ELSEVIER

Image and Visual Computing
journal homepage: https://www.sciencedirect.com/journal/image-and-vision-computing

A Novel Approach for Bias Mitigation of Gender Classification Algorithms using Consistency Regularization

Anoop Krishnan[a], Ajita Rattani[a,b,**]

[a]School Of Computing, Wichita State University, Kansas, USA
[b]Dept. of Computer Science and Engineering, Univ. of North Texas-Denton, Texas, USA

ABSTRACT

Published research has confirmed the bias of automated face-based gender classification algorithms across gender-racial groups. Specifically, unequal accuracy rates were obtained for women and dark--skinned people for face-based automated gender classification algorithms. To mitigate the bias of gender classification and other facial-analysis-based algorithms in general, the vision community has proposed several techniques. However, most of the existing bias mitigation techniques suffer from a lack of generalizability, need a demographically-annotated training set, are application-specific, and often offer a trade-off between fairness and classification accuracy. This means that fairness is often obtained at the cost of a reduction in the classification accuracy of the best-performing demographic sub-group. In this paper, we propose a novel bias mitigation technique that leverages the power of semantic preserving augmentations at the image- and feature-level in a self-consistency setting for the downstream gender classification task. Thorough experimental validation on gender-annotated facial image datasets confirms the efficacy of our bias mitigation technique in improving overall gender classification accuracy as well as reducing bias across all gender-racial groups over state-of-the-art bias mitigation techniques. Specifically, our proposed technique obtained a reduction in the bias by an average of 30% over existing bias mitigation techniques as well as an improvement in the overall classification accuracy of about 5% over the baseline gender classifier. Therefore, resulting in state-of-the-art generalization performance in the intra- and cross-dataset evaluations. Additionally, our proposed technique operates in the absence of demographic labels and is application agnostic, compared to most of the existing bias mitigation techniques.

Keywords:
Fairness in AI, Facial Analytics, Consistency Regularization, Gender Classification, Deep Learning

1. Introduction

As Artificial Intelligence (AI) systems are increasingly used for decision-making in high-stakes scenarios (Yag and Altan, 2022; Ozcelik and Altan, 2023; Kaur et al., 2023), it is vital that they do not exhibit discrimination. However, recent research has raised several fairness concerns, with researchers finding significant accuracy disparities (bias) across demographic groups. Fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics. Thus, an unfair (biased) algorithm is one whose decisions are skewed toward a particular group of people. The facial analysis-based algorithms are at the center stage of this discussion (Krishnan et al., 2020; Albiero et al., 2020). The automated facial analysis-based algorithms include face detection, face recognition, and visual attribute classification (such as gender-, ethnicity-, age-classification, and BMI prediction), and other applications such as deepfake detection (Siddiqui et al., 2022; Nadimpalli and Rattani, 2022; Levi and Hassner, 2015; Almadan et al., 2020; Zhang et al., 2017; Masood

**Corresponding author
e-mail: axupendrannair@shockers.wichita.edu (Anoop Krishnan), ajita.rattani@wichita.edu (Ajita Rattani)

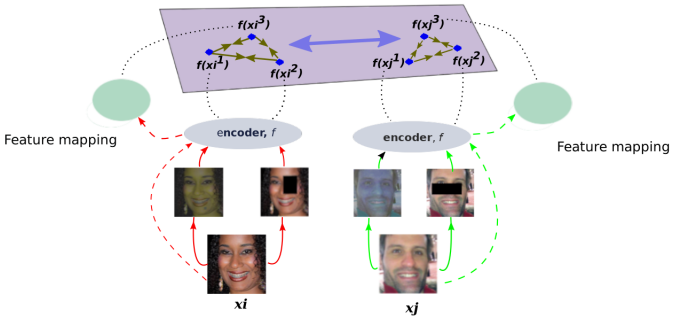

Fig. 1: Illustration of our proposed consistency regularization technique. This figure depicts the proposed method of enforcing consistency using augmented views generated using image-level perturbations and a feature mapping module. The feature mapping module maps the feature vector of the input image to the feature vector of one of its multi-views controlled by pitch, yaw, roll, and field-of-view (fov). This is all possible because these features contain similar semantic information.

et al., 2018; Salim et al., 2021; Kiruthika and Masilamani, 2021). Most of the existing studies related to examining the fairness of facial analysis-based algorithms suggested performance differences for people of color and females (Grother et al., 2011; Klare et al., 2012; Best-Rowden and Jain, 2018; Abdurrahim et al., 2018; Raji and Buolamwini, 2019; Vera-Rodríguez et al., 2019; Albiero et al., 2020; Krishnan et al., 2020; Muthukumar, 2019). These facial-analysis based algorithms are deployed for applications ranging from surveillance, border-control, identity recognition, media authenticity, and consumer products. Thus, bias in these systems is a significant *social problem* that needs immediate attention for the large-scale deployment of fair and trustworthy facial-analysis-based algorithms across demographics.

Among various facial image-based visual attributes such as gender, ethnicity, and age, gender is an essential demographic attribute (Ricanek and Tesafaye, 2006; Levi and Hassner, 2015). Automated gender classification has sparked a lot of interest in a variety of applications including image retrieval, surveillance, and human-computer interaction. Further, gender has been regarded as a soft biometric attribute that has been fused with primary biometric modalities such as face and ocular, for improving their recognition accuracy. An automated face-based gender classifier has been made available in commercial SDKs from tech giants such as Amazon Rekognition (Rekognition, 2022), DeepVision AI (Vision, 2022), FaceX (FaceX, 2022), and Microsoft Azure Cognitive Services (Services, 2022).

Over the last few years, published research has questioned the fairness of these face-based automated gender classification algorithms across gender and ethnicity (Buolamwini and Gebru, 2018; Muthukumar, 2019; Kärkkäinen and Joo, 2019). Specifically, existing gender classification studies obtain unequal accuracy rates for women and dark-skinned people such as African-Americans. Since the majority of the research on this subject adheres to the idea that "gender" is binary, we also stick to it for the sake of fair comparison. We do not intend to belittle people who disagree with this view in the course of conducting this study. Also, in this study, the gender labels were obtained from the gender-annotated publicly available facial datasets.

To mitigate the bias of gender classification and other facial analysis-based algorithms, the vision community has developed several solutions. These solutions are based on regularization strategies (Kamishima et al., 2012), attention mechanism (Majumdar et al., 2021), adversarial debiasing (Zhang et al., 2018; Chuang and Mroueh, 2021), over-sampling the minority class using Generative Adversarial Networks (GANs) (Ramaswamy et al., 2021), multi-task classification methods (Das et al., 2018) and network pruning (Lin et al., 2022).

However, most of these bias mitigation techniques need demographically annotated training sets offer poor generalizability, and are computationally expensive. Further, the fairness techniques based on disentangling the features related to protected attributes from those related to the main classification task (Majumdar et al., 2021; Das et al., 2018) may result in the removal of the features important for the main classification task. Further, most of the proposed debiasing techniques are tailored-made for specific use cases and these techniques require the presence of demographic attributes (Zhang et al., 2018; Ramaswamy et al., 2021; Lin et al., 2022; Majumdar et al., 2021; Das et al., 2018). Also, often the aforementioned mitigation strategies offer a trade-off (Zhang et al., 2018) between fairness and classification accuracy. In other words, overall performance is decreased so as to improve the performance of disadvantaged groups where the algorithm would otherwise be less accurate. A recent study also suggests that existing bias mitigation approaches for facial analysis-based algorithms may improve fairness by degrading the performance of the classifier across all sub-groups (with increased degradation to the best-performing sub-groups). This is called as *pareto-inefficiency* (Zietlow et al., 2022). *An application-agnostic bias mitigation strategy that enhances fairness as well as performance across all inter-sectional subgroups (including the best-performing sub-group), in the absence of demographic labels is still an open issue*.

We **conjecture** that one way to improve fairness without decreasing the performance across sub-groups is by improving the feature representation for each sub-group. It is well-known, and experimentally verified (Li and Vasconcelos, 2019) that the enhanced feature representation improves the generalizability of the classifier and decreases the variance of the worst-off group(s) and hence the bias. Based on this notion, this paper proposes a bias mitigation solution for a gender classifier that leverages the power of augmented views at the image- and feature level using a consistency-based regularization setting during the training stage for enhanced feature representation for each sub-group. This results in the reduction of the variance of the worst-off subgroup(s), and hence the bias. With respect to **significance** over existing bias mitigation techniques, our proposed bias mitigation strategy has the *dual advantage* of enhancing the classifier's performance as well as fairness across all inter-sectional subgroups, as demonstrated through extensive experimental evaluations. Further, our proposed bias mitigation approach can be applied in the absence of demographic labels to any downstream vision-based classification task, not

just gender classification, exhibits faster convergence, and obtains high generalizability in the cross-dataset scenario. Overall, the proposed approach presents a promising direction in advancing representation learning techniques, offering improved performance as well as fairness. Figure 1 depicts our proposed approach to enforcing consistency using augmented views at the image level using image augmentations and at the feature level using a trained feature mapping module, to transform the feature vector of the original sample to the feature vector of one of its multi-views by varying pitch, roll, yaw, and field-of-view using a pre-trained *StyleNeRF* model (Gu et al., 2022). Because these feature vectors contained similar semantic information, which in our case is *gender information*, enforcing consistency regularization was possible which improved the feature representation of each subgroup. The encoder in a convolutional neural network is composed of convolutional layers with an activation function after each layer.

To this aim, we evaluate and address the following research questions **(RQs)** as follows:

- **RQ1: "How effective are augmented views of the training samples in mitigating bias?"** In other words, could augmented views applied at the image or feature level obtain a considerable reduction in bias by reducing the variance of the sub-groups through enhanced feature representation?
- **RQ2: "Could the proposed technique obtain fairness as well as enhanced performance across several intersectional subgroups (including the best-performing sub-group)?"**
- **RQ3: "Can the combination of image-level and feature-level augmentations obtain any advantage over either of them individually?"**

1.1. Our Contribution

In summary, the main contributions of this work are as follows:

- A bias mitigation strategy based on augmented views of training samples at the image level in a consistency-based regularization setting. Image augmentations such as colorjitter, enhancing contrast and brightness, random erasing (Zhong et al., 2020), and random augmentation (Cubuk et al., 2020) were systematically chosen based on the ablation study.
- Apart from image-level augmentations, feature-level augmentations were used and obtained from a feature transform model powered by the neural rendering technique called *StyleNeRF* (Gu et al., 2022). It combines the neural radiance fields algorithm with style transfer techniques to create photo-realistic images and consistent 3D-aware multi-views.
- The use of the combination of both image-level and feature-level semantic preserving augmentations for generating diverse views of the original training samples in a consistency-based regularization setting for learning enhanced and robust feature representation.
- Extensive evaluation of the proposed bias mitigation strategy in the intra- and cross-dataset evaluation scenario. To this front, FairFace (Karkkainen and Joo, 2021) for training, and FairFace, DiveFace (Morales et al., 2021), and UTKFace (Zhang et al., 2017) test sets were used for intra- and cross-dataset evaluations.
- Cross-comparison with the published studies on bias mitigation techniques proposed for gender classification tasks and evaluated on the same datasets.

This paper is organized as follows: The relevant related work on examining and mitigating bias on face-based gender classifiers is discussed in Section 2. Section 3 discusses the proposed bias mitigation strategy. Section 4 discusses the datasets used, implementation details, and the evaluation metrics. Results are discussed in Sections 5, and further analysis and discussion of the obtained results are discussed in Section 6. ANOVA-based hypothesis testing for the statistical significance of the obtained results is detailed in Section 7. Section 8 discusses the conducted ablation study, and the key findings are reported in Section 9. Finally, the concluding remarks are discussed in Section 10.

2. Prior Work

In this section, we will discuss the related work on investigating and mitigating bias in facial analysis based on gender classification algorithms.

On Examining Bias: The following foundational work has identified the systematic failings of gender classification algorithms on particular racial and gender demographics (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019; Muthukumar, 2019; Balakrishnan et al., 2020; Krishnan et al., 2020).

Specifically, Buolamwini and Gebru (Buolamwini and Gebru, 2018) and Raji and Buolamwini (Raji and Buolamwini, 2019) evaluated the fairness of five COTS gender classifiers and suggested unequal accuracy for dark-skinned people and women on the Pilot Parliaments Benchmark (PPB) dataset. Muthukumar (Muthukumar, 2019) and Balakrishnan et al. (Balakrishnan et al., 2020) suggested that age, hair length, and facial hair be the likely cause of the performance differential for women and dark-skinned people when evaluated on PPB dataset. Krishnan et al. (Krishnan et al., 2020) evaluated the efficacy of different CNN architectures (ResNet-50, Inception-V4, VGG-16/19, and VGGFace) in gender classification across gender-racial groups when evaluated on the UTKFace and FairFace datasets, respectively. The authors suggested that architectural differences impact unequal accuracy rates. The high morphological similarity between black males and females is attributed to the high misclassification error rate of the latter.

On Mitigating Bias: Multiple approaches have been proposed to mitigate the bias of gender classification algorithms. Teru and Chakraborty (Teru and Chakraborty, 2019) proposed an adversarially learned encoder for obtaining ethnicity invariant

representation for gender classification when evaluated on the UTKFace dataset. Das et al. (Das et al., 2018) proposed a Multi-Task Convolution Neural Network (MTCNN) to jointly classify gender, age, and ethnicity, as well as to minimize the impact of protected attributes. The proposed model was evaluated on UTKFace and BEFA datasets. Majumdar et al. (Majumdar et al., 2021) proposed the attention-aware debiasing method that uses an attention module focusing on the features important for the main classification task while suppressing the features related to the sensitive attributes. The experimental evaluation was performed on UTKFace and Morph datasets.

Ramachandran and Rattani (Ramachandran and Rattani, 2022) and Ramaswamy et al. (Ramaswamy et al., 2021) proposed methods based on generative views obtained using GAN-based latent vector editing along with structured learning for mitigating bias on gender classification. Lin et al. (Lin et al., 2022) proposed a neural-network pruning technique that computes the per-group importance of each model weight. It then iteratively selects and prunes those weights with small importance values to reduce performance disparity. The effectiveness of the proposed method is demonstrated on FairFace, UTKFace, and CelebA datasets. Park et al. (Park et al., 2022) proposed Fair Supervised Contrastive Loss (FSCL) that enforces the representations of the same class to be closer to each other than that of different classes for fair representation learning. The effectiveness of the proposed method was demonstrated on CelebA and UTKFace datasets.

3. Proposed Approach

In this section, we will discuss the consistency-regularization-based technique used to reduce the bias in the facial image-based gender classification algorithm.

