Exploring Disparity-Accuracy Trade-offs in Face Recognition Systems: The Role of Datasets, Architectures, and Loss Functions

Siddharth Jaiswal[1], **Sagnik Basu**[1], **Sandipan Sikdar**[2], **Animesh Mukherjee**[1]
[1]Indian Institute of Technology, Kharagpur, India
[2]L3S Research Centre, Leibniz University, Hannover, Germany
{siddsjaiswal@kgpian, basusagnik99.24@kgpian, animeshm@cse}.iitkgp.ac.in, sandipan.sikdar@l3s.de

Abstract

Automated Face Recognition Systems (FRSs), developed using deep learning models, are deployed worldwide for identity verification and facial attribute analysis. The performance of these models is determined by a complex interdependence among the model architecture, optimization/loss function and datasets. Although FRSs have surpassed human-level accuracy, they continue to be disparate against certain demographics. Due to the ubiquity of applications, it is extremely important to understand the impact of the three components– model architecture, loss function and face image dataset on the accuracy-disparity trade-off to design better, unbiased platforms. In this work, we perform an in-depth analysis of three FRSs for the task of gender prediction, with various architectural modifications resulting in ten deep-learning models coupled with four loss functions and benchmark them on seven face datasets across 266 evaluation configurations. Our results show that all three components have an individual as well as a combined impact on both accuracy and disparity. We identify that datasets have an inherent property that causes them to perform similarly across models, independent of the choice of loss functions. Moreover, the choice of dataset determines the model's perceived bias– the same model reports bias in opposite directions for three gender-balanced datasets of "in-the-wild" face images of popular individuals. The facial embeddings show that the models are unable to generalize a uniform definition of what constitutes a "female face" as opposed to a "male face", due to dataset diversity. We provide recommendations to model developers on using our study as a blueprint for development and subsequent deployment.

Introduction

The performance of deep learning models is usually determined by a complex interdependency among the model architecture, the objective function being optimized for, and the data, as shown in Figure 1. For example, even a very deep architecture coupled with the optimal loss function cannot perform well if the data is not representative of the true population. Similarly, a model may not perform well on a perfectly sampled dataset if the architecture is too shallow or the loss function is too simple. This is especially true for face recognition models, which have become highly commonplace in society with the development and democratization of deep learning and rival human-level accuracy for a number of tasks (O'Toole and Castillo 2021). These systems ingest very diverse data in the form of face images and attempt to find a general pattern to classify attributes like the perceived gender of the person (Karkkainen and Joo 2021). Use cases for such attribute classification range from targeted advertising (Mennecke and Peters 2013) and customer analytics (Reuters 2010) to surveillance (Hitoshi 2016) and content moderation (Ning et al. 2022). Despite the high accuracy of these classification models, disparities have been reported on commercial (Buolamwini and Gebru 2018; Jaiswal et al. 2022) and open-source FRSs (Jaiswal et al. 2024) against minority demographics, severely impacting access to services (John Dunne 2019), leading to exclusion and unfair treatment. Such disparities are an outcome of the way the dataset, the model architecture and the objective function together behave.

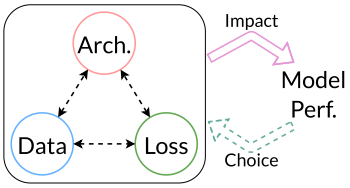


Figure 1: Interdependence between the architecture, data and loss function that determine the performance of a model, which in turn informs the choice of components.

An FRS model can have multiple types of vision backbones viz. CNNs and transformers, use multiple types of loss functions (Wang et al. 2018; Deng et al. 2019) and be trained and tested on various datasets (Raji et al. 2020; Karkkainen and Joo 2021). These components can be combined in an exponential number of ways, with each having an impact on the final accuracy and, more importantly, the disparity in performance amongst classes, especially for minority demographics. In this study, we attempt to unravel this complex relationship and its impact on gender bias in FRSs for the task of gender prediction through an in-depth, large-scale audit study involving three FRSs. The FRSs have two types of vision backbones and four types of loss functions in various combinations and are evaluated over seven benchmark face datasets with more than 86k images in total.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Research questions: It is usually seen that accuracy and bias are often competing objectives (Janssen and Sadowski 2021). Thus, before deploying any FRS for sensitive tasks like gender prediction, it is absolutely imperative to evaluate how the three components – data, model architecture, objective optimised – impact the accuracy vs disparity divide. This brings us to our first research question– **RQ1** *Do the individual components, viz. model architecture, loss function and dataset, impact the disparity along with the accuracy?*

Most studies in the literature show that a majority of FRSs are biased against females. In this work, we attempt to investigate whether this observation is generalizable across datasets, architectures and loss functions or if there exists a more nuanced relationship that determines the direction of disparity. We formalize this into our second research question– **RQ2** *Do all datasets report equally disparate performance against females, irrespective of the model architecture and loss function?* Through this question, we will be able to understand not only the extent of disparity but also the group that is more often discriminated against.

Disparities in FRSs exist because the model performs better for one sub-group over the others. Thus, independent of the dataset, it is important to quantify how the accuracies for each gender group change for different choices of loss functions and architectures. This brings us to our final research question – **RQ3** *What is the relationship between the change in accuracy for males vs. females for different models and their architectures?* Through this question, we seek to understand which models and loss functions prefer which gender group.

Our key observations. In line with the research questions highlighted above, we make the following key observations from our in-depth analysis.

- Our analysis of the accuracy vs disparity trade-off shows that model architecture, loss function and dataset impact not only the accuracy but also the observed disparity in an FRS, irrespective of the choice of backbone– CNN or transformer. Our evaluation shows that independent of other model parameters, in the accuracy vs disparity chart, some datasets like CFD, LFW and CelebSET always cluster around high accuracy with low disparity, while UTKFace and Fairface report low disparity but at the expense of low accuracy and CelebA reports high disparity & low accuracy (answers RQ1).
- We observe that not all FRSs are equally disparate against females in all datasets. In fact, the CNN model LibfaceID reports lower accuracies for males on two standard balanced datasets– CelebSET & CelebA. With increasing model complexity, not only does the overall disparity reduce, but so does the effect of the loss function on the disparity (answers RQ2).
- We note that the two gender groups have different levels of sensitivity toward each model and architecture. For example, the accuracy simultaneously improves for females and degrades for males when residual connections are used in the LibfaceID model. Similarly, using just the CLS embedding in the vision transformer model elicits both positive and negative changes for the males' accuracy and primarily a positive change in accuracy for the females. (answers RQ3).

Related Work

We briefly discuss the existing literature for FRSs, their biases against minority or marginalized groups and existing studies focusing on the interplay among datasets, model architectures and loss functions.

Face recognition systems. These platforms are primarily based on deep learning models and target various tasks like face identification (Yang, Kriegman, and Ahuja 2002) and downstream analysis (Levi and Hassner 2015) like gender, age, emotion and face matching (Taigman et al. 2014). The models can be commercial (Amazon 2021; Face++ 2021; Microsoft 2021) or open-source (Taigman et al. 2014; Deng et al. 2019; Wang et al. 2018). SOTA transformer models (Chen et al. 2016; Dan et al. 2023; Zhong and Deng 2021) are also used for similar tasks, with human-level accuracy. Reduction in deployment costs and optimized models has allowed wide-scale adoption of these platforms at city-scale levels (Livemint 2021). Gender prediction on FRSs comes across as an unobvious task, but FRSs are rampantly used for this task in domains like targeted advertising (Mennecke and Peters 2013), customer analytics (Reuters 2010) and surveillance (Hitoshi 2016). This raises an important concern about the biases that can manifest while using FRSs as gender predictors. Since this is becoming highly normative, such systems need to be monitored, and the untoward biases need to be addressed.

Biases in FRSs. Multiple studies in prior literature have exposed social biases based on sensitive features like gender/race, etc., for the task of gender prediction in traditional CNN-based commercial (Buolamwini and Gebru 2018; Jaiswal et al. 2022; Raji et al. 2020) and open-source (Jaiswal et al. 2024) FRSs. While researchers have studied biases in vision transformers (Liu et al. 2022; Brinkmann, Swoboda, and Bartelt 2023), these have been for a general set of tasks rather than the specific task of gender prediction from face images. As these vision transformers are highly complex models, and used for a wide variety of face recognition tasks, it is important to study and mitigate the biases therein.

Interaction among the three main components of deep learning. Existing literature has studied the interplay between the structure of data and loss functions (d'Ascoli et al. 2021), but have not focused on the impact towards the overall bias in the system. Davidian et al. (2024) study the impact of dataset size and imbalance in CNNs for healthcare while Cherepanova et al. (2023) perform a similar study for FRSs. Cabannes et al. (2023) study the interplay amongst the choices for data augmentation, network architecture and training algorithm, but do not attempt to study the impact on the bias of a model. In this work, we not only study the impact of the choice of datasets, model architecture, and loss functions on the accuracy but also on the disparity in the model, a far more societally impactful problem.

Relevance to web & social media. In this work, we study the fairness and bias implications of FRSs, which are deployed not only in the physical world but also on digital platforms. For example, FRSs have been deployed at a

Name	# Images			Usage
	Male	Female	Total	
Adience	8,107	9,356	17,463	Train/FT/Test
CFD	680	761	1,441	Test
CelebSET	800	800	1,600	Test
FARFace	931	931	1,862	Test
LFW	2,966	2,966	5,932	Test
Fairface	5,788	5,160	10,948	Test
UTKFace	12,390	11,314	23,703	Test
CelebA	20,500	20,500	41,000	Test

Table 1: Summary characteristics of the benchmark datasets.

large scale on both Google Photos (Google 2024) and Facebook (shut down in 2021 (Facebook 2021) and revived in 2024 (Facebook 2024)) for tagging and scammer detection. These services have also shown highly discriminatory biases (Zhang 2015; BBC 2021a,b) indicating the need to audit the models, the datasets and their frameworks in the context of social media applications. On the other hand, the social media research community has shown growing interest in this domain– both in the use of these FRSs for analysing social media platforms (Chakraborty et al. 2017; Vikatos et al. 2017; Messias, Vikatos, and Benevenuto 2017; Pang et al. 2015) as well as studying fairness concerns in image tagging (Kyriakou et al. 2019; Barlas et al. 2019) and attribute analysis (Jaiswal et al. 2022; Jung et al. 2018) applications. We argue that FRSs are highly prevalent in both physical and social networks, and it is important to study them end-to-end to ensure their deployment is indiscriminate and within regulatory frameworks. With FRSs increasingly being deployed in web and social media applications, we believe our study to be both highly relevant and timely.

