In the end, we have 10k pairs of images, each pair containing one negative and one positive example of the biased attribute. The fair dataset synthesis process may further be observed in Figure 3.3.

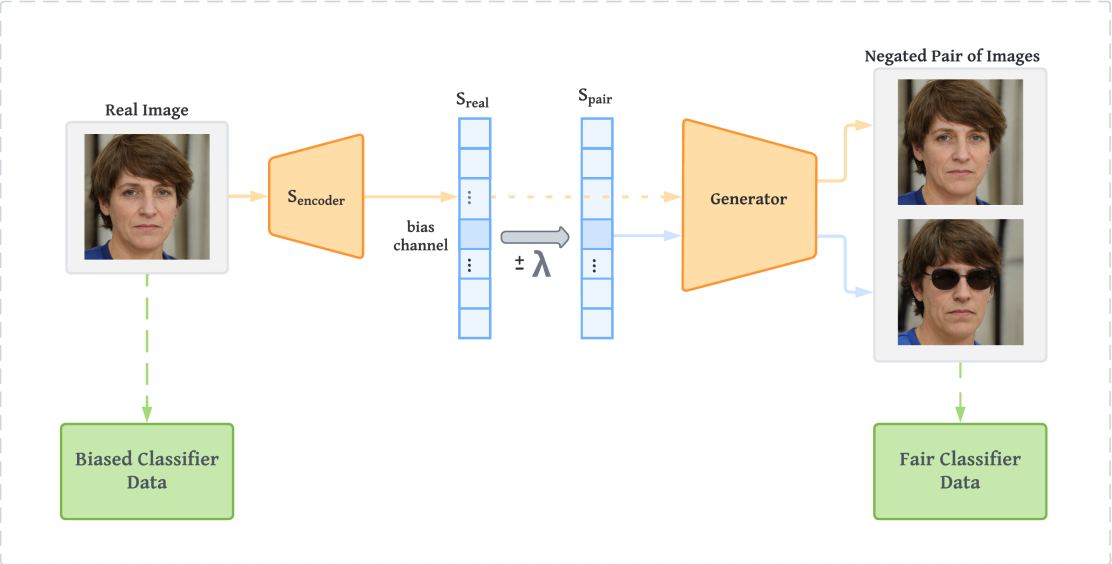


Figure 3.3: **Fair Dataset Generation Pipeline** We generate a pair for each image in the dataset by negating the biased attribute in the original image. The negation is done by shifting the bias channel in the style vector of real image by a certain manipulation strength, λ. The bias channel is previously identified in Section 3.0.4

3.0.6. Debiasing the Classifier with Data Augmentation

With the synthetic and balanced dataset of 20k images generated in Section 3.0.5 we train a new classifier in the same setting with the classifier trained on real images, which we call the *fair* classifier.

3.0.7. Debiasing the Classifier using Ensembling

While retraining on new data can reduce bias, train data is not always available for data augmentation. Therefore, we also propose a method based on ensembling the predictions of the model that can reduce the bias even without train data. On the

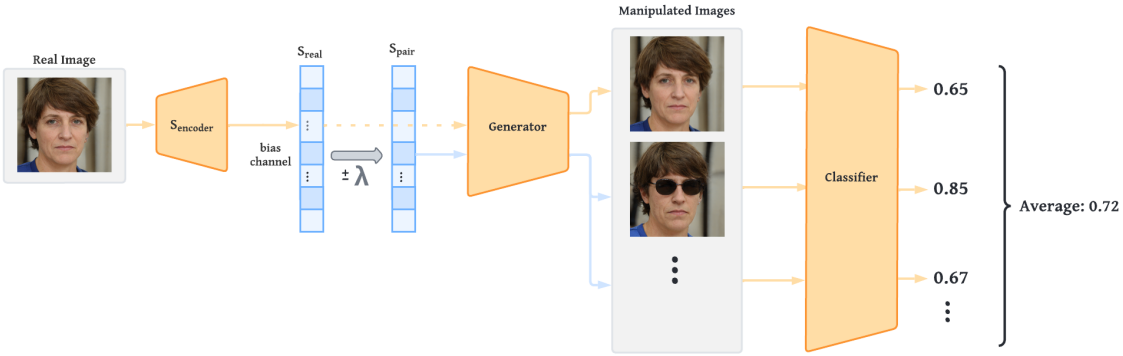


Figure 3.4: **Ensemble Method** We encode the original image and manipulate it in the identified style channels in both positive and negative directions. Then, we calculate the final prediction as the average of the scores on the original and the manipulated images.