3.1. Consistency Regularization Basics

The main aim of consistency regularization is to train a model that is invariant to various data perturbations (Tan et al., 2022; Englesson and Azizpour, 2021; Saunshi et al., 2022). This is done by enforcing the model to obtain consistent predictions for the training sample and its perturbed instances generated using semantics-preserving augmentations. As the consistency regularized model can learn a good feature representation invariant to various perturbations, this method is widely used in semi-supervised and self-supervised learning techniques for harnessing unlabeled data (Nadimpalli et al., 2021; Assran et al., 2023; Huang et al., 2023). Mathematically, consistency regularization can be expressed as:

$$L_c(x) = \| f(x) - f(T(x)) \|^2 \tag{1}$$

where $L_c(x)$ is the consistency loss (which is usually mean square error), x is the original input data, $f(x)$ is the feature representation of x, and $T(x)$ is a randomly perturbed version of x. The consistency loss measures the distance between the feature representation of x and $T(x)$. By minimizing the consistency loss, the model is encouraged to produce consistent feature representations for both x and $T(x)$. Next, we will discuss how we incorporated consistency regularization for learning better feature representation using both image- and feature-level augmentations.

Let I denote the image sample used for training the model, and f be the classification model, then $f_{cls}(I)$ denote the output class probability vector after softmax operation and $f_{feat}(I)$ as the feature vector from the f. Let $\hat{I}$ be the augmented view of I generated using an augmentation. The aim of consistency regularization is to minimize the distance between the predictions from these two images, i.e., $(f_{cls}(I), f_{cls}(\hat{I}))$ and the distance between the feature embeddings $(f_{feat}(I), f_{feat}(\hat{I}))$.

Similarly, at the feature level, the feature embedding of the original image sample I was transformed into the feature embedding of one of its multi-views using a learned mapping function. The new feature embedding obtained was denoted as $f_{\hat{feat}}(I)$. As these feature embeddings carry semantically similar information, the distance between the feature embeddings of I i.e., $f_{feat}(I)$ and $f_{\hat{feat}}(I)$ should be minimum. Therefore, the aim of consistency regularization is to minimize the distance between these feature embeddings.

Accordingly, for image-level perturbation, the consistency loss is computed as:

$$Loss_{imgcons} = distance(f_{cls}(I), f_{cls}(\hat{I})) + distance(f_{feat}(I), f_{feat}(\hat{I})) \tag{2}$$

At feature-level perturbation, the loss is computed as:

$$Loss_{featcons} = distance(f_{feat}(I), f_{\hat{feat}}(I)) \tag{3}$$

The **overall loss for the classification task** is as:

$$TotalLoss = l_{clfloss} + \lambda \times l_{consistency} \tag{4}$$

where $l_{clfloss}$ be the classification loss, $l_{consistency}$ be the regularization loss, (when image-level perturbation was considered, $l_{consistency}$ = Equation 2, and for feature-level perturbation, $l_{consistency}$ = Equation 3 and λ is the weightage between the two loss functions.)

3.2. Consistency Regularization for Bias Mitigation using Augmented Views of Facial Images

In this section, we will discuss the consistency-regularization enforced through image level *augmentations* during the training stage.

Following the existing studies (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019; Muthukumar, 2019; Balakrishnan et al., 2020) on unequal accuracy rates for facial analysis based gender classification algorithms across demographic variations. The unequal accuracy rates are attributed to the under-representation of those specific sub-groups even when the training set is balanced. The under-representation could be due to poor quality training data attributed to environmental conditions such as low lighting conditions, acquisition set-ups such as sensor quality, skin-tone variations, and other factors such as make-up and occlusions. Worth-mentioning in our previous study (Krishnan et al., 2020), we observed that the high similarity in facial morphology between African-American females and males contributes to performance differential of gender classification algorithms across gender-ethnicity groups.

Therefore, apart from the aforementioned factors, facial morphological variations across gender and ethnicity could also contribute to unequal accuracy rates.

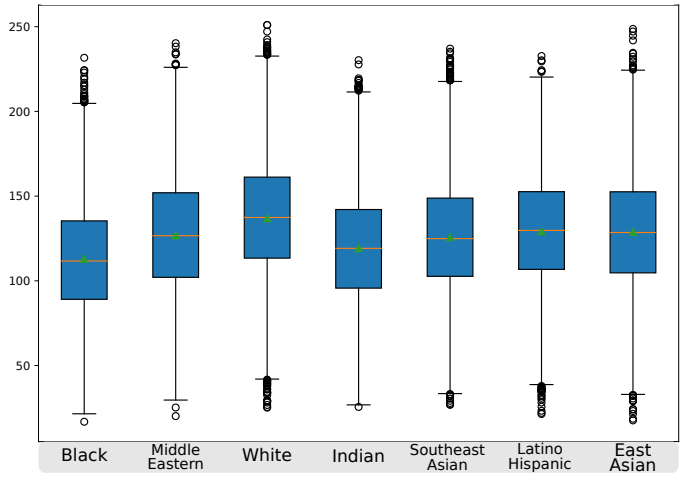

Fig. 2: Box plots of the intensity values of brightness for facial images of females from the FairFace training set. It can be seen that the images of Black females have comparatively low brightness. The average luminance of the image was perceived as its brightness.

To account for these variations in the training data across demographics (across gender and ethnicity), an instance of each training facial image was perturbed to generate augmented views with altered hue, saturation, brightness, contrast, and occlusion generated using randomly erased patches.

These augmentations were selected after a thorough investigation of various semantics-preserving image augmentations (Cubuk et al., 2020). As per our observation from Figure 2, the variation in brightness was significant across facial images from different demographic subgroups such as Black females and males, and also similar observation was noticed for contrast value as shown in Figure 3. Therefore, the augmentations with varying brightness and contrast could enhance the robustness of the model (classifier) to these variations and hence reduce the bias across gender and ethnicity. Further, changing

Fig. 3: Example sample images with poor contrast from the FairFace testing set. The contrast further degrades for darker skin-tone subjects.

the hue, and saturation of the image results in a different skin color effect, Thus augmentations with varying hue and saturation could enhance the robustness of the model to varying skin tones across ethnicities. Occlusion due to random erasing could help the model learn other cues in the facial image apart from facial morphology for the downstream classification task. Thus, all the aforementioned augmentations are intended to render the model invariant to variation in contrast, brightness, skin tone, and facial morphology across demographic sub-groups.

Figure 4 shows examples of image samples obtained after applying color-jitter and random erasing-based augmentations. These augmentations were applied to training images. The original image along with the augmentations was used during the training stage of the model to enforce consistency regularization for the gender classifier using a combination of the loss functions as given in equation 4. No perturbations were applied to the test samples during the evaluation stage.

Fig. 4: Example samples from the FairFace training set along with their augmented views obtained after applying ColorJiT Transform and Random Erase operations

.[Best viewed in color.]

3.3. Consistency Regularization for Bias Mitigation using Perturbations on the Feature Map

In computer vision and computer graphics, photo-realistic free-view image synthesis of real-world scenes has long been a challenge. To create high-quality images, generative models can be trained on a huge number of unstructured images but the majority of GAN models work in a 2D space. As a result, GANs are unable to synthesize images of the same 3D scene with multi-view consistency due to their lack of 3D knowledge of the training images. To tackle this lack of multiview knowledge, a recent technique called *StyleNeRF* (Gu et al., 2022) was introduced for neural rendering that combines two powerful deep learning models, NeRF (Mildenhall et al., 2022) and StyleGAN (Karras et al., 2019), to generate photorealistic images of 3D scenes with rich textures and lighting effects.

NeRF (Neural Radiance Fields) (Mildenhall et al., 2022) is a method for modeling 3D scenes as a continuous function using a deep neural network. The network maps a 3D point and a camera viewpoint to the color and opacity of the point. By integrating the volume of the scene, NeRF can render novel views of the scene from any viewpoint with high quality and detail. Meanwhile, StyleGAN (Karras et al., 2019) is a generative model that can synthesize high-quality images of diverse objects and scenes by learning to model the distribution of images in a low-dimensional latent space. StyleGAN uses a progressive growing architecture and a style-based generator network to control the appearance of the generated images in terms of features such as color, texture, and shape. In *StyleNeRF*, the

idea is to use StyleGAN to control the appearance of the rendered 3D scene by conditioning the NeRF network on a latent code that is generated by StyleGAN. Specifically, *StyleNeRF* uses a hybrid network that combines a modified NeRF network with a StyleGAN generator network. The NeRF network takes as input a 3D point in space and a camera viewpoint and outputs the color and opacity of the point. The StyleGAN generator takes as input a latent code and outputs the style vectors that control the appearance of the rendered scene.

Mathematically, *StyleNeRF* can be expressed as:

$$I(\mathbf{x}) = \sum_{i=1}^{N} \sigma_i(\mathbf{x}) \cdot T_i(\mathbf{x}) \cdot C_i$$

where $I(\mathbf{x})$ is the color of a pixel at 3D point $\mathbf{x}$ in the rendered image, N is the number of views used to capture the scene, $\sigma_i(\mathbf{x})$ is the density of view i at 3D point $\mathbf{x}$, $T_i(\mathbf{x})$ is the transmittance of view i at 3D point $\mathbf{x}$, and C_i is the style of view i. The density and transmittance are computed using NeRF, which learns a function that maps a 3D point to a density and transmittance value based on the training images. The style C_i is obtained by applying style transfer to the training images of view i. *StyleNeRF* trains a network that takes as input a 3D point $\mathbf{x}$ and a view index i, and outputs the color $I(\mathbf{x})$ of the pixel in the image rendered from view i. The network is trained using a combination of supervised and unsupervised losses to ensure that the rendered images match the training images in terms of content and style.

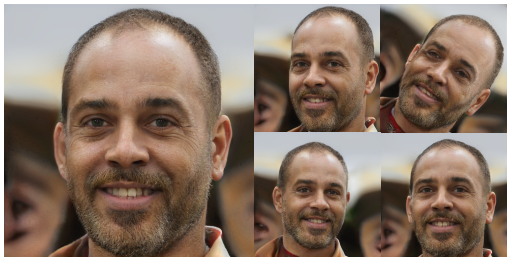

Fig. 5: High-resolution sample and its multi-views by varying explicit camera controls namely pitch, roll, yaw, and field of view on FFHQ dataset (Karras et al., 2021) of size 512×512 generated using *StyleNeRF* model.

Training the Feature Transform MLP: We aimed to improve the feature representation of the image sample at feature space. One way to achieve it with a given dataset was by achieving random transformations or augmentations on the image level, which we discussed before. Another way is to generate different views of each sample, which comes with a lot of challenges, like computational complexity, and mode collapse in GANs, and in *StyleNeRF*, high computational requirements and sensitivity to changes in the input data, which can result in artifacts or inconsistencies in the output images. This can make it difficult to generate high-quality images that are consistent with the desired style or appearance.

To overcome these limitations and to leverage the effectiveness of the multi-view consistency of *StyleNeRF*, we trained a Multi-layer Perceptron (MLP), which we call a feature transform model to generate the difference in the feature embeddings from the multi-views generated by varying yaw, pitch, roll, and fov. We used the *StyleNeRF* model pre-trained on FFHQ 512^2 dataset (Karras et al., 2021) in this study. To achieve it, we extracted the feature of the *StyleNeRF* generated image (Img 1 in the Figure 6) along with these four variables, namely yaw, pitch, roll, and fov. The value of yaw and pitch ranges between -2.5 to 2.5, whereas roll ranges between -1 to 1 and fov ranges between 9 to 15. The values of these four parameters were chosen randomly while generating images and then fed to the feature transform model along with the aforementioned feature vector of the generated image as shown in Figure 6. Further, the feature of its *StyleNeRF* generated multi-view (Img 2 in Figure 6) was also extracted. The features were generated using pre-trained gender classifiers (Baseline models) (denoted as E in Figure 6) as shown in Figure 6. Thus, the feature vectors (obtained from the pre-trained gender classifier) of the generated image and its multi-view, obtained from the *StyleNeRF*, were used to train a feature transform MLP.

Let d be the size of the feature of the generated image ($feat$), along with the four variables (yaw, pitch, roll, and field of view (fov)) and d as the size of the feature of the newly generated image ($\hat{feat}$) with the given viewing angles applied.

Then $feat$ along with yaw, pitch, roll, and fov is input to the MLP, which makes the size of input as $d + 4$, and its generated multi-view feature vector ($\hat{feat}$) of size d as the target. Now it became a supervised learning problem where we could train the feature transform model (MLP), which can approximate this feature mapping from $feat$ and $\hat{feat}$.

To train the feature transform model (MLP) (F in Figure 6), we generated $500,000$ pair which included the feature vector and its transformed feature vector with a dimension of size 1024. On concatenating the other four variables, namely pitch, roll, fov, and yaw, the size of the input to the feature transform MLP became 1028 and 1024 was the size of the target. We employed an empirically chosen three-layer MLP with GELU as non-linear activation applied after the input and the one hidden layer. The model was trained using a batch size of 512 using RTX8000 GPU. The training was performed using an Adam optimizer using cosine annealing with a warm restart for a learning rate schedule with an initial learning rate of 3×10^{-4} along with mixed precision and early stopping mechanism. The mean square error was used as the loss function.

In this work, we trained the models individually with the image level consistency regularization using image augmentation, and with the feature level consistency regularization using feature transform which was powered by the multi-views generated by the *StyleNeRF* which were controlled by four parameters namely, pitch, roll, yaw, and field of view as shown in the Figure 6. This would give an idea of how the feature representation enhanced when image-level and feature-level transformations were applied individually. Then we combined the power of image-level augmentation and feature-level augmentation by generating image-level transformations and feature-level transformations together to understand whether the combined effect can improve the feature representation further by maximizing the diversity in the data manifold. We also experimented by combining image-level augmentations with *StyleNeRF*-generated image-level multi-views of the original

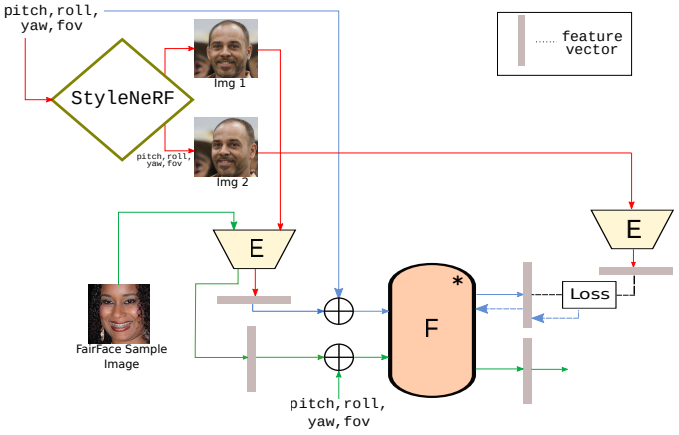

Fig. 6: Training and use case depiction of the Feature Transform framework. E: Encoder; F: Feature Transform MLP. The red line indicates the image flow generated by the *StyleNeRF*. The blue line indicates the training flow of Feature Transform MLP. The green line indicates the use case scenario of the Feature Transform MLP. The red line indicates the data flow for training the Feature Transform MLP. $*$: *The only trainable module.*

image. However, the results were poor due to the difference in the data distribution of *StyleNeRF*-generated multi-view images and those obtained by image-level augmentations.

3.4. Stabilizing the model training process with spectral weight normalization.

One of the disadvantages of the consistency regularization scheme could be the model may be able to generate the same feature distribution for both the original and perturbed data distributions (Arjovsky and Bottou, 2017). Such a model may overfit and lack information about the data distribution. This motivates us to impose some kind of restriction on the model.