Datasets & FRS Models

In this section, we present a brief overview of the datasets and open-source FRS models that we audit in this study.

Datasets

We consider a range of datasets of different sizes with significant variety in gender, race and geographic distributions. We use the Adience (Eidinger, Enbar, and Hassner 2014) dataset, with more than 17k images, to train the CNN model and to fine-tune the pre-trained transformer models. We use five diverse benchmark datasets for evaluation–

- Chicago Face Database (CFD) (Ma, Correll, and Wittenbrink 2015), curated from images volunteered by citizens of the USA, belonging to four different races– White, Black, Asian and Latinx. CFD has two extension sets– CFD-MR (Ma, Kantner, and Wittenbrink 2020), with 88 images and CFD-India (Lakshmi et al. 2021) with 146 images. We combine them with the original dataset and refer to the superset as CFD throughout.
- CelebSET (Raji et al. 2020), curated from IMDB images of Hollywood celebrities belonging to the White and Black races.
- FARFace (Jaiswal et al. 2024), curated from ESPN-CricInfo images of cricketers belonging to the Global North and Global South ($> 50\%$). The authors do not annotate with the race labels.
- LFW (Huang et al. 2007), curated from the web with names and gender of each individual. The authors do not annotate the faces with the race labels.
- Fairface (Karkkainen and Joo 2021), curated from the YFCC-100M Flickr dataset belonging to the following races – White, Black, Latinx, Indian, Southeast Asian, East Asian and Middle Eastern.
- UTKFace (Zhang, Song, and Qi 2017), curated from the Morph dataset and the CACD dataset. This dataset has the following races – White, Black, Asian, Indian and Others (including Hispanic, Latinx, and Middle Eastern).
- CelebA (Liu et al. 2015), curated from celebrity face images around the world collected from the internet. The authors provide fine-grained annotation for gender and other facial attributes, but not race.

All the datasets mentioned above are annotated with the binary gender – male & female. The labels in CFD were self-annotated by the individuals whose photos are part of the dataset, whereas for all other datasets, the labels were annotated by the dataset curators. Consequently, our gender prediction models are also trained to predict a binary gender label. However, we do acknowledge that the notion of gender is fluid, and all predictions should be interpreted as the *perceived* gender. A more detailed discussion is presented in the section on limitations. All datasets except CFD are curated from online sources and have played an important role in the training and deployment of FRSs in the physical world as well as social media platforms (Taigman et al. 2014), thus further reiterating their close association with social media platforms. For example, Google uses FRSs to perform person re-identification for image tagging (Google 2024), and Facebook is using FRSs to identify scammers using celebrity images on their platform (Facebook 2024). The dataset sizes, along with the distribution between the two binary gender labels, are noted in Table 1.

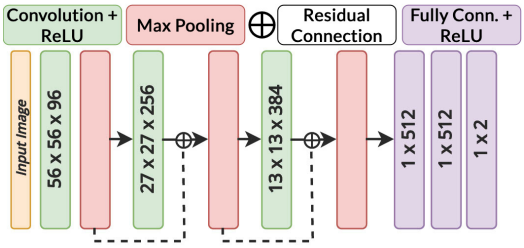
FRS Models

We audit two types of vision backbones – CNNs and vision transformers, by evaluating three types of face recognition models – LibfaceID (Levi and Hassner 2015), ViT-Face (Zhong and Deng 2021) and an instruction-tuned vision language model, InstructBLIP (Dai et al. 2023). The models differ in terms of architectural complexity, size of training data and parameter size, i.e., 8M to 7B. We present a brief description of the models as follows.

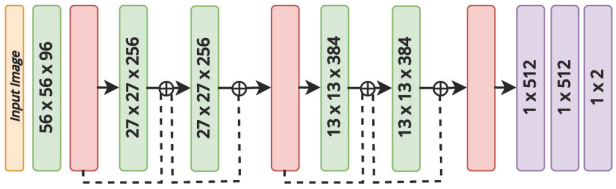
• **Libfaceid**[1] **(Levi and Hassner 2015)** – We use the implementation based on a CNN model proposed by Levi and Hassner (2015). The model has a simple CNN backbone with three `Conv` and three `FC` layers. We train this model from scratch on the Adience (Eidinger, Enbar, and Hassner 2014) dataset. The training hyperparameters are in Table 4, and implementation details[2] are in the Appendix.

[1]https://github.com/richmondu/libfaceid

[2]https://github.com/Sidd0602/ExploringTradeoffsFRS



(a) Residual connections with the base architecture



(b) Residual connections with the deeper architecture

Figure 2: Schematics for the Libfaceid (Levi and Hassner 2015) model modified only with residual connections (2a), and with extra layers and residual connections (2b).

• **ViT-Face (Zhong and Deng 2021)** – This model is based on the classic vision transformer (Dosovitskiy et al. 2020) backbone. The model is deeper and more complex than the CNN architecture and learns image representations by dividing it into patches and linearizing them. The model has an input patch size of 12 with a stride size of 8. The base model is simply trained to generate image embeddings. Hence, we add two linear layers to the final embedding layer to perform gender prediction. The model is fine-tuned on the Adience dataset. We refer to the model as ViT throughout.

• **InstructBLIP (Dai et al. 2023)** – InstructBLIP is an extended version of BLIP-2 (Li et al. 2023) vision-language model. It uses a frozen image encoder (ViT-g/14 (Fang et al. 2023)) and a frozen large language model (LLM) (Vicuna-7B (Chiang et al. 2023)) along with a QFormer (Li et al. 2023) which connects the two frozen models and is pre-trained using task-specific instructions. Similar to the vision transformer above, we add two additional linear layers to the QFormer's output and fine-tune with the Adience dataset.

Experimental Design

We now describe our experimental setup for all the FRS models, using different loss functions along with the architectural changes to evaluate the accuracy and disparity for the task of gender prediction.

LibfaceID Settings

We first describe the architectural changes, followed by the loss functions that we use to train this model.

Architectural changes The base LibfaceID CNN model is shallow with three convolution layers. We refer to this model as LBFC$_B$. Next, we extend this architecture with three changes, resulting in three new models– (a) residual connections between the `Conv` layers (see Fig. 2a) to carry forward information from previous layers (referred to as LBFC$_{B+R}$), (b) two extra `Conv` layers to increase the model depth (referred to as LBFC$_{B+2}$) and, (c) residual connections in the deeper model (see Fig. 2b) (referred to as LBFC$_{B+2+R}$).

For both models in Figure 2, we also experiment with an additional setup where the residual connections are weighted. Details and results for this setup are in the Appendix.

Loss functions The default loss that we train the model with is the Cross-Entropy loss, referred to as "CE" in all our experiments. This is a commonly used classification loss, deployed for its simplicity, especially for binary classification tasks. In this work, we also train the CNN model with two additional SOTA loss functions, explicitly introduced for face recognition models, with the aim of improving inter-class separability– (a) *Triplet* loss (Schroff, Kalenichenko, and Philbin 2015) (T), a contrastive loss function which requires a positive and negative example for every input (known as the anchor). We set the positive example as another image of the same gender and the negative example as an image from the opposite gender. (b) *ArcFace* loss (Deng et al. 2019) (A), an angular margin loss function that attempts to increase inter-class separability and intra-class compactness. These loss functions are used along with the CE loss for all experimental setups in isolation, as well as in combination– "T", "A" and "A+T". The weight for each loss function is decided using grid search (details in the Appendix).

ViT-Face Settings

We do not modify the core architecture of ViT-Face; instead, we choose between different embeddings for fine-tuning the base model.

Choice of embeddings The vision transformer generates embeddings for each of the input patches and an extra embedding for the CLS token at the end of the pipeline. We choose the following embeddings to fine-tune our model– (a) the embedding generated from the CLS token, referred to as ViT$_{CLS}$. (b) the embedding generated by taking a mean of all patch embeddings– ViT$_M$ and, (c) the embedding generated by concatenating the previous two– ViT$_{CLS+M}$.

Loss functions Similar to the CNN model, we use CE as our default fine-tuning loss function here as well. Next, we use the T and A loss functions described above. We also experiment with the SOTA *CosFace* (Wang et al. 2018) (Cos) loss function, designed to use a cosine margin-based function to increase the inter-class distance between different faces. In our experiments, we note that the combination of these losses works well only with ViT$_M$ and ViT$_{CLS+M}$. Thus, we use the loss functions individually for ViT$_{CLS}$ and in combination for the other two models– "CE", "T", "A", "Cos", "A+T", "A+T+Cos". It must be noted that the CE loss is always a part of the pipeline, along with other loss

functions. The weights for each combination are shared in the Appendix.

InstructBLIP Settings

Similar to the vision transformer, we do not modify the core architecture; instead, we fine-tune it with two linear layers. As the vision-language model works best with textual prompts, we provide the following string as input with each query image– *"What is the gender of this person?"*. The choice of loss functions is the same as in ViT$_M$ and ViT$_{CLS+M}$. Further details are present in the Appendix.

Evaluation Metrics

The metric we use to evaluate the models is accuracy– Acc and the fairness metric is disparity between the accuracies for each gender group $Acc_M - Acc_F$, referred to as gender disparity. More formally, we define the two metrics as follows– Acc = $\frac{C_M+C_F}{T_M+T_F}$ where C_G: $G \in \{M \text{ (male)}, F \text{ (female)}\}$ is the number of correct predictions for each gender group, and T_G is the number of total data points for each gender group. Similarly, the disparity is formally defined as $Acc_M - Acc_F$ = $\frac{C_M}{T_M}$ - $\frac{C_F}{T_F}$.

Overall, we consider *ten* models, *seven* datasets and *four* loss functions in various combinations, leading to a total of **266** evaluation configurations. The hyperparameters used in our experiments are noted in the Appendix.

Results & Observations

We now present the results and associated takeaways from our in-depth audit study. Here we only discuss the results of the test benchmark datasets; the results on the test subset for Adience (which was used for training/fine-tuning the models) are present in the Appendix. We reiterate here that the CNN model has been trained from scratch, whereas the transformer and VLM have been fine-tuned.

Accuracy vs. Disparity (RQ1)

Accuracy and disparity are often competing objectives in a model's training procedure. Here, we attempt to get a detailed understanding of how the architecture, loss function and dataset combine to predict the gender of a person. In Figure 3, we present the scatter plots between the accuracy and absolute disparity for each dataset, segregated by the model architecture type.

