A Deep Dive into Dataset Imbalance and Bias in Face Identification

Valeriia Cherepanova *
University of Maryland

Steven Reich *
University of Maryland

Samuel Dooley
University of Maryland

Hossein Souri
Johns Hopkins University

John Dickerson
University of Maryland

Micah Goldblum
New York University

Tom Goldstein
University of Maryland

Abstract

As the deployment of automated face recognition (FR) systems proliferates, bias in these systems is not just an academic question, but a matter of public concern. Media portrayals often center imbalance as the main source of bias, i.e., that FR models perform worse on images of non-white people or women because these demographic groups are underrepresented in training data. Recent academic research paints a more nuanced picture of this relationship. However, previous studies of data imbalance in FR have focused exclusively on the face *verification* setting, while the face *identification* setting has been largely ignored, despite being deployed in sensitive applications such as law enforcement. This is an unfortunate omission, as 'imbalance' is a more complex matter in identification; imbalance may arise in not only the training data, but also the testing data, and furthermore may affect the proportion of identities belonging to each demographic group *or* the number of images belonging to each identity. In this work, we address this gap in the research by thoroughly exploring the effects of each kind of imbalance possible in face identification, and discuss other factors which may impact bias in this setting.

1 Introduction

Automated face recognition is becoming increasingly prevalent in modern life, with applications ranging from improving user experience (such as automatic face-tagging of photos) to security (e.g., phone unlocking or crime suspect identification). While these advances are impressive achievements, decades of research have demonstrated disparate performance in FR systems depending on a subject's race [28, 4], gender presentation [2, 1], age [19], and other factors.

Modern FR systems employ neural networks trained on large datasets and one major source of model bias is disparities in rates of representation of different groups in the dataset. Dataset imbalance is a much more complex and nuanced issue than it may seem at first blush. While a naive conception of 'dataset imbalance' is simply as a disparity in the *number* of images per group, this disparity can manifest itself as either a gap in the number of identities per group, or in the number of images per identity. Furthermore, dataset imbalance can be present in different ways in both the training and testing data, and these two source of imbalance can have radically different (and often opposite) effects on downstream model bias.

*Equal contribution

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

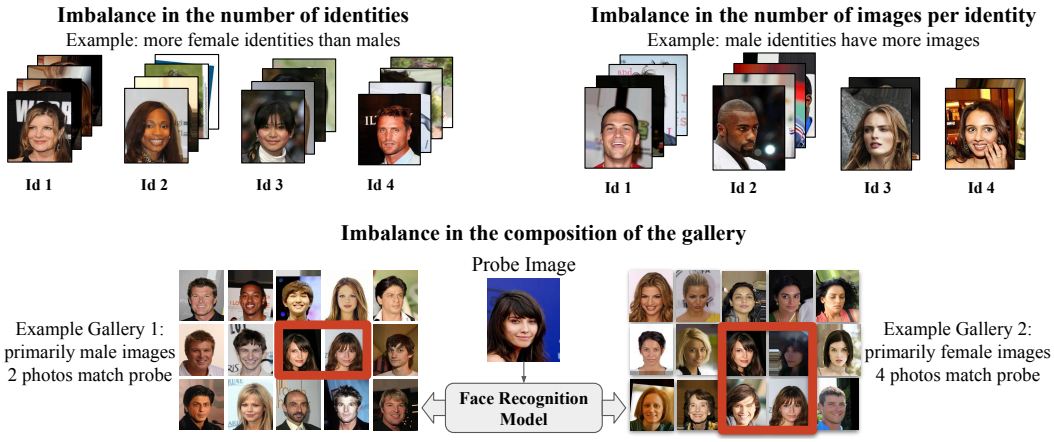

Figure 1: **Examples of imbalance in face identification**. Top left: data containing more female identities than male identities. Top right: data containing the same number of male and female identities, but more images per male identity. Bottom: two possible test (gallery) sets showing how the effects of different kinds of imbalance may interact.