Therefore, on enforcing consistency regularization with augmented views, in order to refrain the model from generating the same feature vectors, we used a technique called spectral normalization (Miyato et al., 2018), which is computationally light and easy to incorporate into existing implementation. We incorporated it with the feature extraction and classification layer. Spectral normalization is a technique that can help stabilize the feature generation by enforcing Lipschitz continuity, which limits the rate of change of the function. The spectral normalization (SN) normalizes the spectral norm of the weight matrix A so that it satisfies the Lipschitz constraint $\sigma(\mathrm{A}) = 1$, where $\sigma(A)$ is the spectral norm of the matrix A. The spectral norm of a matrix is the maximum singular value of the matrix, which gives an upper bound on how much the matrix can stretch any input vector. By constraining the spectral norm of the weight matrices, spectral normalization can limit the Lipschitz constant of the classifier, which can make the training process more stable. Spectral normalization also has a regularization effect, which can help prevent overfitting and improve generalization. In summary, spectral normalization can stabilize the training by enforcing Lipschitz continuity, limiting the rate of change of the function, and enforcing regularization to prevent overfitting.

4. Experimental Setup

In this section, we will discuss the datasets used and the details of model training.

4.1. Dataset

We used a gender- and ethnicity-balanced fairface dataset for training the gender classifier (Refer Table 1). The trained gender classifiers (models) are tested on FairFace, UTKFace, and DiveFace datasets. The images in these datasets vary across age, gender, pose, lighting conditions, and expression. These datasets are discussed below as follows:

FairFace: The Fairface dataset (Karkkainen and Joo, 2021) consists of 108,501 images, with an emphasis on balanced ethnicity composition in the dataset. The dataset is labeled with the seven-ethnicity groups, namely White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino Hispanic across male and female and age groups ranging from 0-9, 10-19, 20-29, 30-39, 40-49 and 50+. The training portion of the FairFace dataset consists of 47% females and 53% males. Table 1 shows the training set distribution of this dataset. Figure 7 shows a few sample images from the dataset.

Fig. 7: Sample images from the FairFace dataset (Karkkainen and Joo, 2021).

Table 1: Training Dataset Distribution of FairFace used for training all the models in this study.

Ethnicity	Female	Male	Total
White	7826 (9%)	8701 (10%)	16527 (19%)
Black	6137 (7%)	6096 (7%)	12233 (14%)
East Asian	6141 (7%)	6146 (7%)	12287 (14%)
Indian	5909 (7%)	6410 (7%)	12319 (14%)
Middle Eastern	2847 (3%)	6369 (8%)	9216 (11%)
Latino Hispanic	6715 (8%)	6652 (8%)	13367 (16%)
Southeast Asian	5183 (6%)	5612 (7%)	10795 (13%)
Total	40758 (47%)	45986 (53%)	86744 (100%)

UTKFace: The UTKFace dataset (Zhang et al., 2017) is a facial image dataset with a long age span (ranging from 0 to 116 years old). It contains over 20,000 face images annotated with age, gender, and ethnicity, namely White, Black, Asian, Indian, and Others (which include Hispanic, Latino, and Middle Eastern) with significant variations across pose, expression, illumination, occlusion, and resolution. Due to the vagueness of the "Other" category, we excluded it from this study. Figure 8 shows a few sample images from the UTKFace dataset.

DiveFace: The DiveFace dataset (Morales et al., 2021) is a facial image dataset and contains a total of 139,677 images. It contains gender and ethnicity annotations equally distributed to three ethnic groups (namely East Asian, Sub-Saharan and South

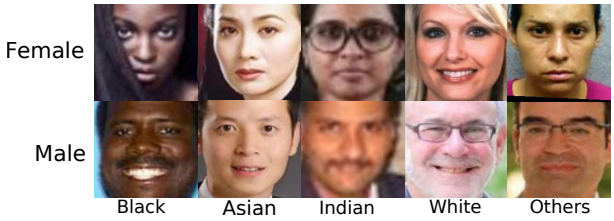

Fig. 8: Sample images from the UTKFace dataset (Zhang et al., 2017).

Indian, and Caucasian). Figure 9 shows a few sample images from the DiveFace dataset.

Fig. 9: Sample images from the DiveFace dataset. (Morales et al., 2021)

4.2. Implementation Details

Model Training: Four different architectures pre-trained on the ImageNet dataset, namely ResNet18 (He et al., 2016), DenseNet121 (Huang et al., 2017), EfficientNetv2 (Tan and Le, 2021), and Vision Transformer (Dosovitskiy et al., 2021) were used for baseline gender classification. Images were input to the Vision Transformer as a sequence of 32×32 fixed-size patches. For the experiments, all images were resized to 224^2.

All the pre-trained weights were obtained from the Timm repository by Ross Wightman (Wightman, 2019) where the weights are already converted from JAX to PyTorch. To all the baseline pretrained architectures, we added a dense layer of size 1024 as a feature extraction layer followed by a final output layer and fine-tuned them on the FairFace training set. Table 2 shows all the abbreviations used in this study.

- **Baseline:** The aforementioned models were fine-tuned on the FairFace training set using the binary cross-entropy loss as given below. We applied Random augmentation (Cubuk et al., 2020) to the training samples.

$$l_{BCE} = -(y \log(p) + (1 - y) \log(1 - p)) \tag{5}$$

where *log* is the natural log, y is the binary label (0 for Female and 1 for Male) and p is the predicted probability. Along with the binary cross-entropy loss, we included a Dirichlet prior to the softmax output to further improve the generalization (Sensoy et al., 2018). Incorporating a Dirichlet prior to the softmax output means that we assume that the class probabilities follow a Dirichlet distribution with a fixed hyperparameter. This prior represents our belief about the distribution of class probabilities before observing any data. When we train the model, we update our prior belief based on the observed data using the maximum likelihood estimation or Bayesian inference. By adding a Dirichlet prior to the softmax output with binary cross-entropy loss, we can encourage the model to learn class probabilities that are close to our prior belief. This can be particularly useful in situations where we have some prior knowledge about the class probabilities or when we want to regularize the model towards a particular distribution of class probabilities.

Table 2: Abbreviations used in this study.

Abbreviation	
CR	Consistency Regularization
CR-Img Aug	Consistency Regularization with Image Augmentation
CR-Feat Trans	Consistency Regularization with Feature Transform
CR-Img Aug+ Feat Trans	Consistency Regularization with Image Augmentation and Feature Transform
CR-Img Aug(SN)	Consistency Regularization with Image Augmentation regularized with Spectral Normalization
CR-Img Aug+ Feat Trans(SN)	Consistency Regularization with Image Augmentation and Feature Transform regularized with Spectral Normalization
MLP	Multi Layer Perceptron
DoB	Degree of Bias
ViT	Vision Transformer
t-SNE	t-distributed stochastic neighbor embedding

The final classification loss as used in this study is given by,

$$l_{clfloss} = l_{BCE} + \lambda \log p(\boldsymbol{p} \mid \boldsymbol{\alpha}) \tag{6}$$

here l_{BCE} is the classification loss (Equation 5), $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ is the vector of class probabilities obtained by applying the softmax function to the model's logits $\boldsymbol{z}$. $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ is the vector of hyperparameters that determine the strength of the Dirichlet prior over the class probabilities. $p(\boldsymbol{p} \mid \boldsymbol{\alpha})$ is the probability density function of the Dirichlet distribution with hyper-parameters $\boldsymbol{\alpha}$, given by:

$$p(\boldsymbol{p} \mid \boldsymbol{\alpha}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \tag{7}$$

where $B(\boldsymbol{\alpha})$ is the normalization constant of the Dirichlet distribution, known as the multivariate beta function. λ is a hyperparameter that controls the strength of the prior and in our case, K is two (Sensoy et al., 2018). In the Tables 3, 4 and 5, the baseline performance of the models is denoted as **Baseline**.

- **Consistency Regularization (CR) using image augmentation and feature transformation:** As mentioned, models were fine-tuned for gender classification with a penultimate layer as a feature extractor of size 1024, followed by the output layer. The models were trained separately with consistency regularization enforced using image augmentations and feature transformations applied individually

(as mentioned in Section 3.2 and Section 3.3, respectively) and in combination. While the training models with consistency regularization enforced using feature level augmentation, we generated only one transformed feature vector. In the Tables 3, 4 and 5, these models are denoted as CR-Img Aug, CR-Feat Trans, and CR-Img Aug+Feat Trans for Consistency Regularization using Image Augmentation, Feature Transformation, and using combined Image Augmentation and Feature Transformation, respectively. The loss functions used for different methods are as follows:

- **CR-Img Aug:** $Eq.2 + Eq.6$
- **CR-Feat Trans:** $Eq.3 + Eq.6$
- **CR-Img Aug+Feat Trans:** $Eq.2 + Eq.3 + Eq.6$

Distance metrics such as $L2$-norm (Euclidean distance) (Grill et al., 2020) or Jensen-Shannon divergence (JSD) (Fuglede and Topsøe, 2004), could be utilized as the consistency regularization loss in equations 2 and 3 . Jeong et al. (Jeong et al., 2019) suggested that consistency regularization techniques perform worse when used with $L2$-based distance metrics as consistency loss. Therefore for our experiments, we used *Jensen-Shannon Divergence* to calculate the distance between the image embedding and its augmented embedding, which essentially captures the information loss between the image and its augmented view.

$$JSD(\hat{y}||y) = \frac{1}{2}(KL(y||\frac{y+\hat{y}}{2}) + KL(\hat{y}||\frac{y+\hat{y}}{2})) \quad (8)$$

where,

$$KL(\hat{y}||y) = \sum_{c=1}^{M} \hat{y}_c \log \frac{\hat{y}_c}{y_c} \quad (9)$$

where KL is the Kullback–Leibler divergence, $\hat{y}$ and y were two feature vectors.

We obtained the best results when equal weightage was given to each loss component (refer λ parameter in equation 4).

- **Spectral normalization along with Consistency Regularization(CR) using image augmentation and feature transformation:** The consistency regularized models were used along with spectral normalization applied on the linear layers to stabilize the training process of the classifier as discussed in Section 3.4. In Tables 3, 4 and 5, these models are denoted as CR-Img Aug(SN) for consistency regularization using image augmentation with spectral normalization on feature and output layers, and CR-Img Aug+Feat Trans(SN) for combining image augmentation and feature transform along with the spectral normalization. The spectral normalization was only applied when image level consistency was used because there could be a chance of overfitting as the feature vectors were generated by the same architecture for the input image and its perturbed instance. The loss functions for CR-Img Aug(SN) and CR-Img Aug+Feat Trans(SN) were the same as those of CR-Img Aug and CR-Img Aug+Feat Trans, respectively.

Other implementation details:

For the model's training, the batch size was set to 128, distributed across 2 RTX 8000 GPUs. Additionally, label smoothing with a value of 0.1 was applied. The RMSprop optimizer was used with a cosine annealing learning rate schedule. A warm restart strategy was employed, starting with an initial learning rate of $1 \times e^{-6}$. A weight decay of $1 \times e^{-5}$ was also incorporated. To enhance the model's performance, stochastic weight averaging, as described in (Izmailov et al., 2018), was utilized along with mixed precision training and an early stopping mechanism. Optimal hyper-parameters for all experiments were determined using grid search, a systematic technique that explores different combinations of hyper-parameter values to identify the best configuration for maximizing performance. All the experiments were implemented using the PyTorch-lightning framework, which provides a convenient and efficient platform for conducting the research.

4.3. Metrics

In order to analyze the performance of all the models, following the existing studies on bias (Lin et al., 2022; Majumdar et al., 2021; Singh et al., 2022) we evaluated the overall classification accuracy of the models, the standard deviation of accuracy across the demographics mentioned as the Degree of Bias (DoB), and the ratio of maximum and minimum accuracy values to quantify bias. A model is called fair when it obtains equivalent classification accuracy values across sub-groups, a low Degree of Bias which is close to 0, and a ratio of maximum and minimum accuracy values close to 1.

5. Results

In this section, we will discuss the performance of the baseline gender classifiers and the proposed bias mitigation techniques when trained on FairFace and evaluated on FairFace, UTKFace, and DiveFace datasets across gender and ethnicity.

Intra-Dataset Evaluation: Table 3 showed the performance of the baseline and proposed methods for gender classification when trained and tested on the FairFace dataset. Overall on applying consistency regularization, the Degree of Bias (DoB) and the ratio of max-min accuracy were significantly reduced and overall classification accuracy was improved across all the sub-groups. The Black Female subgroup performed the least for all the baseline models.

On evaluating the Baseline models based on DenseNet121, Vision Transformer, EfficientNet-V2 and ResNet-18 models, DenseNet-121 obtained the highest DoB of 5.05, followed by 2.85 (ViT), 2.69 (ResNet18) and 2.32 (EfficientNet-V2). Similarly, on evaluation of the ratio of max & min accuracy values across demographics on the Baseline models, DenseNet-121 obtained the highest ratio of 1.29, followed by 1.14 (ResNet18), 1.13 (ViT) and 1.1 (EfficientNet-V2). All four models obtained

higher DoB and max-min ratio. The baseline DenseNet121 obtained the least overall classification accuracy of 88.4% among the four architectures, followed by 92.5% (ResNet18), 92.82% (ViT) and 93.67% (EfficientNet-V2).

On applying consistency regularization technique, the overall classification accuracy of DenseNet121 increased to 93.59% from 88.4% when CR-Img Aug was applied, followed by 93.63%, 93.34% and 93.56% when CR-Img Aug(SN), CR-Feat Trans, and CR-Img Aug+Feat Trans(SN) were applied, respectively. The highest overall accuracy of 93.8% was obtained when CR-Img Aug+Feat Trans was applied. Similarly, the DoB reduced to 2.2 from 5.05 when CR-Img Aug and CR-Feat Trans were applied, followed by 2.02 and 2.03 when CR-Img Aug+Feat Trans and CR-Img Aug+Feat Trans(SN) were applied, respectively. The lowest DoB of 1.98 was obtained using CR-Feat Trans.

Similarly, the overall classification accuracy of ViT increased to 94.48% from 92.82% when CR-Img Aug(SN) was applied, followed by 93.84%, 93.89% and 94.06% obtained when CR-Feat Trans, CR-Img Aug+Feat Trans, and CR-Img Aug+Feat Trans(SN) were applied, respectively. The highest overall accuracy of 94.54% was obtained when CR-Img Aug was applied. Similarly, the DoB reduced to 1.91 from 2.85 when CR-Img Aug was applied, followed by 1.95, 1.98, and 1.87 when CR-Img Aug(SN), CR-Feat Trans, and CR-Img Aug+Feat Trans were applied, respectively. The lowest DoB of 1.8 was obtained using CR-Img Aug+Feat Trans(SN).