Results for LibfaceID The LibfaceID model is a small CNN trained on the Adience dataset for the task of gender prediction. (a) In Figs. 3a–d, we see that the accuracy is always between 60-95% for all architectures, datasets and loss functions, and the highest absolute disparity is 40% and below. (b) Next, we see that, across all architectures, Triplet loss and ArcFace loss compete for the highest disparity. (c) Some of the datasets always report low absolute disparities, whereas others always report high accuracy – FARFace and CelebA report the highest disparity (except in Fig. 3d), whereas UTKFace has the lowest disparity. Similarly, CFD has the highest accuracy, and CelebA has the lowest accuracy. (d) Finally, we see a strong clustering tendency amongst the datasets, especially when residual connections are used, with three primary forms of clusters – (i) high accuracy, low disparity, (ii) low accuracy, low disparity and (iii) low accuracy, high disparity. The last type, reported primarily for the CelebA dataset, is the most adversarial input for the model and exposes its shortcomings.

Results for ViT The complex attention-based architecture of the pre-trained vision transformer is expected to improve the accuracy, but does not provide a guarantee on the reduction of disparity. (a) We see that the accuracy is in a wide range – between 55–90% when only CLS embeddings are used as opposed to 80–98% when the mean of all image patch embeddings is also used. The range of absolute disparity is largest in ViT_M and smallest in ViT_{CLS+M}. (b) Here, we note that the architecture plays a stronger role than the loss functions in determining the accuracy and disparity, with no clear trend across different loss functions. (c) We also note that neither CFD reports the highest accuracy nor UTKFace consistently reports the lowest disparity, thus showing how sensitive each dataset is to the architecture. (d) Finally, we see that the clustering tendency is even higher for this model, especially for ViT_{CLS+M}.

Results for InstructBLIP The VLM includes a frozen language model and a Q-Former block that drives the vision encoder to focus on the salient parts of the image, thereby improving the accuracy of the task even further. (a) It is clear that this is the best-performing model with a high avg. accuracy – $\geq 90\%$ and low avg. disparity – $\leq 12\%$. (b) Similar to LibfaceID, using only triplet loss reduces the accuracy for each dataset. (c) FARFace reports the largest disparity (considering all loss functions), and Fairface reports the lowest accuracy. CFD and CelebSET are the best-performing datasets. (d) The clustering tendency is the highest, with a clear separation into the three types of clusters observed earlier.

Major takeaways We now list the major takeaways from the results in Figure 3.

• *Model architectural complexity impacts performance*– As expected, LibfaceID is the worst-performing model, and InstructBLIP is the best-performing model in terms of both high accuracy and low disparity.

• *Choice of loss function impacts performance*– Triplet loss has been used previously by researchers (Jaiswal et al. 2024) to improve the accuracy of FRSs for gender classification, but our experimental results show that auditing this loss function for different architectures and datasets exposes the shortcomings of this optimization. Thus, for a fixed dataset and architecture, the choice of loss function has a significant impact on the performance (Fig. 1).

• *Choice of dataset impacts performance*– FARFace and CelebA most often report the largest disparity, whereas CFD reports the highest accuracy. This is strongly correlated to the inherent nature of the datasets– CFD is a highly standardized benchmark dataset with every person having the same pose, angle, lighting, and clothing, and FARFace is a new, recently released dataset that has a majority of the

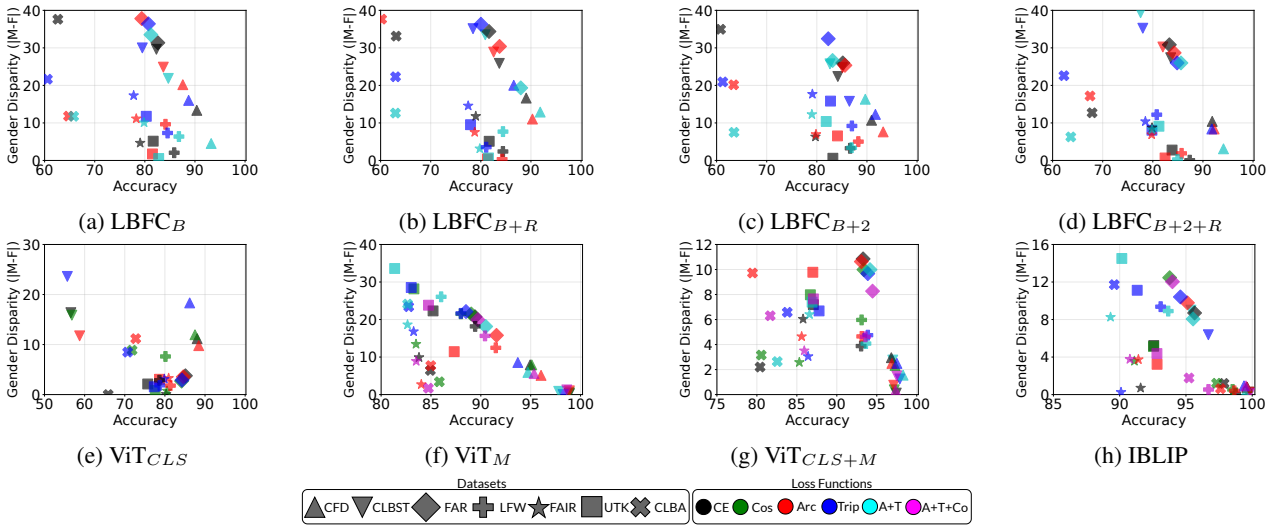


Figure 3: Accuracy vs. absolute disparity for different architectures across all datasets and loss functions. CFD always reports high accuracy, whereas FARFace and CelebA primarily report high disparities. On every architecture, each dataset has a similar performance for all loss functions that it is evaluated with. The shapes refer to the datasets, with colors referring to the various loss functions. Each combination of dataset and loss function refers to one experimental setup. The model complexity, choice of loss function and choice of dataset impact both the accuracy and the disparity.

faces from the Global South, thus providing a new, adversarial challenge, even to large pre-trained models.

For a given architecture, the performances of a dataset for the different loss functions are very close to each other, considering both accuracy and disparity, thus creating a cluster. For example, see Fig. 3g, where each dataset's symbols for all losses are close to each other. A similar observation is true for the other architectures as well. The inherent property of the test dataset seems to strongly influence the fairness of the inference results.

Thus, we see that the three components of a model indeed have a significant impact on the model's performance, impacting both accuracy and disparity, answering RQ1.

Debiased algorithms: To complete the analysis, we investigate the effect of the loss functions discussed earlier on known debiased FRSs. We choose the popular model provided by Karkkainen and Joo (2021) for this purpose. The authors pretrained a ResNet-34 network with the Fairface dataset to ensure that it is debiased. We call this the vanilla version of the model. We then fine-tune the pretrained model with all the loss functions on the Adience dataset. From Table 2, we see that the choice of loss functions positively impact both accuracy and disparity, with our choice of fine-tuning setup *reducing the disparity on 5 out of the 6 datasets* (we ignore Fairface as the model is trained on the same set and has risk of data overlap).

Direction of Disparity (RQ2)

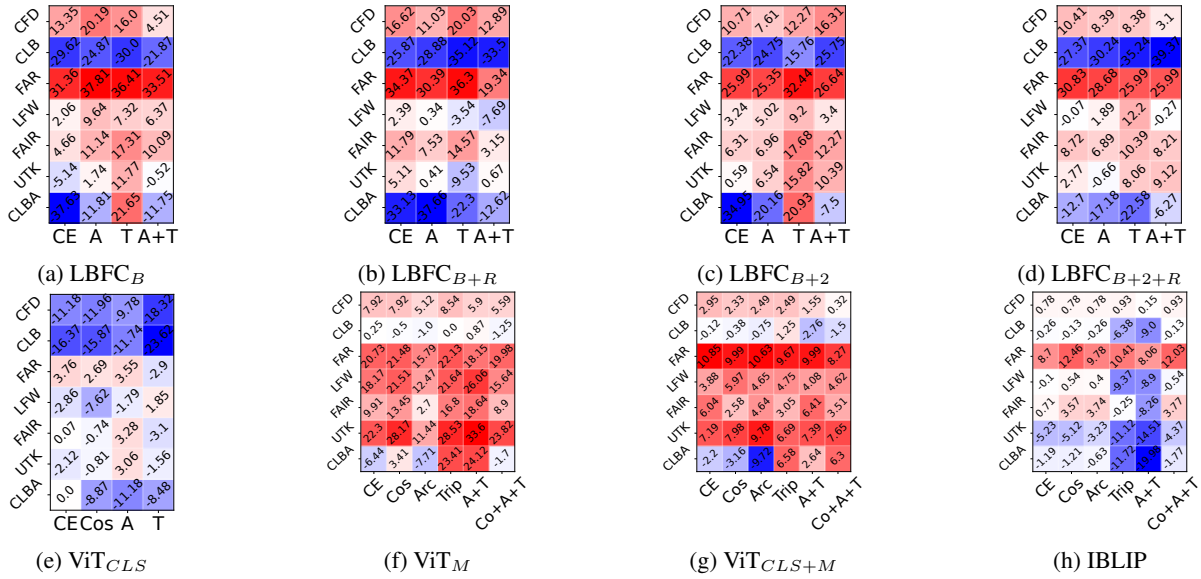
Disparity in an FRS's performance, especially against social minorities, leads to denial of services (Best 2020; John Dunne 2019) and unfair treatment. Multiple studies in literature show that FRSs are heavily biased against females. In Figure 4, we look at heatmaps of disparity which reflect the

Dataset	Accuracy (↑)		Abs. Disparity (↓)	
	Ours	Vanilla	Ours	Vanilla
CFD	**98.33 (A)**	97.29	**0.98 (Cos)**	4.57
CLBST	99.56 (Co+A+T)	**99.69**	0.62 (Co+A+T)	**0.13**
FAR	**93.07 (Co+A+T)**	91.68	**12.78 (Co+A+T)**	16.43
LFW	**97.76 (CE)**	96.94	**0.34 (Cos)**	3.74
UTK	93.70 (Co+A+T)	**94.73**	**0.03 (T)**	0.20
CLBA	90.14 (Co+A+T)	**90.35**	**0.79 (Co+A+T)**	6.78

Table 2: Accuracy & disparity of the vanilla debiased model (Karkkainen and Joo 2021) and the same model fine-tuned with our choice of loss functions. We only show the model+loss with highest accuracy & lowest disparity.

group that has a higher accuracy– red indicates males have a higher accuracy than females, and blue indicates vice versa; the intensity of each color reflects the extent of this disparity.