test time, we encode the given image, manipulate it in the positive and the negative directions of the style channels identified by our method and then calculate the scores for each manipulation and the original image. Finally, we combine all the scores and use the average of the scores as the final prediction. In our case, we manipulated images in the 120th and 228th channels of the third layer, both of which were identified as being related to glasses. Our ensemble method can be seen on Figure 3.4.

4. RESULTS

We evaluate our framework with several experiments focused on individual components. First, we check if our method can identify the underlying biases in the decision process of classifiers. Then, we check if manipulations carried out using style channels result in the expected semantic differences in the images. Thirdly, we evaluate our method based on the distribution of target attribute and biased attribute. Finally, we evaluate the performance of our method using several quantitative metrics.

4.0.1. Experimental Setup

For our experiments, we chose to train and examine age classifiers that label young people as 1 and old people as 0. We use the CelebA [27] dataset, which contains various attribute labels, including age labels. We use MobileNetV2 [25] architecture for our age classifiers, which are trained on real and manipulated images. We train these classifiers with binary cross-entropy loss using the Adam optimizer [28] and a learning rate = 0.0002 for 30 epochs with a batch size of 4. We have noticed that the training dataset is highly imbalanced with respect to the age attribute, with young people being in the majority. This imbalance affects the performance of the classifiers. Therefore, we selected a subset of the data that is balanced with respect to age.

4.0.2. Bias Identification

To check if our method can identify biases, we display the top 9 channels that are identified to be effective in the decision making process. As can be seen from the Figure 3.2 where we display the effects of several style channels using different manipulation strengths with corresponding classifier scores, our method is able to detect several semantic changes that can be associated with aging. For example, in the first row, our method was able to find a channel that controls the length of the hair where the balder hair is classified as older compared to the longer hair by our classifier. In the 6th row, we observe the effect of the 120th channel of the third layer, which adds glasses in the

negative direction and causes the classifier to perceive the person as older. We have focused on eliminating this bias in our framework. Third to the last row represents another bias in the dataset, where the identified channel changes the gender expression and associates masculine attributes with aging. In the last two rows, we can observe the change in the eyebrow and hair color, ranging from white to black where white is associated with aging. Note that, even though the style channel responsible for adding and removing glasses is not at the very top, it is the channel that constitutes a bias. Other style channels control attributes that are indeed correlated with aging, compared to glasses that can be worn at any age. Moreover, second to fourth rows represent channels that do not have any semantic change in the images but have a significant effect on the classifier scores, similar to adversarial attacks. Considering these, it can be concluded that a minimal human supervision is still needed to evaluate and interpret the identified channels. In the future, such redundant channels can be filtered using cognitive metrics and the bias identification process can be improved.

4.0.3. Manipulation Success

We also evaluate the effectiveness of the manipulations done with the identified style channels. Our bias identification module associates the biased glass attribute with the 120th style channel in the 3rd layer. Observing the manipulation strengths in the Figure 3.2, we concluded that a manipulation strength of 25 can be used to manipulate the images since 20 was not sufficient to manipulate some images. We manipulated 10000 images to the negative direction of the glass attribute using this style channel, that is, we tried to remove the glasses if the person wears one and we tried to add a glasses to the person if they do not wear one. We then check if the manipulations change the class of the person accordingly, classes being the wearing glasses and not wearing glasses. For this evaluation, we use the CelebA attribute classifier for glasses provided by [3]. Our manipulations are able to remove existing glasses on the 61.5% of the samples with glasses and able to add glasses successfully 71.5% of the time. Moreover, these manipulations result in disentangled changes and only modify the glasses area without changing the identity and other attributes of the person, as can

Figure 4.1: Manipulations carried out using the 120th style channel of 3rd layer.

be seen from the Figure 4.1. These results show that the style space, with its capability to edit images in disentangled way is the ideal space to carry out manipulations.