Also, the overall classification accuracy of EfficientNet-V2 increased to 93.93% from 93.67% when CR-Img Aug was applied, followed by 93.51%, 93.72% and 93.98% when CR-Img Aug, CR-Img Aug+Feat Trans, and CR-Img Aug+Feat Trans(SN) were applied respectively. The highest overall accuracy of 94% was obtained when CR-Img Aug(SN) was applied. Similarly, the DoB reduced to 1.93 from 2.32 when CR-Img Aug was applied, followed by 2.25, 2, and 2.04 when CR-Feat Trans, CR-Img Aug+Feat Trans, and CR-Img Aug+Feat Trans(SN) were applied, respectively. The lowest DoB of 1.74 was obtained using CR-Img Aug(SN).

Among techniques, CR-Img Aug or CR-Img Aug(SN) obtained the highest overall classification accuracy except on DenseNet121 architecture and increased by around 1-2%. CR-Img Aug+Feat Trans obtained the highest on DenseNet121 and increased by around 5%. Similarly CR-Img Aug+Feat Trans(SN) obtained the lowest DoB on ViT and ResNet18, on EfficientNet-V2 and DenseNet121, the lowest DOB was obtained on applying CR-Img Aug(SN).

Finally, the overall classification accuracy of ResNet18 increased to 93.13% from 92.5% when CR-Img Aug was applied, followed by 92.35%, 93.16% and 93.15% obtained when CR-Feat Trans, CR-Img Aug+Feat Trans, and CR-Img Aug+Feat Trans(SN) were applied, respectively. The highest overall accuracy of 93.47% was obtained when CR-Img Aug(SN) was applied. Similarly, the DoB reduced to 2.53 from 2.69 when CR-Img Aug, followed by 2.43, 2.82, and 2.49 when CR-Img Aug(SN), CR-Feat Trans, and CR-Img Aug+Feat Trans were applied, respectively. The lowest DoB of 2.11 was obtained using CR-Img Aug+Feat Trans(SN).

While looking at the ratio of max & min accuracy values across different methods, Baseline obtained the highest ratio when compared with other methods. The Baseline DenseNet121 obtained max-min ratio of 1.29, followed by 1.13 (Baseline ViT), then 1.1 (Baseline EfficientNetV2) and finally 1.14 on Baseline ResNet18. On applying consistency regularization methods, on DenseNet121, the ratio of max & min accuracy values reduced to the range between 1.07 - 1.1 from 1.29, the lowest was obtained using CR-Img Aug+Feat Trans. Similarly, on ViT, the ratio reduced to the range between 1.07 – 1.08 from 1.13, the lowest again was obtained using CR-Img Aug+Feat Trans. Similarly, on EfficientNetV2, the ratio reduced to the range of 1.07 – 1.09 from 1.1 after applying consistency regularization, the lowest was obtained using CR-Img Aug(SN) and CR-Img Aug+Feat Trans. Finally, on using ResNet18, the ratio after applying consistency regularization reduced to the range of 1.09 – 1.13 from 1.14, and the lowest was obtained using CR-Img Aug(SN).

Overall, we observed the CR-Img Aug+Feat Trans method obtained reduced bias and overall improved classification accuracy over other methods. Our proposed methods obtained an overall improvement of about 2 – 3% in classification accuracy, a reduction in the DoB of about 30%, and a reduction in the ratio of max-min accuracy values by about 7% as observed on the intra-dataset evaluation.

Cross-Dataset Evaluation: Table 4 and Table 5 show the cross-dataset evaluation i.e., the models are trained on FairFace and tested on UTKFace, and DiveFace. Overall, we observed a reduction in the DoB and an improvement in the overall classification accuracy using a consistency regularization-based technique for most of the models. Asian and White subgroups were the least and best-performing groups upon evaluation of Baseline models on the UTKFace dataset. Similarly, for DiveFace, almost similar classification performance was observed across demographic subgroups.

UTKFace: Table 4 tabulates the evaluation of the UTKFace dataset. We observed that for almost all the cases, Baseline models have the lowest overall classification accuracy, and the highest DoB, except for the EfficientNet-V2 Baseline model which obtained the lowest DoB. Across different architectures, the Baseline DenseNet121 model obtained the highest DoB of 4.93 and the lowest overall classification accuracy (90.43%). The DoB, on applying different consistency methods reduced to 2.56 (CR-Img Aug), 2.42 (CR-Img Aug(SN)), 2.6 (CR-Feat Trans), and 2.73 (CR-Img Aug+Feat Trans(SN)). The lowest DoB of 2.13 was obtained on CR-Img Aug+Feat Trans. Similarly, the overall classification accuracy increased to 93.74%, 93.88%, 93.24%, and 93.95% on applying CR-Img Aug, CR-Img Aug(SN), CR-Feat Trans, and CR-Img Aug+Feat Trans respectively. The highest overall classification accuracy of 94.21% was obtained using CR-Img Aug+Feat Trans(SN).

Similar to DenseNet121, the Baseline ViT obtained the least overall classification accuracy of 93.67% and the highest DoB of 3.22. The DoB reduced to 2.13, 1.91, 2.02, and 2 when CR-Img Aug, CR-Img Aug(SN), CR-Feat Trans, and CR-Img Aug+Feat Trans were applied, respectively. The lowest DoB of 1.86 was obtained when CR-Img Aug+Feat Trans(SN) was

applied. While the overall classification performance also improved to 94.6%, 94.74%, 94.06%, and 94.11% when CR-Img Aug, CR-Img Aug(SN), CR-Img Aug+Feat Trans and CR-Img Aug+Feat Trans(SN) were applied, respectively. The highest classification accuracy of 95.07% was obtained using CR-Feat Trans.

Similar to the previous two architectures, Baseline ResNet18 obtained the lowest overall classification and the highest DoB, 92.69% and 3.47, respectively. The overall classification performance increased to 93.40%, 93.64%, 94.01%, and 93.85%, respectively, using CR-Img Aug, CR-Img Aug(SN), CR-Feat Trans, CR-Img Aug+Feat Trans(SN), and finally, the highest accuracy of 94.26% was obtained using CR-Img Aug+Feat Trans. The DoB reduced to 3.08(CR-Img Aug), 2.46(CR-Img Aug(SN)), 2.29(CR-Img Aug+Feat Trans) and 2.22(CR-Img Aug+Feat Trans(SN)). For EfficientNet-V2, contrary to three previous observations, the lowest DoB of 1.81 was observed using the Baseline but neither the highest nor the lowest overall accuracy. The CR-Img Aug+Feat Trans(SN) obtained the lowest overall classification accuracy of 94.01%, followed by 94.23%(Baseline), 94.41%(CR-Img Aug), 94.27%(CR-Feat Trans), 94.37%(CR-Img Aug+Feat Trans) and finally, the highest accuracy of 94.63% obtained using CR-Img Aug(SN). The lowest DoB of 2.21 was obtained using CR-Feat Trans. Meanwhile, the highest DoB of 2.83 was obtained using CR-Img Aug+Feat Trans(SN), followed by 2.67 using both CR-Feat Trans and CR-Img Aug+Feat Trans, and finally, 1.92 and 1.87 was obtained using CR-Img Aug(SN) and CR-Img Aug, respectively.

With regard to max-min ratio, on DenseNet121, the Baseline obtained the highest max-min ratio of accuracy values at 1.17, and similarly, on ViT, it was 1.12, respectively. On applying consistency regularization on DenseNet121, the ratio reduced between $1.08-1.11$, while using ViT, the ratio reduced between $1.06-1.08$. ResNet18 obtained 1.11 and 1.12 using Baseline and CR-Img Aug, respectively. While applying other methods namely CR-Img Aug(SN), CR-Feat Trans, CR-Img Aug+Feat Trans, and CR-Img Aug+Feat Trans(SN), the ratio reduced between $1.07-1.09$. The ratio of max-min accuracy values for EfficientNetV2 was similar to the observations on DoB analysis, i.e., the highest ratio of 1.11 obtained on CR-Feat Trans, CR-Img Aug+Feat Trans and CR-Img Aug+Feat Trans(SN). The least ratio of 1.07 was obtained on Baseline, CR-Img Aug, and CR-Img Aug(SN).

Overall, we observed that on enforcing CR-Img Aug+Feat Trans, an increment of 2% in classification accuracy, a reduction of about 35% in the DoB, and similarly a reduction of 5% in the ratio of max-min accuracy values were obtained on UTKface.

On **DiveFace** dataset as well, the overall classification accuracy is increased and the DoB is reduced when consistency regularization is applied to different architectures as shown in Table 5. The lowest overall classification and the highest DOB were observed for all four Baseline models. The Baseline DenseNet121 obtained an overall accuracy of 96.16% and a DoB of 1.46, while the Baseline ViT obtained an overall accuracy of 97.97% and a DoB of 0.54. Similarly, the Baseline EfficientNetV2 and ResNet18 obtained an overall accuracy of 98.44% and 97.47% and a DoB of 0.63 and 0.88, respectively.

On DenseNet121, on applying different consistency regularization methods, an increment in the overall classification and reduction in DoB was observed. The overall accuracy increased to 98.31%, 97.93%, 98.2% and 97.7% when CR-Img Aug(SN), CR-Feat Trans, CR-Img Aug+Feat Trans and CR-Img Aug+Feat Trans(SN) were applied, respectively. The highest overall classification accuracy of 98.32% was obtained using CR-Img Aug. The DoB reduced to 0.62 from 1.46, followed by 0.55, 0.75, and 1.22 when CR-Img Aug, CR-Feat Trans, CR-Img Aug+Feat Trans and CR-Img Aug+Feat Trans(SN) was applied, respectively. The least DoB of 0.52 was obtained using CR-Img Aug(SN).

Similarly, on ViT, the overall accuracy improved to 98.51% using CR-Img Aug(SN), 98.38% using CR-Feat Trans, 98.41% using CR-Img Aug+Feat Trans and 98.45% using CR-Img Aug+Feat Trans(SN). Finally, the highest accuracy of 98.65% was obtained using CR-Img Aug. In contrary, in DoB reduced to 0.57 from 0.97, then 0.63, 0.8, and 0.76 using CR-Img Aug, CR-Img Aug(SN), CR-Img Aug+Feat Trans and CR-Img Aug+Feat Trans(SN), respectively. The least DOB of 0.54 was obtained on CR-Feat Trans.

On EfficentNet-V2, the overall accuracy improved except on CR-Feat Trans and CR-Img Aug+Feat Trans. These models obtained 98% and 98.44% respectively. The accuracy improved to 98.59% and 98.47% on applying CR-Img Aug(SN) and CR-Img Aug+Feat Trans(SN), respectively. The highest overall accuracy of 98.77% was obtained using CR-Img Aug. When it comes to the DoB, it was reduced except for CR-Feat Trans. On CR-Feat Trans, the DoB increased to 0.83, but it reduced to 0.57, 0.62, and 0.55 from 0.63 when CR-Img Aug, CR-Img Aug(SN), and CR-Img Aug+Feat Trans were applied, respectively. The least DoB of 0.36 was obtained using CR-Img Aug+Feat Trans(SN).

Finally, on ResNet18, the overall classification accuracy increased to 98.31%, 97.49%, 98.27%, and 98.38% when CR-Img Aug, CR-Feat Trans, CR-Img Aug+Feat Trans, and CR-Img Aug+Feat Trans were applied, respectively. The highest overall accuracy of 98.5% was obtained for CR-Img Aug(SN). In the case of the DoB, it reduced to 0.84, 0.69, 0.63, and 0.58 from 0.88 on applying CR-Img Aug, CR-Img Aug(SN), CR-Img Aug+Feat Trans, and CR-Img Aug+Feat Trans(SN), respectively. The least DoB of 0.53 was obtained when CR-Feat Trans was applied.

Overall, we observed an increment in classification accuracy of about 1%, a reduction of an average of 45% in the DoB, and a slight reduction of about 1% in the ratio of max & min accuracy was observed when evaluated on DiveFace. The ratio of maximum and minimum accuracy values across all models was close to 1 (in the range of 1 - 1.04) when consistency regularization techniques were applied. Thus, all the models were fair with respect to the ratio after applying the proposed bias mitigation techniques.

From the aforementioned results on intra- and cross-dataset scenarios, it is evident that on applying consistency regularization, there is an increment in the overall classification accuracy,

a reduction in the DoB, and the ratio of max&min accuracy values moved closer to 1. Therefore, it means the consistency regularization resulted in performance improvement and the reduction in the bias on both intra- and cross-dataset evaluation.

Comparative Analysis with SOTA: We have also compared the performance of our proposed with the popular SOTA bias mitigation techniques based on multi-tasking (Das et al., 2018), adversarial debiasing (Zhang et al., 2018), and deep generative views (Ramachandran and Rattani, 2022), proposed for gender classification and evaluated on Fairface and UTKFace datasets. Study in (Park et al., 2022) did not report gender classification accuracy across gender-ethnicity groups (Park et al., 2022), therefore we did not use it for cross-comparison. The algorithms were trained namely FairFace and tested namely FairFace and UTKFace using the same dataset. We chose ViT-CR Img Aug + Feat Trans(SN) as our proposed model for this comparative study.

For the comparative analysis, we used overall classification accuracy, the ratio of maximum and minimum accuracy values, and the DoB as metrics shown in Table 6. As can be seen in Table 6, our proposed model (ViT-CR Img Aug + Feat Trans(SN)) obtained the least DoB, and max-min ratio, and obtained the highest or the second-highest overall classification accuracy on UTKFace and FairFace test sets.

On the FairFace test set, our proposed method based on consistency regularization obtained the least DoB of 1.59 and the ratio of 1.05. Ramachandran and Rattani (Ramachandran and Rattani, 2022) obtained the highest overall classification accuracy of 94.72% on FairFace. While on UTKFace, our proposed method obtained the highest overall accuracy of 95.03%, the least DoB of 0.95, and the ratio of maximum and minimum accuracy values of 1.02. Therefore, our proposed method obtains state-of-the-art performance.

Worth mentioning, an existing technique based on adversarial debiasing (Zhang et al., 2018) obtained a trade-off between accuracy and fairness. This was due to the addition of the adversarial component which reduced the generalization capacity of the model. Also, adversarial debiasing and multi-tasking (Zhang et al., 2018; Das et al., 2018) based bias mitigation techniques need demographically annotated data. Further, the technique based on deep generative views is computationally expensive and is limited in its ability to synthesize images of the 3D scene with multi-view consistency. Therefore, compared to SOTA, our proposed technique has the advantage of mitigating bias in the absence of protected attributes, is computationally friendly, and is application agnostic. Further, our proposed approach has the advantage of enhanced fairness as well as classification accuracy.

6. Analysis and Discussion of the Results

In this section, we will discuss the three research questions addressed in this study.

To answer the **Research Question 1: "How effective is consistency regularization in mitigating bias?"** The analysis of the effectiveness of consistency regularization in mitigating bias is a crucial aspect of our research. Consistency regularization encourages a neural network to generate consistent outputs for different variations in input data. By reducing the intra-class distance and promoting the clustering of data samples belonging to the same class in the feature embedding space, consistency regularization enhances the network's ability to classify new inputs based on their similarity to well-clustered features. This, in turn, facilitates the linear separation of well-clustered features among different subgroups, thereby improving the classifier's performance by minimizing the overlap between different classes (Huang et al., 2021).