Results for LibfaceID The heatmaps for the LibfaceID model generalize across architectural changes. From Figures 4a–d, we see (a) All datasets report disparity, independent of the choice of architecture and loss function. Thus, the model is biased throughout and has a subpar performance in reducing disparity. (b) From Figs. 3a–d, we already know FARFace and CelebA have the highest disparity. From the heatmaps, we see that this disparity in FARFace is *always* against females, as previously observed (Jaiswal et al. 2024) and in CelebA is primarily against males. UTKFace has the lowest average disparity. (c) The choice of loss function impacts the disparity. Triplet loss has the highest disparity across datasets, independent of the architecture (except CelebA, where CE is the primary contributor to the high-



(a) LBFC$_B$ (b) LBFC$_{B+R}$ (c) LBFC$_{B+2}$ (d) LBFC$_{B+2+R}$

(e) ViT$_{CLS}$ (f) ViT$_M$ (g) ViT$_{CLS+M}$ (h) IBLIP

Figure 4: Heatmaps indicating the extent and the direction of disparity between the two genders for different architectures across all datasets and loss functions. CelebSET is always disparate against males, and FARFace is always disparate against females, independent of all other factors, despite being balanced datasets. The color codes are as follows– red indicates higher accuracy for males, and blue indicates higher accuracy for females. The intensity of the color signals the magnitude of the gender disparity.

est disparity). (d) Interestingly, we see that even though both CelebSET and FARFace are balanced datasets, their disparities are in exact opposite directions. CelebA, another dataset composed of celebrity faces, also reports disparities primarily against males. This shows that the choice of dataset is non-trivial, especially for evaluating FRSs, as it can lead to conflicting inferences.

Results for ViT In Figures 4e–g, the heatmaps give a completely different picture as compared to LibfaceID. We see that (a) The average disparities reduce significantly compared to the simpler CNN model. (b) In contrast to the CNN model, CelebSET and Fairface now report the lowest disparity instead of UTKFace. Using only the CLS embeddings leads to a bias against males, whereas using the mean of the patch embeddings reverts the bias to be against females (except the CelebSET & CelebA dataset). As noted previously, FARFace has the highest disparity against females. (c) The only observable pattern for the loss functions is triplet loss always reporting disparity against males in Fig. 4e (except for LFW) and against females in Figs. 4f-g. (d) Here as well, both CelebSET and FARFace report biases in opposite directions, especially when all embeddings are used – Fig. 4g.

Results for InstructBLIP Finally, we take a look at the heatmap for the VLM in Fig. 4h. (a) We observe the lowest average disparity for this model. (b) The lowest disparities are observed for both CFD and CelebSET and the highest for FARFace and UTKFace. (c) Similar to ViT, there is no generalizable pattern amongst the loss functions. (d) Instead of CelebSET, the highest disparity against males is reported for CelebA (ArcFace + triplet loss). Interestingly, the average disparity against females is always higher than for males.

Major takeaways We now list the major takeaways from the results in Figure 4.

• *Model size determines disparity*– As we increase the model complexity (parameters, layers, etc.), the disparity reduces. Thus, larger models may be the key to reducing disparity.

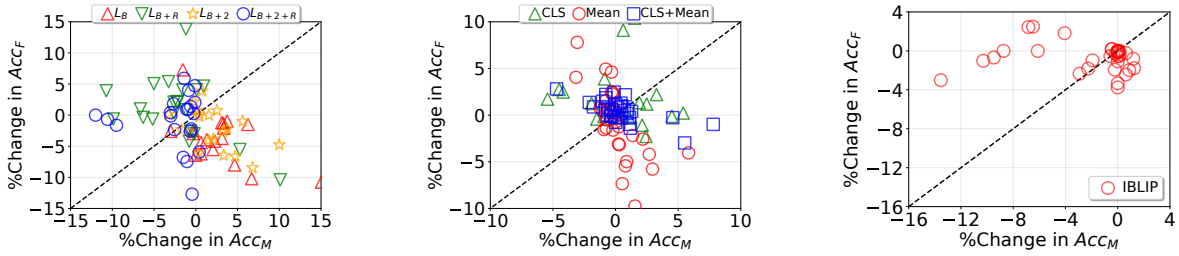
• *Choice of dataset determines perceived disparity*– Some datasets report disparity against females, whereas others report disparity against males, irrespective of the architecture and loss function. Thus, if the test sample changes, a model's direction and intensity of bias can become the opposite, irrespective of the learning involved. This is an important outcome not previously covered in audit studies, where most test datasets have similar distributions. Thus, in-the-wild out-of-distribution test points can make an FRS behave in a completely orthogonal way.

• *Model complexity overpowers loss function*– In simpler CNN models, the loss function plays an important role, whereas in the complex transformer models, the architecture that already has an enormous advantage of the pre-training stage lends more weightage than the loss function in determining the disparity.

Thus, we see that the answer to RQ2. is not monolithic; instead, the direction and intensity of disparity are heavily dependent on the choice of dataset and architecture.

Change in Male and Female Accuracies (RQ3)

We now take a look at the relative change in accuracy for males and females in every architecture for all loss functions when compared against only the CE loss. As mentioned in the previous section, all other losses are added along with the



(a) LBFC (b) ViT-Face (c) IBLIP

Figure 5: Relative change in accuracy for males vs. females, for all losses when compared against only the CE loss, for all FRSs and datasets. The diagonal line indicates an equal relative change for both genders and points on either side imply a larger relative change for the respective gender. The effect of tinkering with the models has an impact on the relative change in accuracy for both genders.

Architecture	Highest accuracy				Lowest disparity			
	Loss	Data	D_M	D_F	Loss	Data	D_M	D_F
LBFC$_B$	A+T	CFD	9.814	**9.952**	A+T	UTKFace	5.390	**5.409**
LBFC$_{B+R}$	A+T	CFD	8.185	**8.330**	A	LFW	8.413	**8.966**
LBFC$_{B+2}$	A	CFD	36.926	**37.569**	A	UTKFace	**21.442**	21.438
LBFC$_{B+2+R}$	A+T	CFD	29.853	**30.153**	A	UTKFace	**19.831**	19.792
ViT$_{CLS}$	A	CFD	**25.086**	24.903	A	LFW	**26.301**	24.969
ViT$_M$	A	CelebSET	**18.993**	18.744	T	CelebSET	**15.270**	15.217
	Cos		**25.381**	24.798				
ViT$_{CLS+M}$	A+T	CFD	21.618	**21.693**	Co+A+T	CFD	23.557	**23.724**
IBLIP	Co+A+T	CelebSET	12.551	**12.635**	Co+A+T	CelebSET	12.551	**12.635**
					Cos		82.551	**82.964**

Table 3: Average Euclidean distance between the anchor embeddings and embeddings obtained with other loss functions for both gender groups (D_M is avg. extent of shift for males and D_F is for females), for the combinations that result in the highest accuracy and lowest disparity, respectively. Maximum values are in bold. On average, the embeddings for females shift more than for males.

CE loss. In Figure 5, we plot this relative change as a scatter plot for each model. A diagonal line indicates an equal relative change for both males and females, and points lying on either side of the diagonal imply a larger relative change for one of the genders. Next, let us call the embeddings obtained using the simple CE loss the anchor embeddings. We also measure the average Euclidean distance between these anchor embeddings and the embeddings obtained when other types of losses are used for each dataset and architecture. This shows the average extent of the shift of the male (D_M) and female (D_F) representations from their respective anchor embeddings. In Table 3, we present the Euclidean distance values for the combinations that result in the highest accuracy and the lowest disparity, respectively (all other Euclidean distance values are in the Appendix).

Results for LibfaceID We study the relative change in the accuracies for males and females for the different architectures in Fig. 5a. (a) The scatter points have a large spread, thus indicating a lack of cohesive trend overall. On closer inspection, we note that the architectures with the residual connections are primarily placed above the diagonal on the top left quadrant, whereas the vanilla architectures are placed below the diagonal on the bottom right quadrant. This implies that adding residual connections favours a reduction in accuracy for males relative to an increase in accuracy for females and vice versa for the vanilla networks. (b) From Table 3, we see that both the highest accuracy and lowest disparity are generated whenever the ArcFace loss is present in the model's objective. (c) For the highest accuracy scenario, D_F is always higher than D_M. This shows that the embeddings for females shift a larger distance away from the anchor embeddings in the representation space as compared to males. On the other hand, the results are mixed for the lowest disparity scenario. In the case of the shallow CNN, D_F is higher (i.e., female representations shift away from the anchor embeddings more), while in the case of the deep CNN, the D_M is higher (i.e., male representations shift away from the anchor embeddings more).

Results for ViT Next, we look at the results for ViT in Fig. 5b. (a) We notice that each architecture behaves in a different manner– the model with only CLS embeddings is spread horizontally (relative change for accuracy of males is higher for negligible change in accuracy for females) on both sides of the diagonal; the model with only the mean

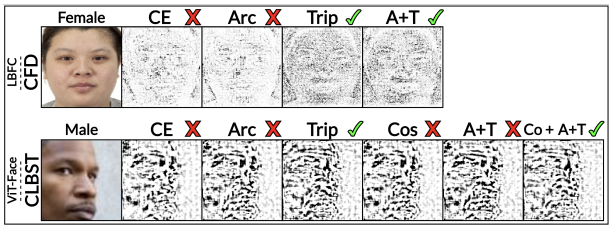


Figure 6: Explainability results on LBFC$_{B+2+R}$ (top row) and ViT$_{CLS+M}$ for sample images from the CFD & CelebSET dataset with all loss functions using the method of integrated gradients (Sundararajan, Taly, and Yan 2017). The original images, predictions and gradient attributions are presented.

embeddings is spread vertically on both sides of the diagonal, and the model which uses a concatenation of both types of embeddings has its' points primarily clustered around the centre (it is minimal, but an equivalent change in the accuracy for both gender groups). (b) From Table 3, we observe that for the high-accuracy scenario, the avg. D_M is higher than D_F for both ViT$_{CLS}$ and ViT$_M$. On the other hand, the opposite is true for ViT$_{CLS+M}$. For the lowest disparity scenario, we make a similar observation.