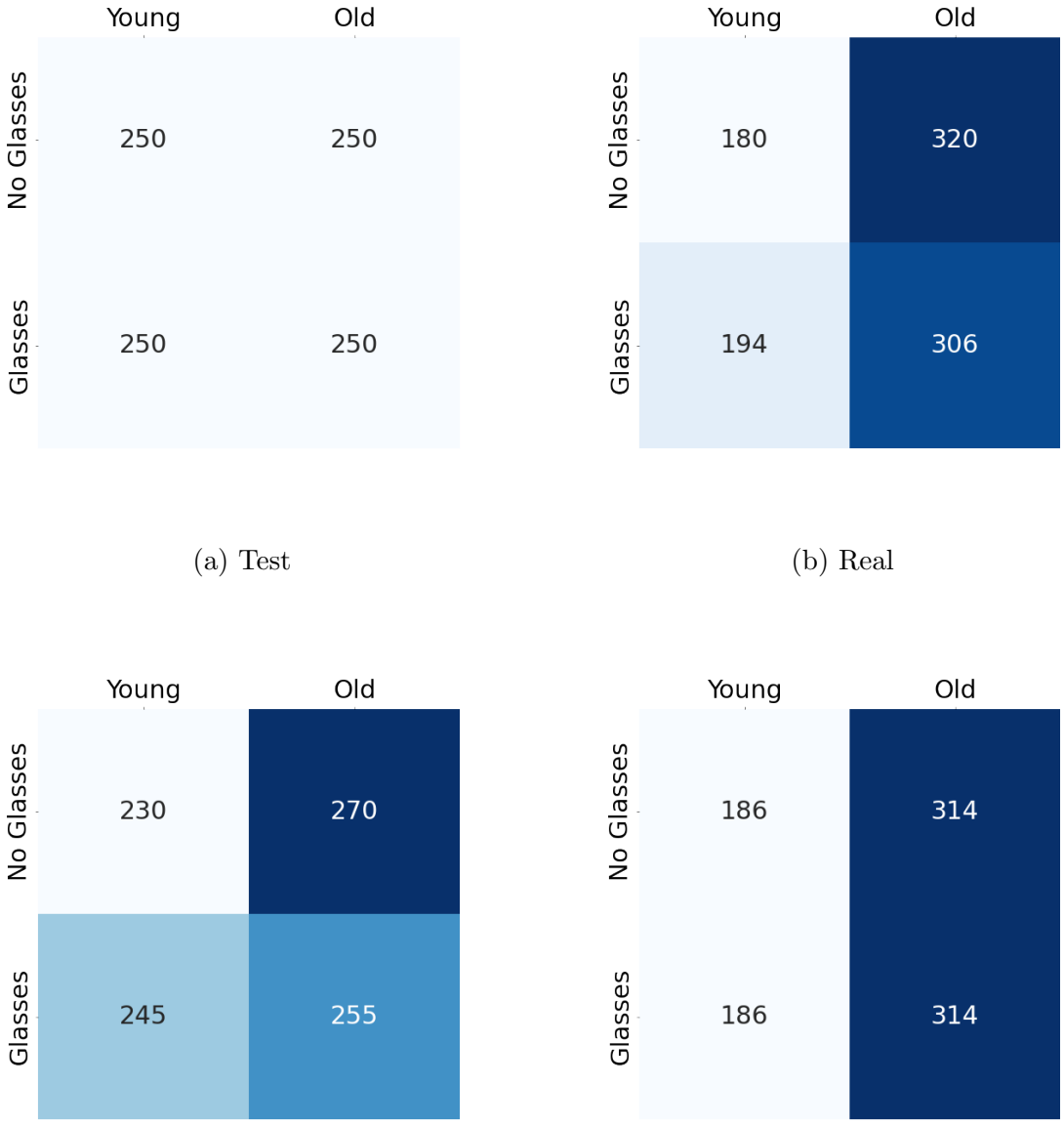
Note that the proportion of successful manipulations can be improved by increasing the manipulation strength; however, doing so would also decrease the image quality as high strengths can cause artifacts in images. This trade-off needs to be investigated further to achieve better results. Furthermore, using a single style channel to manipulate a semantic attribute might be limited since several channels might be related to the same attribute. For example, both the 120th and 228th channels of the 3rd layer are responsible for adding and removing glasses in images, therefore, future work focused on using both channels to manipulate images might result in manipulations that are more successful both quantitatively and qualitatively.