To evaluate the linear separability of feature embeddings for two gender classes, we examined the t-SNE (van der Maaten and Hinton, 2008) plot of feature embeddings obtained from different models: Baseline, CR-Img Aug, CR-Feat Trans, and CR-Img Aug+Feat Trans. For this analysis, we utilized a balanced dataset of 2500 image samples, encompassing various genders and ethnicities, extracted from the FairFace test set. We computed the feature embeddings from ViT-Baseline, ViT-CR-Img Aug, ViT-CR-Feat Trans, and ViT-CR-Img Aug+Feat Trans models.

Analysis of Figure 12 indicates that the feature embeddings from the ViT-Baseline model do not exhibit linear separability between classes, which contributes to performance disparities. However, when consistency regularization was applied, as demonstrated in Figures 15 and 14, the feature embeddings from ViT-CR Img Aug and ViT-CR-Feat Trans models respectively displayed reduced overlap, leading to a decrease in misclassification rates compared to the ViT Baseline model. These findings demonstrate an increase in classification accuracy and a reduction in bias. Similar trends were observed across different architectures and in cross-dataset evaluations.

Furthermore, the results presented in Tables 3, 4, and 5 illustrate that the application of consistency regularization (CR-Img Aug and CR-Feat Trans) resulted in a significant reduction in Degree of Bias (DoB) ranging from 22% to 60% in intra-dataset evaluations and 36% to 64% in cross-dataset evaluations across various architectures. Additionally, the ratio of maximum to minimum accuracy values decreased by approximately 4% and 3% in intra-dataset and cross-dataset evaluations, respectively.

Hence, based on these findings, it is evident that consistency regularization is undeniably effective in mitigating bias. Enhancing the feature representation for each sub-group, not only enhances classification accuracy but also comprehensively tackles bias-related concerns. Therefore, we can confidently assert that consistency regularization is a powerful tool in combating bias.

To answer the **Research Question 2: "Could the proposed technique obtain fairness as well as enhanced performance across several inter-sectional subgroups (including the best-performing group)?"** By employing the proposed technique, we observed a substantial increase in the average classification accuracy of the least-performing demographic group, amounting to approximately 4%. Furthermore, even the best-performing group obtained an improvement of around 1% in their accuracy, both within the dataset under consideration and

Table 3: Gender Classification Accuracy (%) on FairFace testset across different demographics. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and ethnicity; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and ethnicity. The top performance results are highlighted in bold.

Ethnicity	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White				
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F	Max/Min↓	Overall↑	DoB↓
Densenet121																	
Baseline	86.86	73.71	87.77	90.04	92.3	83.75	91.43	90.01	95.33	89.17	89.93	87.66	91.53	88.17	1.29	88.4	5.05
CR-Img Aug	90	88	93.56	94.95	94.16	94.63	93.19	96.38	96.43	95.2	92.65	93.68	94.03	93.46	1.096	93.59	2.2
CR-Img Aug(SN)	90.74	88.24	93.82	95.21	95.22	94.63	93.44	95.43	95.57	94.97	92.38	94.27	94.03	92.84	1.083	93.63	**1.98**
CR-Feat Trans	89.49	88.24	93.44	93.66	93.89	94.63	93.82	95.06	96.8	95.71	92.38	93.53	94.3	91.8	1.097	93.34	2.2
CR-Img Aug+Feat Trans	89.86	90.75	92.92	96.51	93.49	96.07	93.19	96.02	95.7	96	91.56	94.26	93.4	93.46	**1.074**	**93.8**	2.02
CR-Img Aug+Feat Trans(SN)	89.36	89.3	93.56	93.92	94.95	94.63	93.44	95.3	96.31	94.95	91.7	94.12	95.1	93.25	1.078	93.56	2.03
Vision Transformer(ViT)																	
Baseline	89.86	84.94	91.38	93.66	93.09	94.49	93.44	94.71	96.31	94.46	90.75	94.27	94.56	93.57	1.133	92.82	2.85
CR-Img Aug	90.5	90.22	93.82	94.82	95.62	95.81	94.45	95.79	97.17	95.96	94.15	95	95.81	94.4	1.077	**94.54**	1.91
CR-Img Aug(SN)	90.24	90.75	93.56	94.7	95.22	95.54	94.2	96.51	97.42	95.72	93.06	95.74	95.54	94.5	1.08	94.48	1.95
CR-Feat Trans	91.61	90	94.85	92.37	96.06	92.92	95.21	94.22	97.29	93.18	94.15	93.09	96.7	92.11	1.081	93.84	1.98
CR-Img Aug+Feat Trans	90.36	89.83	92.92	94.18	94.29	94.89	93.32	96.02	96.19	95.71	92.65	94.85	95.19	94	**1.071**	93.89	1.87
CR-Img Aug+Feat Trans(SN)	91.24	89.7	93.18	94.57	94.82	95.02	93.44	95.78	96.43	95.71	92.79	94.41	95.63	94.08	1.075	94.06	**1.8**
EfficientNet-V2																	
Baseline	87.86	90.22	93.31	95.6	94.29	95.28	92.43	95.54	96.56	96	92.38	95.15	93.76	93.04	1.1	93.67	2.32
CR-Img Aug	90.74	89.7	93.31	95.6	95.48	92.92	93.44	95.78	96.8	95.71	92.65	94.71	94.12	94	1.079	93.93	1.93
CR-Img Aug(SN)	90.36	90.49	94.34	94.44	96.02	93.97	93.95	95.06	96.8	94.95	92.95	95	94.3	93.35	1.071	**94**	**1.74**
CR-Feat Trans	88.61	90.75	92.28	93.92	94.69	95.15	91.8	95.26	96.31	96.21	91.56	95.44	93.14	94.08	1.087	93.51	2.25
CR-Img Aug+Feat Trans	89.86	90	92.54	94.86	95.62	93.97	94.2	96.02	95.82	95.45	91.16	94.56	93.67	94.3	**1.069**	93.72	2
CR-Img Aug+Feat Trans(SN)	89.49	90.62	94.21	93.79	95.88	93.18	93.69	95.54	96.56	97.22	92.52	94.12	94.21	94.7	1.086	93.98	2.04
ResNet18																	
Baseline	89.61	84.54	92.66	93.01	94.56	92.27	93.44	93.86	96.56	93.94	91.29	93.23	93.85	92.21	1.142	92.5	2.69
CR-Img Aug	87.86	88.24	91.51	95.34	93.89	94.23	93.69	95.9	96.56	94.95	90.88	93.68	93.49	93.56	1.1	93.13	2.53
CR-Img Aug(SN)	89.11	88.9	92.15	94.18	95.35	94.89	93.95	96.14	96.92	95.45	90.07	93.82	94.03	93.67	**1.09**	**93.47**	2.43
CR-Feat Trans	86.86	86.13	91.38	93.01	93.76	93.84	93.57	93.73	97.17	93.18	89.93	93.38	93.58	93.35	1.128	92.35	2.82
CR-Img Aug+Feat Trans	89.11	86.79	94.21	94.05	95.09	93.7 1	93.32	95.06	96.92	93.69	91.29	93.68	94.65	92.73	1.117	93.16	2.49
CR-Img Aug+Feat Trans(SN)	91	87.45	92.41	94.44	93.76	93.71	93.06	95.3	96.19	93.94	90.88	93.68	94.47	93.77	1.1	93.15	**2.11**

Table 4: Gender Classification Accuracy (%) on UTKFace testset across different demographics. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and ethnicity; Overall and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and ethnicity. The top performance results are highlighted in bold.

Ethnicity	Asian		Black		Indian		White				
Gender	M	F	M	F	M	F	M	F	Max/Min↓	Overall↑	DoB↓
DenseNet121											
Baseline	89.18	84.88	98.71	84.62	93.8	87.21	94.15	90.89	1.167	90.43	4.93
CR-Img Aug	92.36	90.81	98.7	93.18	95.57	90.64	95.61	93.04	1.089	93.74	2.56
CR-Img Aug(SN)	91.72	91.89	98.7	93.18	95.13	90.64	95.25	94.56	1.089	93.88	2.42
CR-Feat Trans	90.45	91.89	98.27	90	94.69	91.81	96	92.83	1.092	93.24	2.67
CR-Img Aug+Feat Trans	91.08	90.27	97.4	94.54	95.13	94.15	94.88	94.13	**1.079**	93.95	**2.13**
CR-Img Aug+Feat Trans(SN)	88.54	92.43	98.7	94.1	94.69	95.61	95.25	94.35	1.114	**94.21**	2.73
Vision Transformer(ViT)											
Baseline	91.07	94.05	99.14	88.68	95.57	92.44	95.79	92.62	1.118	93.67	3.22
CR-Img Aug	91.08	92.97	98.7	93.64	95.13	94.15	96.34	94.78	1.084	94.6	2.13
CR-Img Aug(SN)	92.36	92.43	98.27	93.18	95.13	94.74	96.34	95.43	**1.064**	94.74	1.91
CR-Feat Trans	94.27	92.97	99.13	92.73	95.57	95.32	96.89	93.7	1.069	**95.07**	2.02
CR-Img Aug+Feat Trans	90.45	93.51	97.83	93.18	95.13	92.98	95.25	94.13	1.082	94.06	2
CR-Img Aug+Feat Trans(SN)	91.08	92.43	97.4	92.73	94.69	94.15	95.61	94.78	1.069	94.11	**1.86**
EfficientNet-V2											
Baseline	90.45	95.13	96.1	93.18	94.69	96.49	93.23	94.56	1.067	94.23	**1.81**
CR-Img Aug	91.72	93	98.27	93.18	95.57	94.15	95.25	94.13	1.071	94.41	1.87
CR-Img Aug(SN)	93	92.97	98.7	92.73	96.46	94.15	94.88	94.13	**1.064**	**94.63**	1.92
CR-Feat Trans	87.9	94.05	97.4	94.1	95.58	95.91	93.6	95.65	1.108	94.27	2.67
CR-Img Aug+Feat Trans	89.81	90.81	98.7	95.91	95.57	94.74	95.25	94.13	1.099	94.37	2.67
CR-Img Aug+Feat Trans(SN)	94.27	89.73	98.7	92.27	96.02	90.64	96.34	94.13	1.1	94.01	2.83
ResNet18											
Baseline	89.81	88.65	98.7	89.54	97.34	91.81	94.15	91.52	1.113	92.69	3.47
CR-Img Aug	87.9	90.81	98.7	92.73	96.02	94.15	94.7	92.17	1.123	93.4	3.08
CR-Img Aug(SN)	91.08	92.43	98.7	92.27	96.02	91.23	94.52	92.83	1.084	93.64	2.46
CR-Feat Trans	91.08	91.89	97.84	91.82	96.46	94.15	94.7	94.13	**1.074**	94.01	**2.21**
CR-Img Aug+Feat Trans	93	91.35	98.27	91.82	96.9	92.98	95.43	94.35	1.076	**94.26**	2.29
CR-Img Aug+Feat Trans(SN)	90.45	91.89	98.27	93.18	95.57	92.98	94.33	94.13	1.086	93.85	2.22

across different datasets during our evaluation process. These improvements signify the effectiveness of our proposed method in achieving fairness while simultaneously enhancing performance across various inter-sectional subgroups.

Delving deeper, we found that our obtained yielded an overall increment in the classification accuracy ranging from 0.23% to 6% for intra-dataset evaluations and 0.3% to 4% for cross-dataset evaluations. These figures, as presented in Tables 3, 4, and 5, highlight the consistent and positive impact of our proposed technique on classification accuracy across diverse scenarios. Moreover, our approach demonstrated a significant reduction in the Degree of Bias (DoB) by approximately 30%. This reduction indicates the successful mitigation of biases that may exist within the dataset, further emphasizing the fairness obtained by our technique. Additionally, the ratio of the maximum and minimum accuracy values, which serves as a measure of performance disparity, moved closer to 1. This reduction in performance disparity signifies the effectiveness of our tech-

Table 5: Gender Classification Accuracy(%) on DiveFace testset across different demographics. M stands for Male, and F stands for Female. Max/Min is the ratio of maximum and minimum classification accuracy values among gender and ethnicity; Avg and DoB are the overall classification accuracy and the standard deviation of the accuracy values across gender and ethnicity. The top performance results are highlighted in bold.

Ethnicity	East Asian		Sub Saharan South Indian		White				
Gender	M	F	M	F	M	F	Max/Min↓	Overall↑	DoB↓
DenseNet121									
Baseline	94.65	98.29	95.94	94.46	96.6	97	1.041	96.16	1.46
CR-Img Aug	98.05	99.26	97.49	98	98.76	98.33	1.018	**98.32**	0.62
CR-Img Aug(SN)	98.05	99.08	97.4	98.24	98.67	98.41	1.017	98.31	**0.52**
CR-Feat Trans	98.19	98.2	97.27	97.52	98.76	97.66	**1.015**	97.93	0.55
CR-Img Aug+Feat Trans	97.42	99.08	97.17	98.78	98.32	98.41	1.02	98.2	0.75
CR-Img Aug+Feat Trans(SN)	97.6	99.17	97.35	98	95.59	98.46	1.037	97.7	1.22
Vision Transformer(ViT)									
Baseline	97.14	99.17	96.58	98.06	98.76	98.11	1.027	97.97	0.97
CR-Img Aug	98.28	99.52	97.67	98.65	98.98	98.77	1.019	**98.65**	0.57
CR-Img Aug(SN)	98	99.39	97.49	98.38	98.98	98.81	1.019	98.51	0.63
CR-Feat Trans	98.64	98.9	97.31	98.24	98.9	98.28	**1.016**	98.38	**0.54**
CR-Img Aug+Feat Trans	97.46	99.43	97.22	98.74	98.98	98.63	1.023	98.41	0.8
CR-Img Aug+Feat Trans(SN)	97.78	99.52	97.22	98.42	98.94	98.81	1.024	98.45	0.76
EfficientNet-V2									
Baseline	98.23	99.21	97.26	98.83	98.85	98.24	1.02	98.44	0.63
CR-Img Aug	98.59	99.65	97.72	99	98.85	98.81	1.02	**98.77**	0.57
CR-Img Aug(SN)	98.46	99.56	97.49	98.92	98.54	98.59	1.021	98.59	0.62
CR-Feat Trans	97	98.73	96.72	98.78	98.36	98.41	1.021	98	0.83
CR-Img Aug+Feat Trans	98.78	99.21	97.63	98.02	98.85	98.15	1.016	98.44	0.55
CR-Img Aug+Feat Trans(SN)	98.82	98.64	98.31	98.02	98.94	98.06	**1.009**	98.47	**0.36**
ResNet18									
Baseline	97.64	98.47	97.36	96.13	98.54	96.65	1.025	97.47	0.88
CR-Img Aug	97.64	99.47	96.9	98.78	98.23	98.81	1.027	98.31	0.84
CR-Img Aug(SN)	98.05	99.43	97.26	98.51	98.76	99	1.022	**98.5**	0.69
CR-Feat Trans	97.14	98.25	96.63	97.3	97.79	97.84	**1.017**	97.49	**0.53**
CR-Img Aug+Feat Trans	98.19	99.17	97.54	97.52	99	98.19	**1.017**	98.27	0.63
CR-Img Aug+Feat Trans(SN)	98.55	99.17	97.4	97.88	98.67	98.63	1.018	98.38	0.58

Table 6: Comparative Analysis. A: Multi-Tasking (Das et al., 2018), B: Adversarial debiasing (Zhang et al., 2018), D: Deep Generative Views based (Ramachandran and Rattani, 2022). The top performance results are highlighted in bold.