Results for InstructBLIP Finally, from Fig. 5c, we note that– (a) A majority ($> 50\%$) of the points are clustered around the 0 mark, whereas the rest are either in a narrow vertical band, in the bottom right quadrant near 0, or strewn horizontally in the top left quadrant. Thus, most data points imply minimal change in accuracy for both genders. (b) From Table 3, we see that the best performance (high accuracy, low disparity) is observed when the CosFace loss is involved. Also, the values of D_F are always higher than D_M.

Major takeaways The major takeaways from Figure 5 and Table 3 are as follows–

• *Change in accuracy of genders is determined by the choice of model and architecture*– Residual connections in LibfaceID and CLS embeddings in ViT seem to affect the accuracy of the males. On the other hand, the vanilla CNN setup for LibfaceID affects the female accuracy more. Thus, the effect of the tinkering of the models is often tied to the sensitive attributes that the dataset encompasses.

• *Embeddings of females are more sensitive to model choices than males*– The embeddings of females shift a larger distance from the anchor embeddings than males on average, for both the highest accuracy as well as the lowest disparity scenarios. This shows that the embeddings generated from images of females are more sensitive to the model changes– architecture and loss function and could be an avenue for reducing disparity.

Explaining the Observations

We use the method of integrated gradients (Sundararajan, Taly, and Yan 2017) to infer the reason for the predictions for each of the models. In Figure 6, we present the results for the explainability experiments on LBFC$_{B+2+R}$ (top row) and ViT$_{CLS+M}$ (bottom row) models with all loss functions for sample images from the CFD and CelebSET datasets respectively. We see how each model & loss function elicits a starkly different gradient attribution. Whenever the CNN model predicts a person as male, the attributions are more "localized", and the focus is on the eyebrows, nose, jawline or facial hair, and whenever the model predicts a person as female, the focus is more "globalized", distributed and balanced with a focus on the upper facial contours. In the top row, the correct prediction is thus obtained for the T and A+T loss functions that enforce localized attention. In the ViT model, we notice patch-based grid-like attributions wherein the attention is more on the eyes and upper face region when predicting female and more uniform throughout the face when predicting male. In the bottom row, the correct prediction is thus obtained for the T and Co+A+T loss functions that enforce uniform attention. While we present only a subset of results here due to paucity of space, the inferences from the other dataset-model-loss function combinations are very similar.

Discussion and Recommendations

We now discuss the high-level learnings from our extensive audit as well as the recommendations for future auditors, developers and users of such FRSs.

Summary

In this work, we attempted to understand how the three major components of an FRS impact the model's performance when observing two opposing objectives – accuracy and disparity, for the task of gender prediction. To address RQ1, we studied the accuracy vs disparity trade-off amongst the different architectures, loss functions and datasets from Figure 3. We trivially observed that with increasing model complexity, the accuracy increases, and the disparity reduces as models with deeper layers learn more salient discriminative features in the face images. Next, we stress on the importance of choosing diverse datasets – CFD (Ma, Correll, and Wittenbrink 2015), the most standard face dataset in our benchmark, consistently reports the highest accuracies (especially on the shallow CNN). CelebSET (Raji et al. 2020), a dataset of Hollywood celebrity faces, reports the highest accuracies on both ViT and InstructBLIP. We hypothesize that this is possibly due to data leakage from the pre-training stage. Finally, CelebA & FARFace (Jaiswal et al. 2024), a newly released dataset with significantly more images from the Global South who have different skin tones and facial features, prove to be the most adversarial, reporting high disparity. Hence, the choice of dataset may propagate a false sense of confidence or fear in the model's abilities and stakeholders need to better engage with the datasets in order to understand the FRS's true performance. This is especially true for social media platforms as they are deployed worldwide, and disparities can impact millions of users. We observed a strong tendency of clustering for each face dataset, with regards to the performance metrics, that is independent of the model and loss function under consideration.

Next, to address RQ2, we studied the heatmaps in Figure 4 for the extent and direction of disparity in each model for all architectures, loss functions and datasets. We discovered that the model size and dataset determine not only the extent of disparity but also the direction of disparity. For example, both CelebSET and FARFace are gender-balanced datasets of "in-the-wild" photos of public-facing individuals (Hollywood celebrities and cricketers, respectively), yet CelebSET consistently reports disparity against males and FARFace reports disparities against females. This again confirms that the choice of dataset has a huge impact on *the kind of bias a model is perceived to have*. We also trivially observed that in complex models, the loss functions play a less important role in determining the extent and direction of disparity.

In Figure 5 and Table 3, we attempted to quantify the relationship between the change in accuracy for the two gender groups to address RQ3. Our results show that the combination of architecture and loss function determines which gender group reports a higher accuracy and to what extent for all datasets. For example, ArcFace loss always gives the highest accuracy and lowest disparity, independent of the architecture; in the two most optimal models– VIT_{CLS+M} and InstructBLIP, the embeddings for females shift more than males. This shows that existing datasets have a lot of variety in what constitutes a "female face", and there is no general pattern for the same as for males. We hypothesize here that this is the leading cause of bias in FRSs, and ML developers need to design better models that can find more generalizable features, especially for female faces, to address the problem of disparity.

Recommendations for the Community

We perform an in-depth investigation into the relationship between model architecture, loss function and data and its ensuing impact on disparity for the task of gender prediction. Our experimental results show that all three components, individually as well as collectively, impact both accuracy and disparity. Any human-facing technology, especially one with the sensitive nature of deployment like an FRS, has multiple stakeholders – developers, users and regulators, all of whom need to perform due diligence before the technology can be released to the general public. From our study, we make some recommendations for developers and users. First, the model developers need to be cognizant of the various architectures and how they interact with each dataset (or the population where the model will be deployed)[3]. Similarly, the users need to be aware of the deployment scenario and which model suits their use case (Cherepanova et al. 2023). For example, a model to be deployed in the USA should be audited and designed with the CelebSET dataset as a template, whereas one that has to be deployed in India or the West Indies should use the FARFace dataset. We believe our study will provide the necessary awareness and blueprint that the community can benefit from and reduce the disparity and unfair outcomes that result from the bias in these models.

[3]https://www.paravision.ai/news/addressing-critical-inclusion-questions-for-face-recognition-technology-buyers/

Recommendations for social media platforms: Our results are especially useful for social media platforms that use FRSs for image tagging, advertising and scammer detection applications. We show in our experiments how each model performs differently over a given dataset and the associated disparities differ. Thus, social media platforms deployed across the world should take cognizance of this report and deploy more localized models that are trained on data reflective of the local geography to reduce instances of disparate performance. Further, various FRS companies, e.g., Clearview AI[4] routinely use billions of datapoints scraped from social media platforms to build their database to be shared with law enforcement agencies thus putting everyone into a "perpetual police line-up". Biases in this type of technology would only exacerbate the risks of false arrests or detentions. To minimize such risks, the social media platforms thus need to be extra vigilant perhaps making amendments to their data scraping/download policies.

Limitations of our Study

Similar to many other studies in this field, ours also has certain limitations, which we acknowledge and clarify here.

Definition of gender: Historically, gender has been classified as binary– masculine and feminine. In this model, gender is aligned with the sex assigned at birth. Such definitions often conform to legal, governmental, political, societal and even technological expectations, wherein having the lowest common denominator definition suits most use cases. Most government IDs only recognize the binary male or female gender identity, which technology like face recognition systems has to adhere to, especially in security applications like airport entry. This binary definition has been criticized by scholars of intersectionality, who maintain that such a structure maintains patriarchal and supremacist norms (Scaptura and Hayes 2023). We acknowledge and reiterate that gender is fluid and can include various groups like transgender, agender, intersex and other non-binary identifying individuals. Due to the reasons stated here, we are limited by the choice of only binary gender labels available in all existing datasets. Thus, predictions should be interpreted as the "*perceived*" binary gender.

Dataset artifacts: We understand that each face dataset has certain artifacts like representation of a narrow socioeconomic demographic, age group and other factors, which may limit the generalizability of our findings. To circumvent this issue to an extent we have selected a multitude of datasets roughly covering a mix of both the well-studied and the newly introduced ones. Further construction of large-scale datasets encompassing different sociodemographic backgrounds seems imperative. This points to an interesting future direction of research.

Conclusion

In this study, we investigate the role of datasets, model architectures and loss functions in shaping the accuracy-disparity landscape in FRSs for the task of gender prediction. We use three different models– LibfaceID (a CNN architecture),

[4]https://tinyurl.com/mrxbtetx

ViT-Face (a vision transformer architecture) and InstructBLIP (a vision-language model architecture), seven diverse face datasets – CFD, CelebSET, FARFace, LFW, UTKFace, Fairface and CelebA, and four loss functions– Cross Entropy, Triplet, ArcFace and CosFace loss functions. We modify the FRSs into multiple variants to create ten models that are then trained and fine-tuned with a combination of the loss functions mentioned here, resulting in 266 evaluation configurations.

Our large-scale study shows that the choice of the loss function and evaluation dataset is very important to understand the true bias of the FRS. Highly standard datasets like CFD result in high accuracy and low disparity, whereas adversarial datasets like FARFace & CelebA expose the true shortcomings of the FRS. Triplet loss, used in previous studies to improve the performance of FRSs does not work as well when used on different architectures and datasets. Similarly, model size and dataset combine to determine which gender performs better, with two "in-the-wild" datasets with popular individuals reporting disparities in exactly opposite directions. Finally, we observe that the architecture and loss function also combine to determine which gender group has a higher accuracy and to what extent for all datasets. Our most interesting observation is the fact that existing datasets have a lot of variety for "female faces" but not so much for male faces, which could be a leading cause of bias in FRSs.

Future Work

We plan to extend this work to larger and more diverse vision architectures and a more diverse choice of loss functions and datasets. We also plan to use our findings to devise FRS debiasing algorithms that can improve the performance for the genders reporting lower accuracy and bring more parity in the model's performance.

Ethics Statement

Our in-depth study shows that FRSs have disparity irrespective of the dataset, choice of architecture or loss functions. Thus, existing algorithms and datasets are still not apt for large-scale deployment in society. Our study attempts to shed light on the reason for this disparity, which model developers and users can take cognizance of for their own deployment use cases. We acknowledge that gender is a spectrum but limit ourselves to binary gender prediction due to the gender labels available in the benchmark datasets. We have shared the model cards (Mitchell et al. 2019) of all our FRS models on Github [2].

Acknowledgments

S. Jaiswal is grateful for the financial assistance by PMRF, Govt. of India and AI4CPS, IIT Kharagpur. S. Basu acknowledges the DAAD KOSPIE fellowship.