Past work has only considered the *verification* setting of FR, where testing consists of determining whether a pair of images belongs to the same identity [28, 19, 1, 15]. As such, 'imbalance' between demographic groups is not a meaningful concept in the test data. Furthermore, the distinction between imbalance of identities belonging to a certain demographic group versus that of images per identity in each demographic group has not been carefully studied in either the testing or the training data. All of these facets of imbalance are present in the face *identification* setting, where testing involves matching a probe image to a gallery of many identities, each of which contains multiple images, see Figure 1.

In this work, we unravel the complex effects that dataset imbalance can have on model bias for face identification systems. We separately consider imbalance (both in terms of identities or images per identity) in the train set and in the test set. We also consider the realistic social use case in which a large dataset is collected from an imbalanced population and then split at random, resulting in similar dataset imbalance in both the train and test set. We specifically focus on imbalance with respect to gender presentation. **Our findings show that each type of imbalance has a distinct effect on a model's performance on each gender presentation. Furthermore, in the realistic scenario where the train and test set are similarly imbalanced, the train and test imbalance have the potential to interact in a way that leads to systematic underestimation of the true bias of a model during an audit. Thus any audit of model bias in face identification must carefully control for these effects.**

2 Face Identification Setup

Face recognition has two tasks: face verification and face identification. The first refers to verifying whether a person of interest (called the *probe image*) and a person in a reference photo are the same. This is the setting that might be applied, e.g., to phone unlocking or other identity confirmation. In contrast, face identification involves matching a probe image against a set of images (called the *gallery*) with known identities. This application is relevant to search tasks, such as identifying the subject of a photo from a database of driver's license or mugshot photos.

Our experiments use state-of-the-art face recognition models. We train MobileFaceNet and ResNet-152 feature extractors each with a CosFace and ArcFace heads. For training and evaluation we use the CelebA dataset [25], which provides annotations for gender presentation. We create a balanced training set with equal number of identities and total number of images from each gender presentation, as well as perfectly balanced test set. The identities in the train and test sets are disjoint. We call these the *default train* and *default test* sets. All models are trained with class-balanced sampling to ensure equal contribution of identities to the loss. We additionally include results for models trained without over-sampling in Appendix D.2.

Recall that our research question is to investigate how class imbalances affect face identification. In order to answer this question, we train models on a range of deliberately imbalanced subsamples of the default training set, and test models on a range of deliberately imbalanced subsamples of the default test set. To evaluate the models, we compute rank-1 accuracy over the test set. Specifically, for each test image we treat the rest of the test set as gallery images and find if the closest gallery image in the feature space (as defined by cosine similarity) of a model is an image of the same person.

3 Experimental Setup

First, to explore the effect of train set balance in the number of identities on gender presentation bias, we construct train data splits with different ratios of female and male identities, while ensuring that the average number of images per identity is the same across gender presentations. Then, to explore the effect of train set balance in the number of images per identity we change the average number of images per male and female identity, but fix the number of identities of each gender presentation. In both experiments we evaluate the models on the (perfectly balanced) default test set and report rank-1 face identification accuracy.

Analogous to the train set experiments, we conduct similar experiments with test set (gallery) by splitting the data with different ratios of female and male identities or number of images per identity. We train all models on default balanced train set and evaluate on constructed imbalanced test sets.

4 Balance in the Train Set

Balancing the number of identities. We compute accuracy scores separately for male and female test images for models trained on each of the train splits and depict them in Figure 6 in Appendix D.1 with solid lines. We observe that a higher proportion of male identities in the train set leads to an increase in male accuracy and decrease in female accuracy, with the most significant drops occurring near the extreme $10:0$ imbalance. This indicates that it is very important to have at least a few identities from the target demographic group in the train set; once the representation of the minority group reaches 10%, the marginal gain of additional identities becomes less. Additionally, the accuracy gap is closed for all models when the train set consists of about 10% male and 90% female identities.

Balancing the number of images per identity. The dashed lines in Figure 6 illustrate the accuracy of the models trained on each data split. From the first row plots we see that, similar to the previous experiment, increasing the number of male images in the train set leads to increased accuracy on male and decreased accuracy on female images. Interestingly, we observe a decrease in performance for both demographic groups when the images of that group constitute more than 60% of train data. However, we find that this effect results from the widely used class-balanced sampling training strategy, and models trained without the default oversampling are more robust to imbalance in the number of images per identity, see details in Appendix D.2. The “fair point" where female accuracy is closest to male accuracy occurs when around 20% of images are of males.