Method	Accuracy								DoB↓	Max/Min↓
	Black	East Asian	Indian	Latino Hispanic	Middle Eastern	Southeast Asian	White	Overall↑		
FairFace										
A	91.26	94.45	95.05	95.19	97.35	94.2	94.96	94.64	1.81	1.067
B	87.66	91.93	93.67	93.8	95.96	91.81	93.96	92.69	2.62	1.095
D	91.64	95.29	95.38	95.32	97.11	93.5	94.92	**94.72**	1.72	1.06
Ours	90.83	93.6	94.48	94.7	95.94	93.64	94.57	94	**1.59**	**1.056**
UTKFace										
B	94.62	-	93.65	-	-	91.89	94.97	93.78	1.38	1.03
Ours	95.85	-	95.43	-	-	93.67	95.16	**95.03**	**0.95**	**1.02**

nique in equalizing the performance across different subgroups, leading to fairer outcomes.

Hence, to address the second research question, we have obtained compelling and comprehensive results from intra- and cross-dataset evaluations. These results unambiguously support our claim that our proposed technique not only achieves fairness but also significantly enhances performance across various inter-sectional subgroups, including the top-performing group. The substantial improvements in classification accuracy, reduction in bias, and decreased performance disparity serve as robust evidence of the effectiveness and importance of our approach.

To answer the final **Research Question 3: "Can the combination of image-level and feature-level augmentation techniques obtain any advantage over either of them individually?"** The experimental results presented in Tables 3, 4, and 5 highlight the substantial advantages gained through the combination of image-level and feature-level augmentation techniques compared to their individual usage. By employing the CR-Img Aug+Feat Trans technique across various models, we observed an average reduction of 8% in the ratio of maximum and minimum accuracy values on intra-dataset evaluations and approximately 3% on cross-dataset evaluations. In contrast, when applying CR-Img Aug or CR-Feat Trans techniques individually, we obtained a 6% reduction in the ratio of max-min accuracy on intra-dataset evaluations and an average of 2% reduction on cross-dataset evaluations.

Additionally, the t-SNE plots depicted in Figures 12, 13, 14, and 15, based on feature embeddings from ViT-Baseline, ViT-CR-Img Aug, ViT-Feat Trans, and ViT-Img Aug+Feat Trans models, respectively, further strengthen the significance of the combined augmentation approach. While the classes were not linearly separable in the feature embedding space for the former models, the t-SNE plot obtained from the ViT-Img Aug+Feat Trans model showcased linear separability between the two classes (males and females) with minimal overlap. Moreover, this model effectively distinguished features among different ethnicities but displayed the most overlap among male and female kids, indicating that gender-specific features are less prominent in younger age groups. Similar trends were observed across all architectures and proposed mitigation techniques, both in intra-dataset and cross-dataset evaluations.

Hence, to address the research question at hand, we can confidently state that combining image-level and feature-level augmentations within a consistency-based regularization framework significantly reduces bias. This novel approach enhances the discriminatory power of the models and diminishes the po-

tential for biases based on gender, ethnicity, or other factors, as demonstrated by compelling empirical evidence.

Overall, the findings demonstrate that consistency regularization, combined augmentation techniques, and the proposed framework effectively mitigate bias, enhance performance, and achieve substantial fairness in classification tasks.

7. Statistical Inference

In this section, we will discuss the performed hypothesis testing for the statistical significance of the obtained results due to the proposed consistency regularization technique (Image level, Feature level, and Combination of both) in bias reduction.

Hypothesis testing: In statistics, hypothesis testing is one way of doing statistical inference by determining if survey or experiment results are relevant. By calculating the likelihood that the results occurred by chance, it can be determined whether the observed results are genuine/ significant. *The experiments will not be repeatable and therefore will not be of much help if the results have been obtained by accident/ chance.* Therefore, we used hypothesis testing to validate whether the claimed bias reduction was obtained by accident or not using our proposed bias reduction technique.

To begin with, we have to come up with an appropriate null hypothesis (H_o) and an alternate hypothesis (H_a). Through our proposed bias mitigation technique, we aimed to obtain **demographic parity**, which means an equal proportion of true positive predictions in each group (in this case, demographics). Given the aim, the **Null Hypothesis,** H_o and the **Alternate Hypothesis,** H_a for our statistical inference will be,
H_o: *There is no significant difference in the overall true positive predictions across the groups*;
H_a: *Overall true positive predictions differs in at least two groups in the population.* Therefore to claim that our results are fair, we should accept the null hypothesis or reject the alternate hypothesis.

How to know whether the null hypothesis is to be accepted or rejected?

The p-value or probability value in statistics is the probability that the observations have been obtained under the null hypothesis. The p-value is used to determine the smallest level of significance at which the null hypothesis would be rejected. A lower p-value indicates that there is more evidence supporting the alternative hypothesis. The second step of hypothesis testing is to collect data. We have used the overall gender classification accuracy values across demographics as the data. We chose the overall gender classification accuracy across ethnicities from the ViT architecture, compared across Baseline, CR-Img Aug, CR-Feat Trans, and CR-Img Aug+Feat Trans from intra-dataset and cross-dataset evaluations.

The collection of data is followed by performing the statistical test. The statistical test is essentially based on the comparison of within-group variance (the spread of the data within a group) versus between-group variance (how different the categories are from one another). We used **One-Way ANOVA (ANalysis Of VAriance)** test (Girden, 1992) for hypothesis testing. It uses F-distribution to compare the means of two or more independent groups in order to determine whether there is statistical evidence that associated gender classification accuracies from the population are significantly different.

To perform the test, we have split the test datasets into five equal-sized subsets, also known as folds, and then evaluated the overall classification accuracy between each subgroup across five folds as shown in Table 7, 8, and 9. This gives an idea of how the overall classification accuracy varies between the subgroup and within the subgroup on intra and cross-dataset evaluations. Once the variation between the subgroups and within the subgroups was obtained, its ratio gave the F-statistic. Then from the F-distribution table, the p-value was obtained.

Analysis: According to the inferential evaluation on the intra and cross-dataset evaluations, we have obtained p-values from the observations of four models when used on ViT namely, Baseline, CR-Img Aug, CR-Feat Trans, and CR-Img Aug+Feat Trans given in Table 10.

From Table 10, on the intra-dataset evaluation with the FairFace test set, we observed that the p-value was increased as consistency regularization was applied, Baseline obtained a p-value of $2.88 \times 10^{-}7$, followed by $2.27 \times 10^{-}7$ on CR-Img Aug. Whereas on applying CR-Feat Trans, p-value increased to $9.93 \times 10^{-}7$. Finally, when image- and feature-level were combined together (CR-Img Aug+Feat Trans), the p-value increased by a factor of 10. The Baseline obtained the least p-value, which means it supports the alternative hypothesis which indicates the presence of bias. The p-values increased by a factor of 10^2 and 10^3, respectively, when CR-Img Aug and CR-Feat Trans were applied. When the combination of both CR was applied, the p-value was further increased by a factor of 10^3. This indicates CR-Img Aug+Feat Trans has the most impact on bias reduction.

On cross-dataset evaluation with UTKFace, from Table 10, we observed the Baseline model obtained the second least p-value of $2.88e-07$, and after combining CR-Img Aug+Feat Trans, it obtained the highest p-value, which was increased by a factor of 10 when compared with other models namely, Baseline, CR-Img Aug, and CR-Feat Trans. This indicates CR-Img Aug+Feat Trans has resulted in significant bias reduction on the UTKFace dataset.

But on the cross-dataset evaluation with DiveFace, we observed a difference in the trend. Evaluations across different models and across architecture have given near-perfect results on the DiveFace, from Table 10, we observed the lowest value for p-value was obtained with CR-Img Aug+Feat Trans of 0.122, and CR-Img Aug obtained the highest value of 0.678. This suggests that on DiveFace, CR-Img Aug has more impact on bias reduction as compared with CR-Feat Trans and CR-Img Aug+Feat Trans. A similar observation was made for the models along with spectral normalization applied.

The higher p-value for models based on consistency regularization when compared to the baseline model (where no consistency regularization was applied) confirms the acceptance of the null hypothesis. This confirms that the bias reduction obtained using our proposed consistency regularization-based techniques is significant and not obtained by chance .

Table 7: Overall Gender Classification Accuracy (%) on FairFace testset across five folds on each demographic group using ViT architecture across five folds.

	Black	East Asian	Indian	Latino Hispanic	Middle Eastern	Southeast Asian	White
Baseline							
Fold 1	86.44	92.04	94.47	93.53	95.88	92.9	93.32
Fold 2	86.97	93.45	94.45	94.35	95.62	92.84	93.71
Fold 3	87.69	92.87	94.7	92.31	95.22	92.54	95.65
Fold 4	89.53	93.03	94.38	94.87	96.3	93.12	94.53
Fold 5	87.27	93.82	94.45	93.19	96.03	92.42	94.33
CR-Img Aug							
Fold 1	87.85	94.08	95.29	94.75	97.18	95.74	94.7
Fold 2	89.58	95.63	94.26	95.97	96.72	94.68	95.05
Fold 3	90.04	94.65	96.54	95.34	96.52	94.96	95.27
Fold 4	90.96	94.22	95.38	96.65	97.61	93.96	95.31
Fold 5	89.1	95.02	95.48	93.79	97.36	93.89	95.84
CR-Feat Trans							
Fold 1	88.26	93.47	94.67	94.14	95.88	93.91	94.3
Fold 2	89.58	94.44	94.85	95.77	95.19	93.66	94.86
Fold 3	90.82	94.45	96.33	94.53	95.43	92.94	95.46
Fold 4	90.55	93.62	94.78	95.86	96.52	92.29	94.53
Fold 5	90	94.42	95.28	93.59	96.47	92.42	94.52
CR-Img Aug + Feat Trans							
Fold 1	87.25	92.86	94.67	93.74	96.96	93.91	94.7
Fold 2	90.18	94.44	93.86	95.36	95.62	93.86	94.86
Fold 3	90.23	94.45	95.93	94.33	95.43	93.75	94.9
Fold 4	92	94.62	94.18	96.45	96.96	93.54	94.14
Fold 5	90.3	94.22	94.66	94.79	96.26	93.89	96.41

Table 8: Overall Gender Classification Accuracy (%) on UTKFace testset across five folds on each demographic group using ViT architecture across five folds.

	Asian	Black	Indian	White
Baseline				
Fold 1	90.49	94.95	95.5	94.02
Fold 2	90.59	94.6	93.78	95.56
Fold 3	89.27	95.1	94.33	95.52
Fold 4	91.06	94.94	97.26	93.82
Fold 5	89.32	94.3	94.88	95.2
CR-Img Aug				
Fold 1	92.14	95.67	96.04	95.56
Fold 2	92.07	96.22	94.88	95.73
Fold 3	90.4	96.2	95.25	96.21
Fold 4	92.15	97.29	97.08	95.02
Fold 5	92.08	95	95.79	96.56
CR-Feat Trans				
Fold 1	92.69	96.03	96.22	95.04
Fold 2	91.51	95.5	94.7	95.56
Fold 3	91.34	96.19	94.7	96.04
Fold 4	91.97	96.38	96.71	93.65
Fold 5	90.24	95.19	96.52	95.36
CR-Img Aug+Feat Trans				
Fold 1	91.59	95.85	95.86	95.04
Fold 2	92.25	95.5	94.33	95.56
Fold 3	90.77	95.46	95.43	96.56
Fold 4	91.61	96.38	97.08	94.34
Fold 5	92.08	94.12	95.8	95.88

8. Ablation Study

8.1. How are the models trained with consistency regularization performed on facial feature occlusion?

We have evaluated the performance of the baseline models (Baseline) and consistency regularized models (CR-Img Aug, CR-Feat Trans & CR-Img Aug+Feat Trans) on test face images with occlusions. These occlusions were generated by masking various facial features such as cheeks, eyes, nose, forehead, mouth, and chin (See Figure 10) by locating the facial landmarks using dlib (King, 2009). The intuition behind this study was to validate the claim that consistency regularization-based models are more robust to variations, therefore should outperform the baseline models on test samples with occluded facial features.

Table 9: Overall Gender Classification Accuracy (%) on DiveFace testset across five folds on each demographic group using ViT architecture across five folds.

	East Asian	Sub Saharan South Indian	White
Baseline			
Fold 1	98.6	98.03	98.2
Fold 2	98.24	96.15	98.05
Fold 3	99.11	98.02	98.61
Fold 4	97.92	99.28	98.02
Fold 5	98.03	97.88	98.39
CR-Img Aug			
Fold 1	98.78	98.92	98.74
Fold 2	99.12	97.72	98.94
Fold 3	99.29	98.74	98.96
Fold 4	98.61	99.46	98.92
Fold 5	98.92	98.76	98.75
CR-Feat Trans			
Fold 1	98.6	98.03	98.92
Fold 2	99.12	96.85	98.94
Fold 3	98.94	97.12	98.09
Fold 4	98.44	98.74	98.02
Fold 5	97.67	98.23	98.57
CR-Img Aug+Feat Trans			
Fold 1	98.6	98.92	98.92
Fold 2	98.42	97.38	98.94
Fold 3	98.75	98.2	98.61
Fold 4	98.61	98.74	98.56
Fold 5	98.21	98.23	99.28

Table 10: p-values obtained using One-way ANOVA test for statistical validation of the results for different bias mitigation techniques in intra- and cross-dataset evaluation.

	Baseline	CR-Img Aug	CR-Feat Trans	CR-Img Aug Feat Trans
FairFace	6.9e-15	7.24e-13	4.44e-12	**2.99e-09**
UTKFace	2.88e-07	2.27e-07	9.93e-07	**1.34e-06**
DiveFace	0.525	**0.678**	0.132	0.122

Table 11 shows the average accuracy, DOB, and the ratio of max-min accuracies for the ViT-Baseline model and the ViT (CR-Img Aug, Img Feat Trans, and CR-Img Aug+Feat Trans). It can be observed that overall classification accuracy was improved over the baseline after consistency regularization was applied to the test samples with occluded facial regions. This indicates the robustness of the model trained with consistency regularization to the facial region occlusion over the baseline.

Overall classification accuracy was reduced by an average of 3% when Baseline models were evaluated on occluded test samples. Using consistency regularization, the average classification accuracy increased by an average of $2-3\%$ using CR-Img Aug, CR-Feat Trans & CR-Img Aug+Feat Trans over the baseline. Further, the reduction in the DOB and the max-min ratio was about 25%, 5% for CR-Img Aug, CR-Feat Trans, and CR-Img Aug+Feat Trans with respect to the Baseline.

Fig. 10: a. Test Sample; b. Nose region masked; c. Mouth region masked; d. Forehead masked; e. Eyes masked; f: Chin masked & g: Cheeks masked.