References

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In *ACM SIGKDD*.

Amazon. 2021. Amazon AWS Rekognition. https://aws.amazon.com/rekognition/faqs/. Accessed: 2021-04-01.

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2019. Social B(eye)as: Human and Machine Descriptions of People Images. *AAAI ICWSM*, 13.

BBC. 2021a. Facebook apology as AI labels black men 'primates'. https://tinyurl.com/346n29ys. Accessed: 2022-08-20.

BBC. 2021b. Twitter finds racial bias in image-cropping AI. https://www.bbc.com/news/technology-57192898. Accessed: 2022-08-20.

Best, S. 2020. Woman's lips mistaken for open mouth by 'racist' online passport checker. https://tinyurl.com/6jar9969. Accessed: 2024-01-01.

Brinkmann, J.; Swoboda, P.; and Bartelt, C. 2023. A multidimensional analysis of social biases in vision transformers. In *IEEE/CVF ICCV*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**.

Cabannes, V.; Kiani, B.; Balestriero, R.; LeCun, Y.; and Bietti, A. 2023. The ssl interplay: Augmentations, inductive bias, and generalization. In *ICML*.

Chakraborty, A.; Messias, J.; Benevenuto, F.; Ghosh, S.; Ganguly, N.; and Gummadi, K. 2017. Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. In *AAAI ICWSM*.

Chen, D.; Hua, G.; Wen, F.; and Sun, J. 2016. Supervised transformer network for efficient face detection. In *ECCV*.

Cherepanova, V.; Reich, S.; Dooley, S.; Souri, H.; Dickerson, J.; Goldblum, M.; and Goldstein, T. 2023. A Deep Dive into Dataset Imbalance and Bias in Face Identification. In *AAAI/ACM AIES*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Dai, W.; Li, J.; LI, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *NeurIPS*.

Dan, J.; Liu, Y.; Xie, H.; Deng, J.; Xie, H.; Xie, X.; and Sun, B. 2023. TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective. In *IEEE/CVF ICCV*.

d'Ascoli, S.; Gabrié, M.; Sagun, L.; and Biroli, G. 2021. On the interplay between data structure and loss function in classification problems. *NeurIPS*.

Davidian, M.; Lahav, A.; Joshua, B.-Z.; Wand, O.; Lurie, Y.; and Mark, S. 2024. Exploring the interplay of dataset size and imbalance on CNN performance in healthcare: Using X-rays to identify COVID-19 patients. *Diagnostics*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eidinger, E.; Enbar, R.; and Hassner, T. 2014. Age and gender estimation of unfiltered faces. *IEEE TIFS*.

Face++. 2021. Face++ Detect. https://www.faceplusplus.com/face-detection/. Accessed: 2021-04-01.

Facebook. 2021. An Update On Our Use of Face Recognition. https://tinyurl.com/4zzkmr9v. Accessed: 2024-12-12.

Facebook. 2024. Testing New Ways to Combat Scams and Help Restore Access to Compromised Accounts. https://tinyurl.com/34s47vbj. Accessed: 2024-12-12.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF CVPR*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*.

Google. 2024. Set up & manage your face groups. https://tinyurl.com/3zrnjz88. Accessed: 2024-12-12.

Hitoshi, I. 2016. Video face recognition system enabling real-time surveillance. *NEC Technical Journal*.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical report, University of Massachusetts, Amherst.

Jaiswal, S.; Duggirala, K.; Dash, A.; and Mukherjee, A. 2022. Twoface: Adversarial audit of commercial face recognition systems. In *AAAI ICWSM*.

Jaiswal, S.; Ganai, A.; Dash, A.; Ghosh, S.; and Mukherjee, A. 2024. Breaking the Global North Stereotype: A Global South-centric Benchmark Dataset for Auditing and Mitigating Biases in Facial Recognition Systems. In *AAAI/ACM AIES*.

Janssen, P.; and Sadowski, B. M. 2021. Bias in Algorithms: On the trade-off between accuracy and fairness. In *ITS*.

John Dunne. 2019. Man stunned as passport photo check sees lips as open mouth. https://tinyurl.com/bdcrb4h8.

Jung, S.-g.; An, J.; Kwak, H.; Salminen, J.; and Jansen, B. 2018. Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race. *AAAI ICWSM*.

Karkkainen, K.; and Joo, J. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF WACV*.

Kyriakou, K.; Barlas, P.; Kleanthous, S.; and Otterbacher, J. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *AAAI ICWSM*.

Lakshmi, A.; Wittenbrink, B.; Correll, J.; and Ma, D. S. 2021. The India Face Set: International and Cultural Boundaries Impact Face Impressions and Perceptions of Category Membership. *Frontiers in Psychology*.

Levi, G.; and Hassner, T. 2015. Age and gender classification using convolutional neural networks. In *IEEE CVPRW*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Liu, R.; Liu, H.; Li, G.; Hou, H.; Yu, T.; and Yang, T. 2022. Contextual debiasing for visual recognition with causal mechanisms. In *IEEE/CVF CVPR*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*.

Livemint. 2021. This Indian City has the highest number of CCTVs in the World. https://tinyurl.com/39kc87wy. Accessed: 2024-01-01.

Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*.

Ma, D. S.; Kantner, J.; and Wittenbrink, B. 2020. Chicago Face Database: Multiracial expansion. *Behavior Research Methods*.

Mennecke, B. E.; and Peters, A. 2013. From avatars to mavatars: The role of marketing avatars and embodied representations in consumer profiling. *Business Horizons*.

Messias, J.; Vikatos, P.; and Benevenuto, F. 2017. White, Man, and Highly Followed: Gender and Race Inequalities in Twitter. In *ACM WI*.

Microsoft. 2021. Microsoft Azure Face. https://azure.microsoft.com/en-in/services/cognitive-services/face/. Accessed: 2021-04-01.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *ACM FAT**.

Ning, X.; Xu, S.; Nan, F.; Zeng, Q.; Wang, C.; Cai, W.; Li, W.; and Jiang, Y. 2022. Face editing based on facial recognition features. *IEEE TCDS*.

O'Toole, A. J.; and Castillo, C. D. 2021. Face recognition by humans and machines: three fundamental advances from deep learning. *ARVS*.

Pang, R.; Baretto, A.; Kautz, H.; and Luo, J. 2015. Monitoring adolescent alcohol use via multimodal analysis in social multimedia. In *IEEE Big Data*.

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *AAAI/ACM AIES*.

Reuters. 2010. Japan vending machine recommends drinks to buyers. https://tinyurl.com/apbh6p5e. Accessed: 2024-01-01.

Scaptura, M. N.; and Hayes, B. E. 2023. Systems of Power and Femicide: The Intersection of Race, Gender, and Extremist Violence. Routledge.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE CVPR*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *ICML*.

Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*.

Vikatos, P.; Messias, J.; Miranda, M.; and Benevenuto, F. 2017. Linguistic Diversities of Demographic Groups in Twitter. In *ACM HT*.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *IEEE CVPR*.

Yang, M.-H.; Kriegman, D. J.; and Ahuja, N. 2002. Detecting faces in images: A survey. *IEEE TPAMI*.

Zhang, M. 2015. Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software. https://tinyurl.com/2p9a72h. Accessed: 2022-08-20.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *IEEE CVPR*.

Zhong, Y.; and Deng, W. 2021. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Conclusion
   (e) Did you describe the limitations of your work? Yes, see Conclusion.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethics Statement
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. All hyperparameters and experimental details are shared. The code is available on Github [2].
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Table 4 and Appendix (Results on Adience, Experimental Details)
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Appendix (Results on Adience, Experimental Details)
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, Datasets– (Eidinger, Enbar, and Hassner 2014; Raji et al. 2020; Karkkainen and Joo 2021; Jaiswal et al. 2024; Zhang, Song, and Qi 2017; Huang et al. 2007; Liu et al. 2015), and FRSs– (Levi and Hassner 2015; Zhong and Deng 2021; Dai et al. 2023)
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Additional details on experimental design

Architectures and loss functions for LibfaceID In addition to the models described in the main draft, we create two more models with weighted residual connections. The number of layers and connections are the same as described in Fig. 2. The weight of a residual connection determines how much information is passed on to the next layer through this connection, calculated using the given formula–

$$\alpha \times O_{L-1} + (1 - \alpha) \times O_L$$

where O_L is the output for layer L and α is a heuristic. We perform a grid search using Optuna (Akiba et al. 2019) to calculate the value of α as 0.3.

For all models where we use ArcFace and Triplet loss along with Cross-Entropy loss, the weights are– $CE+\alpha T+\beta A$, where $(\alpha, \beta) = (0.7, 0.1)$.

Loss functions for vision transformer For ViT$_M$ and ViT$_{CLS+M}$, we use CosFace, ArcFace and Triplet loss with Cross-Entropy loss in various combinations, whose weights are determined as follows– (a) $CE + \alpha T + \beta A$, where $(\alpha, \beta) = (0.6, 0.6)$ for ViT$_M$ and $(\alpha, \beta) = (0.8, 0.9)$ for ViT$_{CLS+M}$, and (b) $CE + \alpha T + \beta A + \gamma Cos$, where $(\alpha, \beta, \gamma) = (0.9, 0.4, 1)$ for ViT$_M$ and $(\alpha, \beta, \gamma) = (0.4, 1, 0.8)$ for ViT$_{CLS+M}$.

Loss functions for InstructBLIP For InstructBLIP, the combination of CosFace, ArcFace and Triplet loss with Cross-Entropy loss are weighted in the following manner– (a) $CE + \alpha T + \beta A$, where $(\alpha, \beta) = (0.9, 0.4)$, (b) $CE + \alpha T + \beta A + \gamma Cos$, where $(\alpha, \beta, \gamma) = (0.8, 1, 0.5)$.

Hyperparameters

In Table 4, we present the key hyperparameters – the learning rate (LR), the optimizer and the number of epochs – used in our experiments.

Model	LR	Optimizer	#Epochs
LBFC$_B$	9×10^{-3}	SGD	100
LBFC$_{B+R}$	2×10^{-3}		
LBFC$_{B+R\alpha}$	9×10^{-3}		
LBFC$_{B+2}$	1×10^{-3}		
LBFC$_{B+2+R}$	9×10^{-3}		
LBFC$_{B+2+R\alpha}$	6×10^{-4}		
ViT$_{CLS}$	7×10^{-4}	AdamW	
ViT$_M$	3×10^{-5}		
ViT$_{CLS+M}$	1×10^{-4}		
IBLIP	2×10^{-3}		

Table 4: Training and fine-tuning hyperparameters.