5 Balance in the Gallery Set

Balancing the number of identities. We observe that increasing the proportion of identities of a target demographic group in the gallery set hurts the model’s performance on that demographic group, and this trend is similar for male and female images. The results are shown in the solid lines of Figure 7 in Appendix D.1. Intuitively, this is because face recognition models rarely match images to one of a different demographic group; therefore by adding more identities of a particular demographic group, we add more potential false matches for images from that demographic group, which leads to higher error rates.

Balancing the number of images per identity. Unlike the results with identity balance, increasing the average number of images per identity leads to performance gains, since this increases the probability of a match with an image of the same person. Also, image balance affects the performance more significantly than identity balance, and these trends are similar across all the models and both gender presentations. Finally, we note that the “fair point" for image balance in the test set occurs at

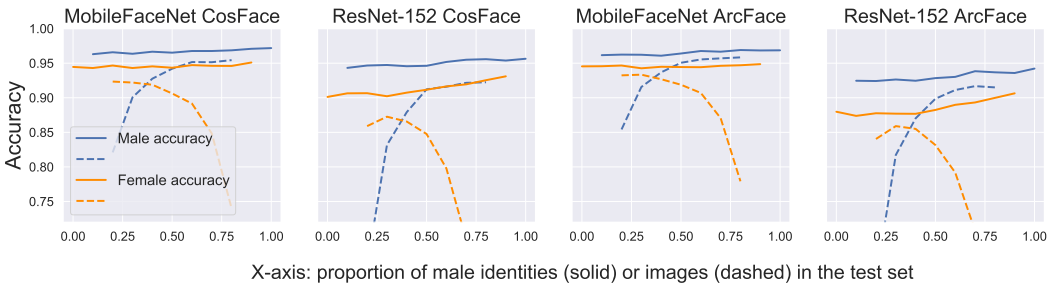

Figure 2: **Train & Test Set Imbalance.** Results of experiments that adjust the gender presentation balance in both the train and test set. Male (blue) and female (orange) accuracy are plotted against the proportion of male data used in both the train and test set.

about 30% male images; contrast this with identity balance, for which no fair point appears to exist. These results are recorded as dashed lines in Figure 7

6 A cautionary tale: matching the balance in the train and gallery data

Using our findings from above, we conclude that common machine learning techniques to create train and test splits can lead to Simpson's paradoxes which lead to a false belief that a model is unbiased. It is standard practice to make random train/test splits of a dataset. If the original dataset is imbalanced, as is commonly the case, the resulting splits will be imbalanced in similar ways. As we have seen above, the effects of imbalance in the train and test splits may oppose one another, causing severe underestimation of model bias when measured using the test split. This occurs because the minority status of a group in the train split will bias the model towards low accuracy on that group, while the correspondingly small representation in the test split will cause an increase in model accuracy, partially or entirely masking the true model bias. The results for these experiments are presented in Figure 2.

Balancing the number of identities We create train and test sets with identical distributions of identities. Recalling the results from prior experiments, increasing the number of identities for the target group in the training stage improves accuracy on that group, while adding more identities in the gallery degrades it. Interestingly, when we increase the proportion of male identities in *both* train and test sets, we observe gains in both male and female accuracy, and that trend is especially strong for ResNet models.

Balancing the number of images per identity Having more images is beneficial in both train and test stages. Therefore, the effect of image balance is amplified when both train and test sets are imbalanced in a similar way. Similar to the train set experiments, having more than 70% female images in both train and test sets leads to slight drops in female accuracy on ResNet models, which again is a result of the default class-balanced oversampling strategy.

7 Bias comparisons

We ask two concluding questions: one about whether class imbalance captures all the inherent bias and the other about how the bias we see compares to human biases.