4.0.4. Distribution of Target and Biased Attribute

We evaluate the performances of the real, fair and ensemble classifiers on the basis of the distributions of their predictions. We display the distribution of glasses and age attributes on the test set, on the predictions of the real model, the fair model and the ensemble model in Figure 4.2. We considered a subset of the test set where 1000 samples are equally distributed into young/old and glasses/no glasses categories. As can be seen on the figure, our real model, trained directly on the training set, predicts a less balanced distribution by associating the glasses with older age. Fair model, on the other hand, predicts a much more balanced distribution with respect to the age attribute, although slight bias still exists given the unbalanced distribution. Ensemble method is not as successful as the fair model; however, it can predict the same number of old people whether or not they wear glasses.

4.0.5. Quantitative Analysis

For the quantitative analysis, three metrics are used, which were also used by [20]. The first metric is *difference in equality of opportunity (DEO)* [29], which is the absolute difference between false-negative rates. The second metric is *divergence between*



(a) Test

(b) Real

(c) Fair

(d) Ensemble

Figure 4.2: Distribution of the target and biased attribute among the test dataset, predictions of the classifier trained with real images, the predictions of the debiased classifier, the predictions of the ensemble method.

Model	AP ↑	Correlation ↓	DEO ↓	KL ↓
Real	89.49 ±0.51	**-0.210**	23.88 ±4.72	0.31 ±0.62
Fair	**92.70 ±0.45**	-0.228	**21.60±4.41**	**0.26±0.53**
Ensemble	89.54 ±0.53	-0.212	23.24 ±4.32	0.30 ±0.60

Comparison of the real and fair model with respect to average precision, correlation, difference in equality of opportunity, KL divergence metrics.

score distributions (KL) [30], which measures how much one probability distribution differs from another distribution. Another metric is average precision (AP), which measures the accuracy of the classifiers. AP is a threshold-invariant accuracy metric that summarizes the precision and recall curves. Finally, we check the Pearson correlation between the glasses attribute and young attribute for the predictions of the models. All models yield negative results because these attributes are negatively correlated. The evaluations for these metrics can be seen in Table 4.1. Our fair retraining method and ensemble methods outperform the real model on most metrics, demonstrating their ability and impact in reducing bias. Between the two methods, fair retraining method performs better than ensemble. Even though ensemble method does not produce better results, it also does not require per bias training therefore it is faster.

5. CONCLUSION AND DISCUSSION

In this work, we propose two debiasing approaches that aim to both identify the biases of classifiers by uncovering the agents in the decision making process and eliminating these biases. While one method focuses on retraining a classifier with newly synthesized dataset, other method performs test time manipulations and makes an ensemble prediction. We perform several quantitative and qualitative experiments and show that both our methods are able to reduce biases. We saw that fair retraining method performs better on multiple experiments. However it requires training a classifier per attribute from scratch which may take several hours. Ensemble method does not require training and can be used for applications that do not have access to training resources and data. This trade-off should be taken into account before making a decision on which method to use.

6. FUTURE WORK

In the retraining method, we augmented the data with only the manipulations of a single style channel. This can be extended into multiple style channels to increase the variations; however, doing so would also cause the model to encounter similar images many times and might hurt the performance. In the ensemble method, we considered the effects of multiple style channels. However, in both methods, we did not consider the effects of multiple style channels at the same time by considering the combinations of positive and negative manipulations for each channel. This might improve the accuracy at the cost of a complexity of $\Theta(2^n)$, where n is the number of style channels. Finally, we can improve the ensemble method by changing the ensembling strategy, which is averaging. We might consider implementing a weighted average ensemble and change the weights dynamically by following a reinforcement learning approach.

REFERENCES

1. Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and improving the image quality of stylegan", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

2. Rudd, E. M., M. Günther and T. E. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes", *European Conference on Computer Vision*, pp. 19–35, Springer, 2016.