Table 11: Evaluation of the robustness of the gender classification models trained with CR-Img, CR-Feat Trans, and CR-Img Aug+Feat Trans techniques against Baseline on test samples with facial region "occlusion". ViT architecture and FairFace testsets were used for this experiment.

Masked Facial Feature	Cheeks	Chin	Eye	Forehead	Mouth	Nose
Baseline						
Overall↑	87.8	91.33	88.92	89.77	91.44	90.62
DoB↓	5.58	4.2	5.51	4	4.08	4.48
Max/Min↓	1.36	1.24	1.22	1.2	1.23	1.27
CR-Img Aug						
Overall↑	91.34	93.26	91.77	92.29	93.11	92.53
DoB↓	3.85	3.49	4.56	2.76	4.36	4.66
Max/Min↓	1.22	1.18	1.26	1.14	1.22	1.24
CR-Feat Trans						
Overall↑	90.6	93.1	91.82	92.12	92.88	93.03
DoB↓	3.57	3.34	3.04	2.83	3.8	2.74
Max/Min↓	1.17	1.18	1.14	1.13	1.2	1.13
CR-Img Aug+Feat Trans						
Overall↑	90.55	92.97	91.52	91.7	93	93.05
DoB↓	4.48	3.06	4	3.2	3.49	2.98
Max/Min↓	1.26	1.16	1.16	1.14	1.19	1.15

8.2. Choosing the right set of image augmentations for enforcing consistency regularization.

The aim of this study was to find out the right set of augmentation to improve performance and reduce bias. We did experiment with a different set of augmentations. Table 12 showed various combinations of augmentations for ViT-CR-Img Aug against no augmentations (baseline model). The chosen combination is labeled as "Best" with the lowest DoB and that ratio of max-min accuracy values. On training with the multiple set of augmentations, we chose the set with the lowest DoB on the FairFace test set for all the experiments. The fairest performance was obtained when the combination of random erase, colorjit augmentations along with enhanced brightness and contrast were applied, where we obtained the overall classification accuracy as 94.48%, the DoB as 1.95, and the ratio of maximum and minimum accuracies as 1.08. From Table 12, we observed the higher values of DoB and the ratio of maximum and minimum accuracy values on various other combinations of augmentations. The reason why the combination of random erase, colorjit along with enhanced brightness and contrast obtained the fair result could be because of reduced similarity in facial morphology when the random erase was used. Further, applying colorjit, enhanced brightness, and contrast-based augmentation reduces the impact of skin color. This indicates the importance of choosing the right set of augmentations for consistency regularization for the downstream classification task.

Worth mentioning we also did an experiment with a set of augmentations automatically obtained using AutoAugment (Cubuk et al., 2019) based pretrained policies. The obtained results were poor. For instance, we obtained a Degree of Bias of 2.422, 2.087, and 0.927 on FairFace, UTKFace and DiveFace testsets using ViT-CR Img AutoAug (ViT with AutoAugment). In comparison, using CR-Img Aug, we obtained a DoB of 1.95, 1.91, and 0.63 on FairFace, UTKFace, and DiveFace, respectively. Similarly, CR-Img Aug+Feat Trans obtained a DoB of 1.87, 2, and 0.8 on FairFace, UTKFace, and DiveFace, respectively. This suggests the efficacy of our approach over AutoAugment. This could be due to the fact that Auto-Augment policies learned on ImageNet for general image classification tasks did not transfer well to obtain significant improvements for our case.

Further, our proposed bias mitigation technique has a significant advantage over random data augmentation applied to the training set. This is because our proposed approach based on systematic image and feature-level augmentations using *StyleNeRF* has a significant impact on enhancing the underlying data manifold of the samples. Further, consistency-based regularization enforces the feature embeddings from these perturbations (from the underlying data manifold) to be similar. This significantly enhances the feature representation for each sub-group. Consequently, enhancing the accuracy and reducing the bias of the classifier. These set of best augmentations were also combined with the transformed feature vector with the control variables (pitch, yaw, roll, and fov) randomly selected (obtained using the trained MLP) in a consistency-based regularization setting as a separate experiment.

Table 12: The evaluation of the ViT-CR-Img Aug trained with different combinations of augmentations. The evaluation was performed on the FairFace test set. We observed similar observations on other models as well. A: Random Perspective; B: Random Rotate; C: ColorJitter; D: Random Erase; E: Gaussian Blur; F: GrayScale; G: Brightness+Contrast; H: Random Flip.

Augmentations	Overall↑	DoB↓	Max/Min↓
No Augmentations (baseline)	93.66	2.65	1.14
A	93.33	3.17	1.16
E+F	93.82	2.57	1.09
E+G	93.74	2.54	1.1
A+C	93.51	2.53	1.11
A+D	93.59	2.49	1.12
H+C	93.96	2.38	1.08
H+D	94.3	2.36	1.07
G+D	94.71	2.33	1.08
G+C	93.97	2.31	1.08
H+A	93.78	2.24	1.1
B+C+D	94.25	2.58	1.09
A+G+D	94.05	2.54	1.11
H+D+C	94.45	2.48	1.08
D+G+C (Best)	94.48	1.95	1.08

8.3. On varying the weightage between the consistency regularization loss and the classification loss?

In all the aforementioned experiments, equal weightage was given to the consistency and classification loss (refer equation 4). We have evaluated the impact of varying the weightage between these two loss functions. Table 13 shows the performance of the ViT-CR-Img Aug+Feat Trans. As can be seen from Table 13, overall classification accuracy was consistently increased for the lambda(λ) value in the range of 2 & 12. At the same time, DOB and the max-min ratio were also reduced for this range of lambda. The classification accuracy was reduced and the bias of the classifier was increased on the lambda value over 12. However, we obtained the best result when the classification loss and consistency regularization loss were given equal weightage. From Table 3, we obtained an overall accuracy of 94.06%, a DoB of 1.8, and a ratio of maximum and minimum accuracy as 1.075. A similar observation was made for different architectures across proposed bias mitigation techniques and the datasets.

This experiment validates the importance of balance in the weightage between the classification loss and the consistency

regularization loss. If the weight constant of the consistency regularization loss is too low, then the regularization effect may not be sufficient to prevent overfitting, resulting in poor generalization performance. If the classification loss is too strong, dominating the training process and causing the network to focus too much on the training data at the expense of low generalization.

Table 13: Varying the contribution of consistency and classification loss by changing the lambda value(Refer Eq: 4). These results are obtained for the ViT-CR-Img Aug+Feat Trans evaluated on the FairFace test set. We observed the same trend for other models as well.

Lambda	1	2	5	10	12	100	1000
Overall↑	94.06	93.81	94.24	94.27	**94.32**	93.37	83.5
DoB↓	**1.8**	2.07	2	2	2.086	2.75	11.92
Max/Min↓	**1.075**	1.094	1.065	1.085	1.0945	1.143	1.837

8.4. Did Feature Transformer MLP learn the gender cues for the classification task?

The aim of this experiment was to verify the capability of feature transformer MLP in retaining the gender cues in the transformed feature vector for the gender classification task. For the experiment, we used ViT-Baseline and the feature transform model which was trained with the features extracted from the FairFace training set using the ViT-Baseline model and evaluated on the FairFace, UTKFace, and DiveFace test sets.

As shown in Figure 11, the feature vector of the input face image was extracted from the feature extraction layer. This feature vector of the image along arbitrarily chosen pitch, roll, yaw, and fov was input to the feature transform model. We generated only one feature vector for each test image (from the testing part of the dataset), and then the transformed feature vector from the feature transform model was used as an input to the classification layer (final output layer of the ViT-Baseline model) (Training of the feature transform was discussed in Section 3.3).

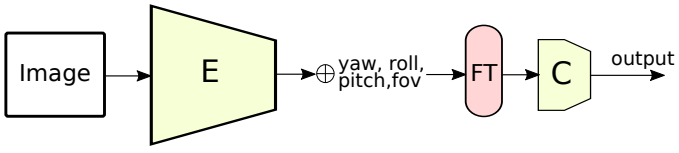

Fig. 11: Depiction of the experiment. E: Encoder, C: Binary Classification Layer(Output Layer). E & C were trained together. FT: Feature Transform MLP (Refer the Figure 6 for more details). The output from the encoder was concatenated with arbitrarily chosen yaw, pitch, roll, and fov, and it was input to the MLP to obtain the transformed feature, the output of the feature transform model was then input to the classification layer for binary output (gender classification).

From Table 14, it was observed that the classifier was able to classify the transformed feature vector with an overall accuracy of around 87% on intra-class evaluation. The overall accuracy of 93% was obtained with an end-to-end baseline model. Similarly, on UTKFace, the transformed feature vector obtained an overall accuracy of 92.3% whereas 93.67% overall accuracy was obtained by the end-to-end baseline classifier. Further on DiveFace, we obtained 97.42% gender classification accuracy for the feature transform model whereas 97.97% accuracy was obtained by the end-to-end model. It shows that the classification layer was able to distinguish the gender class from the transformed feature which is comparable with the end-to-end baseline model. This comparable performance indicates that the transformed feature vector has captured relevant gender cues, thereby supporting the consistency regularization techniques (CR-Feat Trans, CR-Img Aug+Feat Trans) in using semantic preserving augmentations for enhanced performance. We obtained similar observations with other architectures.

Table 14: Evaluation of Gender Classification with the transformed feature of FairFace(FF), UTKFace(UTKF), and DiveFace(DF) testsets was given as the input to the classification layer. The evaluation was done using the ViT-Baseline. A: End-to-end image protocol, B: Transformed feature protocol as shown in Figure 11.

Ethnicity	Black		East Asian		Indian		Latino Hispanic		Middle Eastern		Southeast Asian		White	
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F
A(FF)	89.86	84.94	91.38	93.66	93.09	94.49	93.44	94.71	96.31	94.46	90.75	94.27	94.56	93.57
B(FF)	82.38	76.01	79.2	92.126	88.526	85.954	87.97	90.02	92.78	92.24	83.33	89.45	88.493	89.538
A(UTKF)	99.14	88.68	-	-	95.57	92.44	-	-	-	-	91.07	94.05	95.79	92.62
B(UTKF)	99.57	84.38	-	-	94.24	93.019	-	-	-	-	83.43	98.43	92.71	92.62
A(DF)	96.58	98.06	-	-	97.14	99.17	-	-	-	-	-	-	98.76	98.11
B(DF)	96.31	97.165	-	-	95.163	99.47	-	-	-	-	-	-	98.68	97.71

9. Key Findings

In this section, we will discuss the key findings derived from our research.

- Systematic addition of image or feature level augmentations to the training data in a consistency regularization framework can enhance the feature representation for each demographic sub-group, hence, enhancing the performance as well as the fairness of the system.

- The combination of image-level and feature-level augmentations can further enhance the performance and fairness of image- and feature-level augmentations, individually. This is due to the enhancement of the data manifold by combining image-level perturbations with the variations in pitch, roll, yaw, and field of view (due to feature-level perturbations).

- Visual analysis of t-SNE plots of feature embeddings from our proposed techniques demonstrates the clear linear separability between the features from each demographic sub-group, resulting in enhanced performance and reduced bias.

- ANOVA-based hypothesis testing of the obtained results confirms that the bias reduction due to our proposed techniques is significant and not obtained by chance.

- Additionally, the ablation studies demonstrate the robustness of our proposed techniques to facial occlusions, the capability of feature transformer MLP in retaining the gender cues, the importance of the right set of augmentations, and the optimum weightage to the consistency and classification loss for the enhanced performance of our proposed techniques.

10. Conclusion and Future Work

Much of the existing machine learning-based fairness literature assumes the presence of protected attributes such as ethnicity and sex for bias mitigation. However, in practice, the collection of protected features, or their use for training or inference is often precluded due to privacy and regulation. This severely limits the applicability of traditional fairness research. Further, existing approaches to mitigating bias may offer a trade-off between fairness and classification performance.

We proposed a novel bias mitigation technique based on consistency-based regularization that can mitigate bias in the absence of demographic (protected attributes). It leverages the power of augmented views generated using image perturbation and *StyleNeRF* based multi-views. Thorough experimental validation supported by ablation studies confirms that carefully chosen augmentations in a consistency-based regularization setting can help improve the fairness as well as classification accuracy of the model by enhancing feature representation and reducing variance for the demographic sub-groups. Our proposed method obtained an overall reduction in the bias of about 30% over SOTA bias mitigation techniques and an improvement in classification accuracy of about 5% over the baseline. Thus, obtaining state-of-the-art performance.

Further, our proposed techniques using image- and feature-level augmentations have an advantage over random augmentations and AutoAugment, based on pre-trained policies, in significantly enhancing the underlying data manifold of the samples for rich feature representation learning. Thus enhancing performance as well as reducing bias at the same time. Further, our proposed technique can be applied to any downstream image recognition task. To further validate the proposed bias mitigation technique, additional rigorous empirical analysis across diverse problem domains is warranted. Future work should evaluate the efficacy of alternative biometric modalities (e.g. ocular biometrics (Krishnan et al., 2020, 2021))), heterogeneous data spectra (Near-Infrared (Krishnan et al., 2022)), and expanded application areas (such as Deepfake detection (Nadimpalli and Rattani, 2022)). Thorough intersectional analysis accounting for combinations of demographic sub-groups should also be undertaken. Insights from expanded evaluations will inform requisite algorithmic enhancements and modifications to optimize fairness and generalization performance.

11. Acknowledgements

This work is supported in part by National Science Foundation (NSF) award no. 2129173.

Fig. 12: t-SNE plot for projection of the feature embedding of the 2500 test samples from FairFace using the ViT Baseline. The green bounding box indicates correct classification, and the red indicates incorrect classification.[Best viewed in color.]

Fig. 13: t-SNE plot for projection of the feature embedding of the 2500 test samples from FairFace using ViT CR-Img Aug. The green bounding box indicates correct classification, and the red indicates incorrect classification.[Best viewed in color.]

References

I. Yag, A. Altan, Artificial intelligence-based robust hybrid algorithm design and implementation for real-time detection of plant diseases in agricultural environments, Biology 11 (2022). URL: https://www.mdpi.com/2079-7737/11/12/1732.

Y. Ozcelik, A. Altan, Classification of diabetic retinopathy by machine learning algorithm using entropy-based features, in: Cankaya International Congress on Scientific, 2023, pp. 1–11.

R. Kaur, D. Gabrijelcic, T. Klobucar, Artificial intelligence for cybersecurity: Literature review and future research directions, Inf. Fusion 97 (2023) 101804. URL: https://doi.org/10.1016/j.inffus.2023.101804. doi:10.1016/j.inffus.2023.101804.

A. Krishnan, A. Almadan, A. Rattani, Understanding fairness of gender classification algorithms across gender-race groups, in: 19th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2020, pp. 1028–1035. URL: https://doi.org/10.1109/ICMLA51294.2020.00167. doi:10.1109/ICMLA51294.2020.00167.

V. Albiero, K. K. S, K. Vangara, K. Zhang, M. C. King, K. W. Bowyer, Analysis of gender inequality in face recognition accuracy, in: IEEE WACV Workshops, IEEE, 2020, pp. 81–89. URL: https://doi.org/10.