Results on Adience

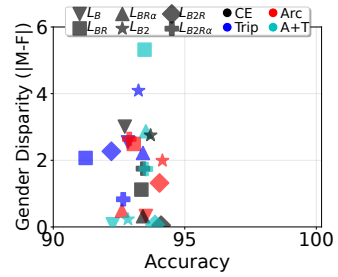
Experimental Details We split the Adience dataset into train, validation, and test sets in an 80:10:10 ratio. Thus, we have 810 males and 935 females in the test set for Adience. We present the results for this set as follows.

We train/fine-tune our FRS models on an Ubuntu 18.04 LTS Intel(R) Xeon(R) Gold 6126 CPU server with NVIDIA Tesla P100 GPU (CUDA v12.2) $\times$ 2, 128 GB RAM and 48 cores. The model hyperparameters are present in Table 4, chosen using Optuna.

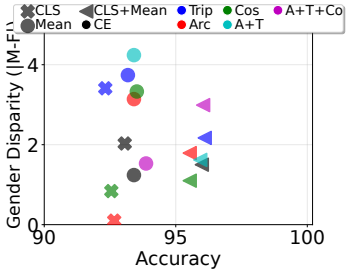
Accuracy vs disparity From Fig. 7, we can see that for all FRSs, the scatter points lie in a vertical line, indicating a relatively stable accuracy but diverse absolute disparity values. We also note that with increasing model complexity, the accuracy value increases as well. Interestingly, the disparity remains consistent, upper bounded at 6%. For ViT (Fig. 7b), we also note that when the type of embedding changes from trivial to more detailed, the model accuracy increases.

Direction of disparity In Fig. 8, we look at the heatmaps of disparity for the different loss functions and FRSs. The color indicates the direction of disparity, with red implying higher accuracy for males and blue implying higher accuracy for females. Since the LibfaceID model is trained from scratch on the Adience dataset, which has more female images, the test accuracy for females is naturally higher (as evident from the first 6 rows of Fig. 8, especially for models not using Triplet loss). For ViT, we note that only ViT$_M$ reports disparity against females. Finally, InstructBLIP reports disparities against males as well.

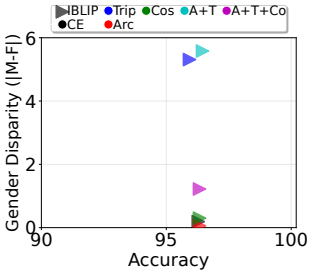
Change in male and female accuracies In Figure 9, we observe the change in male and female accuracies for each FRS-architecture-loss combination on the Adience test set. For all FRSs, we note that the scatter points are distributed



(a) LBFC



(b) ViT



(c) IBLIP

Figure 7: Accuracy vs. absolute gender disparity for the Adience test set, for each type of model backbone, independent of the loss functions. The accuracy increases with increasing model complexity. The absolute disparity remains consistently upper bounded at 6%.

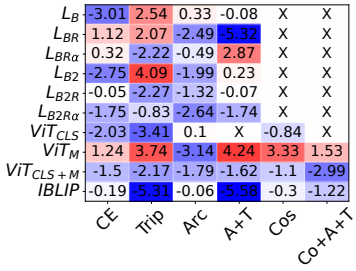


Figure 8: Heatmap for each type of FRS and loss functions tested on Adience. This plot shows the direction of disparity– against females (+ve) or against males (-ve) across all the models and losses.

along the cross-diagonal, implying that there is an inverse relationship between the change in accuracies for males and females. For InstructBLIP, we note that all scatter points are above the diagonal– the change in accuracy for males is always negative, and for females is always positive.

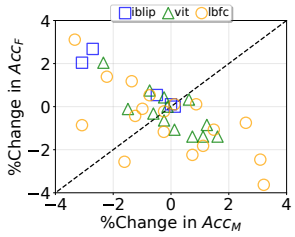
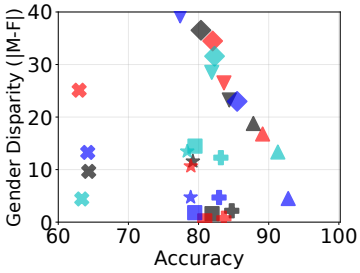
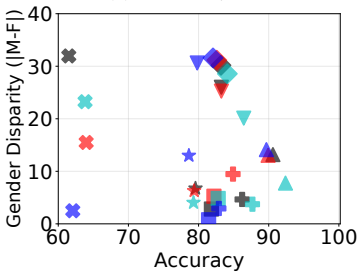


Figure 9: Change in male and female accuracies for each FRS tested on Adience.

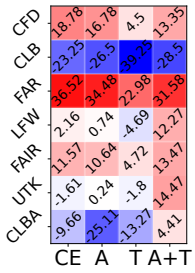


(a) LBFC$_{B+R\alpha}$

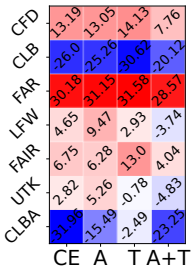


(b) LBFC$_{B+2+R\alpha}$

Figure 10: Accuracy vs absolute gender disparity for the LibfaceID FRS with weighted residual connections.



(a) LBFC$_{B+R\alpha}$



(b) LBFC$_{B+2+R\alpha}$

Figure 11: Direction of disparity for all datasets and all loss functions on the LibfaceID FRS with weighted residual connections.

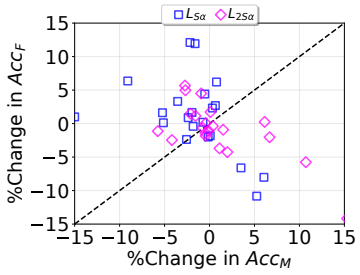


Figure 12: Change in male and female accuracies for the LibfaceID FRS with weighted residual connections.

Results on LibfaceID with weighted residual connections

We now look at the experimental results on the LibfaceID FRS with weighted residual connections in Figures 10, 11 and 12. The results are similar to those observed for other LibfaceID models (discussed in the main draft). We discuss each result as follows–

- In Fig. 10, we can see that FARFace always reports the highest disparity, whereas UTKFace reports the lowest. Next, CFD reports the highest accuracy and CelebA the lowest. We also notice a strong clustering in the results for each dataset with regard to one of the objectives.
- In the heatmaps in Figure 11, we see that CelebSET and FARFace always report the highest disparity across architecture and loss function variations. Interestingly, CelebSET and FARFace are both balanced datasets, however their disparities are in exact opposite directions. UTKFace always reports the lowest disparity among all datasets (Figure 10).
- In Figure 12, we look at the change in male and female accuracies. Here, we can observe that the points lie primarily on or above the diagonal, indicating that either the accuracies for males and females do not change much, and if they do, then the accuracy for males increases less often than the accuracy for females increases.

Change in embeddings for males and females

In Tables 5–14, we look at the change in the position of the embeddings generated for males and females for each FRS in the different architecture, loss function and dataset combination. The metric we use for calculating this change is the Euclidean distance between the embeddings generated with only the Cross-Entropy loss and other loss function combinations. The key takeaways are as follows –

- Among all the datasets, CFD always has a higher Euclidean distance for females than males, whereas FARFace always has a higher Euclidean distance for males instead of females. This indicates the sensitivity of each gender group in individual datasets to the model architecture and loss function. FARFace reports the largest difference between the distance measures for males and females.
- The results for all LibfaceID models (Tables 5–10) report a higher Euclidean distance value for males than females.

Loss	Data	D_M	D_F
T	Adience	6.065	**6.310**
	CelebSET	**6.802**	6.663
	CFD	9.816	**9.964**
	FairFace	**5.414**	5.251
	FARFace	**9.911**	5.335
	UTKFace	5.417	**5.436**
	LFW	5.344	**5.627**
	CelebA	4.232	**4.859**
A	Adience	**8.621**	8.451
	CelebSET	**9.225**	9.023
	CFD	13.471	**13.640**
	FairFace	**7.614**	7.169
	FARFace	**13.755**	7.644
	UTKFace	7.452	**7.469**
	LFW	**7.512**	7.458
	CelebA	6.304	**6.739**
A+T	Adience	6.070	**6.195**
	CelebSET	**6.769**	6.631
	CFD	9.814	**9.953**
	FairFace	**5.397**	5.172
	FARFace	**9.970**	5.264
	UTKFace	5.390	**5.410**
	LFW	5.369	**5.539**
	CelebA	4.194	**4.829**

Table 5: Average Euclidean distance for LBFC$_B$.

On adding residual connections to the architecture (Tables 6 and 9), the number of datasets following this trend increases, as opposed to non-residual networks.

- From Table 11, we see that the change in embedding positions for males and females in ViT$_{CLS}$ is closer than observed for LibfaceID. Tables 12 and 13 show that across different loss functions, the behaviour of the datasets remains the same. For both models, Adience, CelebSET and FARFace always report higher Euclidean distance measures for males than females, indicating that male images observe a larger shift in the embedding space than females.
- In Table 14, we observe the results for InstructBLIP. The trend in the change of embedding position is constant for males and females across all the loss functions. Interestingly, upon applying the CosFace loss to the FRS, the Euclidean distance is observed to be the highest. This shows that the CosFace loss changes the positions of the embeddings the most out of all loss functions.

Loss	Data	D_M	D_F
T	Adience	**5.563**	5.257
	CelebSET	**5.532**	5.478
	CFD	8.094	**8.238**
	FairFace	**4.913**	4.448
	FARFace	**8.422**	4.636
	UTKFace	**4.592**	4.588
	LFW	4.360	**4.668**
	CelebA	3.515	**4.039**
A	Adience	**10.049**	9.896
	CelebSET	**10.386**	10.345
	CFD	15.375	**15.486**
	FairFace	**8.912**	8.485
	FARFace	**15.482**	9.269
	UTKFace	8.435	**8.450**
	LFW	8.003	**8.628**
	CelebA	7.203	**7.945**
A+T	Adience	**5.556**	5.432
	CelebSET	**5.637**	5.597
	CFD	8.186	**8.330**
	FairFace	**4.898**	4.542
	FARFace	**8.455**	4.721
	UTKFace	**4.645**	4.638
	LFW	4.358	**4.756**
	CelebA	3.553	**4.071**

Table 6: Average Euclidean distance for LBFC$_{B+R}$.