3. Karras, T., S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks", *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

4. Buolamwini, J. and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification", *Conference on fairness, accountability and transparency*, pp. 77–91, PMLR, 2018.

5. Chen, J., N. Kallus, X. Mao, G. Svacha and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved", *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

6. Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets", Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (Editors), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, Curran Associates, Inc., 2014, `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

7. Karras, T., S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", *CoRR*, Vol. abs/1812.04948, 2018, `http://arxiv.org/abs/1812.04948`.

8. Goetschalckx, L., A. Andonian, A. Oliva and P. Isola, "Ganalyze: Toward visual definitions of cognitive image properties", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.

9. Brock, A., J. Donahue and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis", *arXiv preprint arXiv:1809.11096*, 2018.

10. Shen, Y., C. Yang, X. Tang and B. Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

11. Noble, W. S., "What is a support vector machine?", *Nature biotechnology*, Vol. 24, No. 12, pp. 1565–1567, 2006.

12. Wu, Z., D. Lischinski and E. Shechtman, "StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation", *arXiv preprint arXiv:2011.12799*, 2020.

13. Jahanian, A., L. Chai and P. Isola, "On the" steerability" of generative adversarial networks", *arXiv preprint arXiv:1907.07171*, 2019.

14. Härkönen, E., A. Hertzmann, J. Lehtinen and S. Paris, "Ganspace: Discovering interpretable gan controls", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9841–9850, 2020.

15. Shen, Y. and B. Zhou, "Closed-form factorization of latent semantics in gans", *arXiv preprint arXiv:2007.06600*, 2020.

16. Voynov, A. and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space", *International Conference on Machine Learning*, pp. 9786–9796, PMLR, 2020.

17. Yüksel, O. K., E. Simsar, E. G. Er and P. Yanardag, "LatentCLR: A Contrastive Learning Approach for Unsupervised Discovery of Interpretable Directions", *arXiv*

preprint arXiv:2104.00820, 2021.

18. Lang, O., Y. Gandelsman, M. Yarom, Y. Wald, G. Elidan, A. Hassidim, W. T. Freeman, P. Isola, A. Globerson, M. Irani *et al.*, "Explaining in Style: Training a GAN to explain a classifier in StyleSpace", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 693–702, 2021.

19. Li, Z. and C. Xu, "Discover the Unknown Biased Attribute of an Image Classifier", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14970–14979, 2021.

20. Ramaswamy, V. V., S. S. Kim and O. Russakovsky, "Fair attribute classification through latent space de-biasing", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9301–9310, 2021.

21. Nam, J., H. Cha, S. Ahn, J. Lee and J. Shin, "Learning from failure: De-biasing classifier from biased classifier", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 20673–20684, 2020.

22. Breiman, L., "Random forests", *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

23. Chen, T. and C. Guestrin, "Xgboost: A scalable tree boosting system", *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

24. Chai, L., J.-Y. Zhu, E. Shechtman, P. Isola and R. Zhang, "Ensembling with Deep Generative Views.", *CVPR*, 2021.

25. Sandler, M., A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

26. Tov, O., Y. Alaluf, Y. Nitzan, O. Patashnik and D. Cohen-Or, "Designing an

encoder for stylegan image manipulation", *ACM Transactions on Graphics (TOG)*, Vol. 40, No. 4, pp. 1–14, 2021.

27. Liu, Z., P. Luo, X. Wang and X. Tang, "Deep Learning Face Attributes in the Wild", *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

28. Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

29. Lokhande, V. S., A. K. Akash, S. N. Ravi and V. Singh, "Fairalm: Augmented lagrangian method for training fair models with little regret", *European Conference on Computer Vision*, pp. 365–381, Springer, 2020.

30. Chen, M. and M. Wu, "Towards threshold invariant fair classification", *Conference on Uncertainty in Artificial Intelligence*, pp. 560–569, PMLR, 2020.

APPENDIX A: DATA AVAILABILITY STATEMENT

- The datasets analysed during the current study are available in the FFHQ repository (https://github.com/NVlabs/ffhq-dataset) and CelebA repository (https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html).