Fig. 14: t-SNE plot for projection of the feature embedding of the 2500 test samples from FairFace using ViT CR-Feat Trans. The green bounding box indicates correct classification, and the red indicates incorrect classification.[Best viewed in color.]

Fig. 15: t-SNE plot for projection of the feature embedding of the 2500 test samples from FairFace using ViT CR-Img Aug+Feat Trans. The green bounding box indicates correct classification, and the red indicates incorrect classification.[Best viewed in color.]

1109/WACVW50321.2020.9096947. doi:10.1109/WACVW50321.2020.9096947.

H. Siddiqui, A. Rattani, K. Ricanek, T. J. Hill, An examination of bias of facial analysis based BMI prediction models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022, IEEE, 2022, pp. 2925–2934. URL: https://doi.org/10.1109/CVPRW56347.2022.00330. doi:10.1109/CVPRW56347.2022.00330.

A. V. Nadimpalli, A. Rattani, GBDF: gender balanced deepfake dataset towards fair deepfake detection, CoRR abs/2207.10246 (2022). URL: https://doi.org/10.48550/arXiv.2207.10246. doi:10.48550/arXiv.2207.10246. arXiv:2207.10246.

G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 34–42.

A. Almadan, A. Krishnan, A. Rattani, Bwcface: Open-set face recognition using body-worn camera, in: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 1036–1043. doi:10.1109/ICMLA51294.2020.00168.

K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, B. Li, Age group and gender estimation in the wild with deep ror architecture, IEEE Access 5 (2017) 22492–22503. URL: https://doi.org/10.1109/ACCESS.2017.2761849. doi:10.1109/ACCESS.2017.2761849.

S. Masood, S. Gupta, A. Wajid, S. Gupta, M. Ahmed, Prediction of human ethnicity from facial images using neural networks, in: S. C. Satapathy, V. Bhateja, K. S. Raju, B. Janakiramaiah (Eds.), Data Engineering and Intelligent Computing, Springer Singapore, Singapore, 2018, pp. 217–226.

N. R. Salim, N. Sankaranarayanan, U. Jayaraman, Gender classification beyond visible spectrum using shallow convolution neural network, in: 2021 IEEE Madras Section Conference (MASCON), 2021, pp. 1–7. doi:10.1109/MASCON51689.2021.9563425.

S. Kiruthika, V. Masilamani, Retinal image quality assessment using sharpness and connected components, in: B. Raman, S. Murala, A. S. Chowdhury, A. Dhall, P. Goyal (Eds.), Computer Vision and Image Processing - 6th International Conference, CVIP 2021, Rupnagar, India, December 3-5, 2021, Revised Selected Papers, Part II, volume 1568 of *Communications in Computer and Information Science*, Springer, 2021, pp. 181–191. URL: https://doi.org/10.1007/978-3-031-11349-9_16. doi:10.1007/978-3-031-11349-9_16.

P. Grother, G. W. Quinn, P. J. Phillips, Report on the evaluation of 2d still-image face recognition algorithms, in: NIST Report, 2011, p. 61.

B. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, A. K. Jain, Face recognition performance: Role of demographic information, IEEE Trans. Inf. Forensics Secur. 7 (2012) 1789–1801. doi:10.1109/TIFS.2012.2214212.

L. Best-Rowden, A. K. Jain, Longitudinal study of automatic face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 148–162. URL: https://doi.org/10.1109/TPAMI.2017.2652466. doi:10.1109/TPAMI.2017.2652466.

S. H. Abdurrahim, S. A. Samad, A. B. Huddin, Review on the effects of age, gender, and race demographics on automatic face recognition, Vis. Comput. 34 (2018) 1617–1630. URL: https://doi.org/10.1007/s00371-017-1428-z. doi:10.1007/s00371-017-1428-z.

I. D. Raji, J. Buolamwini, Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2019, pp. 429–435. URL: https://doi.org/10.1145/3306618.3314244. doi:10.1145/3306618.3314244.

R. Vera-Rodríguez, M. Blázquez, A. Morales, E. Gonzalez-Sosa, J. C. Neves, H. Proença, Facegenderid: Exploiting gender information in dcnns face recognition systems, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, Computer Vision Foundation / IEEE, 2019, pp. 2254–2260. doi:10.1109/CVPRW.2019.00278.

V. Muthukumar, Color-theoretic experiments to understand unequal gender classification accuracy from face images, in: IEEE CVPR Workshops, Computer Vision Foundation / IEEE, 2019, pp. 2286–2295. doi:10.1109/CVPRW.2019.00282.

K. Ricanek, T. Tesafaye, Morph: "a longitudinal image database of normal adult age-progression", in: IEEE Int'l Conf. on Automatic Face and Gesture Recognition (FG), Southampton, 2006, pp. 341–345.

A. Rekognition, Amazon Rekognition face api, 2022. URL: https://aws.amazon.com/rekognition/.

D. Vision, Deep Vision face api, 2022. URL: https://deepvisionai.in/.

FaceX, FaceX face api, 2022. URL: https://facex.io/.

M. A. C. Services, Microsoft Azure Cognitive Services face api, 2022. URL: https://azure.microsoft.com/en-in/pricing/details/cognitive-services/face-api/.

J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: ACM Conference on Fairness, Accountability, and Transparency, 2018, p. 77–91.

K. Kärkkäinen, J. Joo, Fairface: Face attribute dataset for balanced race, gender, and age, 2019. arXiv:1908.04913.

T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: P. A. Flach, T. D. Bie, N. Cristianini (Eds.), Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II, volume 7524 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 35–50. URL: https://doi.org/10.1007/978-3-642-33486-3_3. doi:10.1007/978-3-642-33486-3_3.

P. Majumdar, R. Singh, M. Vatsa, Attention aware debiasing for unbiased model prediction, in: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021, IEEE, 2021, pp. 4116–4124. URL: https://doi.org/10.1109/ICCVW54120.2021.00459. doi:10.1109/ICCVW54120.2021.00459.

B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: J. Furman, G. E. Marchant, H. Price, F. Rossi (Eds.), Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018, ACM, 2018, pp. 335–340. URL: https://doi.org/10.1145/3278721.3278779. doi:10.1145/3278721.3278779.

C. Chuang, Y. Mroueh, Fair mixup: Fairness via interpolation, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021, p. 15. URL: https://openreview.net/forum?id=DNl5s5BXeBn.

V. V. Ramaswamy, S. S. Y. Kim, O. Russakovsky, Fair attribute classification through latent space de-biasing, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 9301–9310. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Ramaswamy_Fair_Attribute_Classification_Through_Latent_Space_De-Biasing_CVPR_2021_paper.html. doi:10.1109/CVPR46437.2021.00918.

A. Das, A. Dantcheva, F. Brémond, Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach, in: L. Leal-Taixé, S. Roth (Eds.), Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I, volume 11129 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 573–585. URL: https://doi.org/10.1007/978-3-030-11009-3_35. doi:10.1007/978-3-030-11009-3_35.

X. Lin, S. Kim, J. Joo, Fairgrape: Fairness-aware gradient pruning method for face attribute classification, in: S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII, volume 13673 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 414–432. URL: https://doi.org/10.1007/978-3-031-19778-9_24. doi:10.1007/978-3-031-19778-9_24.

D. Zietlow, M. Lohaus, G. Balakrishnan, M. Kleindessner, F. Locatello, B. Schölkopf, C. Russell, Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 10400–10411. URL: https://doi.org/10.1109/CVPR52688.2022.01016. doi:10.1109/CVPR52688.2022.01016.

Y. Li, N. Vasconcelos, REPAIR: removing representation bias by dataset resampling, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 9572–9581. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Li_REPAIR_Removing_Representation_Bias_by_Dataset_Resampling_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00980.

J. Gu, L. Liu, P. Wang, C. Theobalt, Stylenerf: A style-based 3d aware generator for high-resolution image synthesis, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022, p. 24. URL: https://openreview.net/forum?id=iUuzzTMUw9K.

Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 13001–13008. URL: https://ojs.aaai.org/index.php/AAAI/article/view/7000.

E. D. Cubuk, B. Zoph, J. Shlens, Q. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020, p. 13. URL: https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

K. Karkkainen, J. Joo, Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1548–1558.

A. Morales, J. Fiérrez, R. Vera-Rodríguez, R. Tolosana, Sensitivenets: Learning agnostic representations with application to face images, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 2158–2164. URL: https://doi.org/10.1109/TPAMI.2020.3015420. doi:10.1109/TPAMI.2020.3015420.

Zhang, Zhifei, Song, Yang, H. Qi, Age progression/regression by conditional adversarial autoencoder, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, p. 9.

G. Balakrishnan, Y. Xiong, W. Xia, P. Perona, Towards causal benchmarking of bias in face analysis algorithms, in: A. Vedaldi, H. Bischof, T. Brox, J. Frahm (Eds.), 16th ECCV, volume 12363 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 547–563. URL: https://doi.org/10.1007/978-3-030-58523-5_32. doi:10.1007/978-3-030-58523-5_32.

K. K. Teru, A. Chakraborty, Towards reducing bias in gender classification, CoRR abs/1911.08556 (2019). URL: http://arxiv.org/abs/1911.08556. arXiv:1911.08556.

S. Ramachandran, A. Rattani, Deep generative views to mitigate gender classification bias across gender-race groups, CoRR abs/2208.08382 (2022). URL: https://doi.org/10.48550/arXiv.2208.08382. doi:10.48550/arXiv.2208.08382. arXiv:2208.08382.

S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, H. Byun, Fair contrastive learning for facial attribute classification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 10379–10388. URL: https://doi.org/10.1109/CVPR52688.2022.01014. doi:10.1109/CVPR52688.2022.01014.

C. Tan, Z. Gao, L. Wu, S. Li, S. Z. Li, Hyperspherical consistency regularization, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 7234–7245. URL: https://doi.org/10.1109/CVPR52688.2022.00710. doi:10.1109/CVPR52688.2022.00710.

E. Englesson, H. Azizpour, Consistency regularization can improve robustness to label noise, CoRR abs/2110.01242 (2021). URL: https://arxiv.org/abs/2110.01242. arXiv:2110.01242.

N. Saunshi, J. T. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. M. Kakade, A. Krishnamurthy, Understanding contrastive learning requires incorporating inductive biases, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 19250–19286. URL: https://proceedings.mlr.press/v162/saunshi22a.html.

A. V. Nadimpalli, N. Reddy, S. Ramachandran, A. Rattani, Harnessing unlabeled data to improve generalization of biometric gender and age classifiers, in: IEEE Symposium Series on Computational Intelligence, SSCI 2021, Orlando, FL, USA, December 5-7, 2021, IEEE, 2021, pp. 1–7. URL: https://doi.org/10.1109/SSCI50451.2021.9660182. doi:10.1109/SSCI50451.2021.9660182.

M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, N. Ballas, Self-supervised learning from images with a joint-embedding predictive architecture, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15619–15629.

C. Huang, H. Goh, J. Gu, J. Susskind, Mast: Masked augmentation subspace training for generalizable self-supervised priors, in: ICLR, 2023, pp. 1–10. URL: https://arxiv.org/abs/2303.03679.

B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, Nerf: representing scenes as neural radiance fields for view synthesis, Commun. ACM 65 (2022) 99–106. URL: https://doi.org/10.1145/3503250. doi:10.1145/3503250.

T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 4401–4410. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00453.

T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 4217–4228. URL: https://doi.org/10.1109/TPAMI.2020.2970919. doi:10.1109/TPAMI.2020.2970919.

M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Con-

ference Track Proceedings, OpenReview.net, 2017, p. 17. URL: https://openreview.net/forum?id=Hk4_qw5xe.

T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018, p. 26. URL: https://openreview.net/forum?id=B1QRgziT-.

K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 2261–2269. URL: https://doi.org/10.1109/CVPR.2017.243. doi:10.1109/CVPR.2017.243.

M. Tan, Q. V. Le, Efficientnetv2: Smaller models and faster training, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10096–10106. URL: http://proceedings.mlr.press/v139/tan21a.html.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021, p. 22. URL: https://openreview.net/forum?id=YicbFdNTTy.

R. Wightman, Pytorch image models, https://github.com/rwightman/pytorch-image-models, 2019. doi:10.5281/zenodo.4414861.

M. Sensoy, L. M. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 3183–3193. URL: https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent - A new approach to self-supervised learning, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020, p. 35. URL: https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

B. Fuglede, F. Topsøe, Jensen-shannon divergence and hilbert space embedding, in: Proceedings of the 2004 IEEE International Symposium on Information Theory, ISIT 2004, Chicago Downtown Marriott, Chicago, Illinois, USA, June 27 - July 2, 2004, IEEE, 2004, p. 31. URL: https://doi.org/10.1109/ISIT.2004.1365067. doi:10.1109/ISIT.2004.1365067.

J. Jeong, S. Lee, J. Kim, N. Kwak, Consistency-based semi-supervised learning for object detection, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 10758–10767. URL: https://proceedings.neurips.cc/paper/2019/hash/d0f4dae80c3d0277922f8371d5827292-Abstract.html.

P. Izmailov, D. Podoprikhin, T. Garipov, D. P. Vetrov, A. G. Wilson, Averaging weights leads to wider optima and better generalization, in: A. Globerson, R. Silva (Eds.), Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018, AUAI Press, 2018, pp. 876–885. URL: http://auai.org/uai2018/proceedings/papers/313.pdf.

R. Singh, P. Majumdar, S. Mittal, M. Vatsa, Anatomizing bias in facial analysis, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 12351–12358. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21500.

W. Huang, M. Yi, X. Zhao, Towards the generalization of contrastive self-supervised learning, CoRR abs/2111.00743 (2021). URL: https://arxiv.org/abs/2111.00743. arXiv:2111.00743.

L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

E. R. Girden, ANOVA: Repeated measures, 84, Sage, 1992.

D. E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758. URL: https://dl.acm.org/doi/10.5555/1577069.1755843. doi:10.5555/1577069.1755843.

E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation strategies from data, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 113–123. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00020.

A. Krishnan, A. Almadan, A. Rattani, Probing fairness of mobile ocular biometrics methods across gender on VISOB 2.0 dataset, in: A. D. Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, R. Vezzani (Eds.), Pattern Recognition. ICPR International Workshops and Challenges - Virtual Event, January 10-15, 2021, Proceedings, Part VIII, volume 12668 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 229–243. URL: https://doi.org/10.1007/978-3-030-68793-9_16. doi:10.1007/978-3-030-68793-9_16.

A. Krishnan, A. Almadan, A. Rattani, Investigating fairness of ocular biometrics among young, middle-aged, and older adults, in: 2021 International Carnahan Conference on Security Technology, ICCST 2021, Hatfield, United Kingdom, October 11-15, 2021, IEEE, 2021, pp. 1–7. URL: https://doi.org/10.1109/ICCST49569.2021.9717383. doi:10.1109/ICCST49569.2021.9717383.

A. Krishnan, B. Neas, A. Rattani, Is facial recognition biased at near-infrared spectrum as well?, CoRR abs/2211.00129 (2022). URL: https://doi.org/10.48550/arXiv.2211.00129. doi:10.48550/arXiv.2211.00129. arXiv:2211.00129.