Loss	Data	D_M	D_F
T	Adience	**4.829**	4.573
	CelebSET	**4.948**	4.943
	CFD	7.360	**7.470**
	FairFace	**4.257**	3.841
	FARFace	**7.558**	4.057
	UTKFace	3.988	**3.991**
	LFW	3.896	**4.008**
	CelebA	**3.220**	3.096
A	Adience	**6.765**	6.455
	CelebSET	6.889	**6.892**
	CFD	10.253	**10.400**
	FairFace	**6.027**	5.510
	FARFace	**10.496**	5.820
	UTKFace	5.672	**5.687**
	LFW	5.514	**5.820**
	CelebA	4.863	**5.038**
A+T	Adience	**5.075**	4.664
	CelebSET	**5.044**	5.039
	CFD	7.515	**7.623**
	FairFace	**4.458**	3.934
	FARFace	**7.814**	4.145
	UTKFace	4.141	**4.141**
	LFW	**4.136**	4.098
	CelebA	**3.330**	3.165

Table 7: Average Euclidean distance for LBFC$_{B+R\alpha}$.

Loss	Data	D_M	D_F
T	Adience	4.186	**4.296**
	CelebSET	**4.182**	4.162
	CFD	5.439	**5.514**
	FairFace	**3.767**	3.630
	FARFace	**5.790**	3.700
	UTKFace	**3.775**	3.771
	LFW	3.652	**3.857**
	CelebA	3.137	**3.280**
A	Adience	24.221	**25.856**
	CelebSET	**26.519**	26.190
	CFD	36.926	**37.570**
	FairFace	20.708	**20.929**
	FARFace	**35.723**	22.702
	UTKFace	**21.442**	21.438
	LFW	19.846	**22.800**
	CelebA	15.693	**16.190**
A+T	Adience	4.205	**4.290**
	CelebSET	**4.219**	4.195
	CFD	5.522	**5.581**
	FairFace	**3.823**	3.674
	FARFace	**5.786**	3.790
	UTKFace	**3.818**	3.816
	LFW	3.732	**3.860**
	CelebA	3.247	**3.391**

Table 8: Average Euclidean distance for LBFC$_{B+2}$.

Loss	Data	D_M	D_F
T	Adience	**21.188**	18.646
	CelebSET	**20.688**	20.279
	CFD	29.515	**29.814**
	FairFace	**17.330**	15.228
	FARFace	**33.828**	15.780
	UTKFace	**16.531**	16.460
	LFW	15.739	**17.592**
	CelebA	**10.902**	10.481
A	Adience	**24.422**	22.504
	CelebSET	**24.300**	23.899
	CFD	33.741	**34.097**
	FairFace	**20.516**	18.762
	FARFace	**37.753**	19.571
	UTKFace	**19.831**	19.793
	LFW	18.819	**20.938**
	CelebA	**14.626**	14.214
A+T	Adience	**21.344**	19.186
	CelebSET	**21.085**	20.682
	CFD	29.853	**30.153**
	FairFace	**17.491**	15.628
	FARFace	**33.992**	16.119
	UTKFace	**16.822**	16.756
	LFW	15.886	**18.048**
	CelebA	**11.154**	10.842

Table 9: Average Euclidean distance for LBFC$_{B+2+R}$.

Loss	Data	D_M	D_F
T	Adience	**7.230**	6.961
	CelebSET	**6.982**	6.867
	CFD	9.736	**9.841**
	FairFace	**6.399**	5.893
	FARFace	**10.492**	6.056
	UTKFace	5.977	**5.988**
	LFW	**6.080**	6.002
	CelebA	4.677	**5.008**
A	Adience	10.358	**10.478**
	CelebSET	**10.359**	10.215
	CFD	14.157	**14.331**
	FairFace	**9.321**	8.886
	FARFace	**15.395**	9.041
	UTKFace	8.958	**8.982**
	LFW	**9.082**	8.892
	CelebA	7.487	**7.793**
A+T	Adience	**7.039**	6.924
	CelebSET	**6.926**	6.817
	CFD	9.621	**9.721**
	FairFace	**6.222**	5.841
	FARFace	**10.321**	6.018
	UTKFace	5.879	**5.889**
	LFW	5.951	**5.986**
	CelebA	4.572	**4.949**

Table 10: Average Euclidean distance for LBFC$_{B+2+R\alpha}$.

Loss	Data	D_M	D_F
T	Adience	18.272	**18.916**
	CelebSET	**15.907**	15.757
	CFD	**17.979**	17.939
	FairFace	16.237	**17.444**
	FARFace	**17.535**	17.324
	UTKFace	16.119	**16.140**
	LFW	16.626	**16.703**
	CelebA	**15.893**	15.697
A	Adience	**32.986**	32.878
	CelebSET	**19.849**	19.509
	CFD	**25.086**	24.904
	FairFace	31.482	**33.344**
	FARFace	**25.876**	25.748
	UTKFace	32.525	**32.543**
	LFW	**26.301**	24.969
	CelebA	**21.908**	20.727
Cos	Adience	27.578	**28.406**
	CelebSET	**20.541**	20.434
	CFD	24.320	**24.821**
	FairFace	**25.632**	26.828
	FARFace	**24.161**	23.258
	UTKFace	**26.103**	26.051
	LFW	**23.704**	22.962
	CelebA	**22.486**	20.575

Table 11: Average Euclidean distance for VIT$_{CLS}$.

Loss	Data	D_M	D_F
T	Adience	**14.772**	14.722
	CelebSET	**15.270**	15.218
	CFD	14.854	**14.906**
	FairFace	**14.552**	14.488
	FARFace	**15.241**	14.495
	UTKFace	14.514	**14.524**
	LFW	14.206	**14.466**
	CelebA	13.223	**13.545**
A	Adience	17.158	**17.186**
	CelebSET	**18.993**	18.745
	CFD	17.800	**17.936**
	FairFace	**16.778**	16.509
	FARFace	**20.238**	15.430
	UTKFace	17.242	**17.298**
	LFW	16.410	**17.612**
	CelebA	14.129	**15.229**
Cos	Adience	**21.674**	18.266
	CelebSET	**25.381**	24.799
	CFD	22.761	**23.237**
	FairFace	**20.233**	16.986
	FARFace	**28.699**	17.275
	UTKFace	20.099	**20.108**
	LFW	**24.187**	18.723
	CelebA	15.048	**15.845**
A+T	Adience	**14.733**	14.264
	CelebSET	**14.990**	14.880
	CFD	14.730	**14.786**
	FairFace	**14.492**	14.079
	FARFace	**15.693**	14.049
	UTKFace	14.476	**14.492**
	LFW	**14.055**	13.968
	CelebA	12.828	**12.935**
Co+A+T	Adience	**13.616**	13.352
	CelebSET	**14.344**	14.281
	CFD	14.024	**14.081**
	FairFace	**13.382**	13.140
	FARFace	**14.690**	13.209
	UTKFace	13.457	**13.472**
	LFW	13.217	**13.474**
	CelebA	12.393	**12.738**

Table 12: Average Euclidean distance for VIT$_M$.

Loss	Data	D_M	D_F
T	Adience	**24.426**	24.327
	CelebSET	**25.198**	25.060
	CFD	25.249	**25.314**
	FairFace	23.557	**23.562**
	FARFace	**25.939**	24.298
	UTKFace	23.370	**23.413**
	LFW	23.390	**24.059**
	CelebA	21.535	**21.716**
A	Adience	**27.325**	26.912
	CelebSET	**26.790**	26.070
	CFD	27.478	**28.000**
	FairFace	24.922	**25.736**
	FARFace	**30.054**	24.276
	UTKFace	25.881	**25.979**
	LFW	24.670	**26.402**
	CelebA	20.238	**22.034**
Cos	Adience	**34.509**	33.279
	CelebSET	**34.650**	33.981
	CFD	35.711	**36.541**
	FairFace	30.560	**30.724**
	FARFace	**40.118**	30.873
	UTKFace	30.130	**30.332**
	LFW	32.158	**32.271**
	CelebA	27.341	**27.588**
A+T	Adience	**21.087**	20.593
	CelebSET	**21.845**	21.620
	CFD	21.618	**21.693**
	FairFace	**20.249**	19.858
	FARFace	**22.792**	20.493
	UTKFace	19.803	**19.836**
	LFW	20.777	**20.792**
	CelebA	**19.713**	19.554
Co+A+T	Adience	**22.915**	22.811
	CelebSET	**23.911**	23.658
	CFD	23.558	**23.724**
	FairFace	**22.024**	21.746
	FARFace	**24.840**	22.293
	UTKFace	21.575	**21.635**
	LFW	22.474	**22.752**
	CelebA	**21.102**	20.643

Table 13: Average Euclidean distance for VIT$_{CLS+M}$.

Loss	Data	D_M	D_F
T	Adience	10.077	**12.524**
	CelebSET	12.243	**12.308**
	CFD	12.595	**12.737**
	FairFace	**10.968**	10.593
	FARFace	**13.529**	10.987
	UTKFace	**11.550**	11.526
	LFW	11.256	**12.106**
	CelebA	10.462	**10.830**
A	Adience	14.040	**16.225**
	CelebSET	15.771	**15.834**
	CFD	16.490	**16.631**
	FairFace	**15.622**	14.735
	FARFace	**17.888**	15.010
	UTKFace	**15.536**	15.496
	LFW	14.609	**14.766**
	CelebA	13.268	**13.451**
Cos	Adience	75.745	**78.757**
	CelebSET	82.551	**82.964**
	CFD	82.158	**83.226**
	FairFace	**81.578**	64.050
	FARFace	**110.671**	58.004
	UTKFace	**78.920**	78.233
	LFW	**80.738**	70.964
	CelebA	**67.583**	64.406
A+T	Adience	9.938	**12.611**
	CelebSET	12.215	**12.287**
	CFD	12.625	**12.771**
	FairFace	**10.859**	10.659
	FARFace	**13.501**	11.019
	UTKFace	**11.546**	11.525
	LFW	11.149	**12.052**
	CelebA	10.271	**10.717**
Co+A+T	Adience	10.298	**12.916**
	CelebSET	12.551	**12.635**
	CFD	12.988	**13.127**
	FairFace	**11.303**	10.984
	FARFace	**13.943**	11.433
	UTKFace	**11.874**	11.853
	LFW	11.416	**12.300**
	CelebA	10.549	**10.991**

Table 14: Average Euclidean distance for IBLIP.