First, we explore how data imbalances cause biases in random networks. Given a network with random initializations, we would expect that evaluation on a balanced test set would result in equal performance on males and females.However, from Figure 3 we observe that randomly initialized models have higher male performance when the test set is perfectly balanced, and that performance on males is higher when they make up 80% of the test set than female performance when they make up 80% of the test set. This provides strong evidence

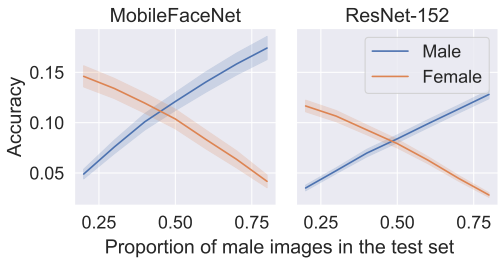

Figure 3: **Random Feature Extractors.** The plot illustrates male (blue) and female (orange) accuracy of random feature extractors against the proportion of male images in the test set. The standard deviation is computed across 10 random initializations.

that there are sources of bias that lie outside what we explore here and which are potential confounders to a thorough study of bias in face identification; further work on this is warranted.

Then, we ask how class imbalances in machines compare to how humans exhibit bias on face identification tasks. Please, see the details of this experiment in Appendix C. We observe that although models and humans both find identifying females more difficult than identifying males, the *particular* males and females that are easier or harder to identify appear to differ between models and humans, which suggests the *reasons* for bias in humans and machines are different.

References

[1] V. Albiero, K. KS, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 81–89, 2020.

[2] M. Alvi, A. Zisserman, and C. Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[3] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[4] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[6] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

[7] N. V. Chawla, N. Japkowicz, and A. Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[8] A. Chouldechova and A. Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[9] C. Dalitz. Reject options and confidence measures for knn classifiers. *Schriftenreihe des Fachbereichs Elektrotechnik und Informatik Hochschule Niederrhein*, 8:16–38, 2009.

[10] D. Danks and A. J. London. Algorithmic bias in autonomous systems. In *IJCAI*, volume 17, pages 4691–4697, 2017.

[11] S. Dooley, T. Goldstein, and J. P. Dickerson. Robustness disparities in commercial face detection. *arXiv preprint arXiv:2108.12508*, 2020.

[12] S. Dooley, R. Downing, G. Wei, N. Shankar, B. Thymes, G. Thorkelsdottir, T. Kurtz-Miott, R. Mattson, O. Obiwumi, V. Cherepanova, et al. Comparing human and machine bias in face recognition. *arXiv preprint arXiv:2110.08396*, 2021.

[13] C. Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[14] P. J. Grother, M. L. Ngan, K. K. Hanaoka, et al. Face recognition vendor test part 3: demographic effects. 2019.

[15] M. Gwilliam, S. Hegde, L. Tinubu, and A. Hanson. Rethinking common assumptions to mitigate racial bias in face recognition datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4123–4132, 2021.

[16] P. Hensman and D. Masko. The impact of imbalanced training data for convolutional neural networks. *Degree Project in Computer Science, KTH Royal Institute of Technology*, 2015.

[17] J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[18] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[19] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[20] B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[21] H. Lee, M. Park, and J. Kim. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In *2016 IEEE international conference on image processing (ICIP)*, pages 3713–3717. IEEE, 2016.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[23] C. X. Ling and V. S. Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011:231–235, 2008.

[24] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[26] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[27] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. Frvt 2006 and ice 2006 large-scale experimental results. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):831–846, 2009.

[28] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11, 2011.

[29] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen, et al. Dynamic sampling in convolutional neural networks for imbalanced data classification. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 112–117. IEEE, 2018.

[30] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.

[31] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, 40(12):3358–3378, 2007.

[32] H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2, 2019.

[33] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942, 2007.

[34] M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020.

[35] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy. Training deep neural networks on imbalanced data sets. In *2016 international joint conference on neural networks (IJCNN)*, pages 4368–4374. IEEE, 2016.

[36] Y. Zhang and W. Deng. Class-balanced training for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 824–825, 2020.

A X-Risk Sheet

Individual question responses do not decisively imply relevance or irrelevance to existential risk reduction. Do not check a box if it is not applicable.

A.1 Long-Term Impact on Advanced AI Systems

In this section, please analyze how this work shapes the process that will lead to advanced AI systems and how it steers the process in a safer direction.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?
   Answer: While modern face recognition systems advanced significantly in the last decades, it is known that these systems have disparate performance on groups of subjects sharing particular sensitive attributes, such as gender, race or age, leading to massive distrust of these systems and an unwillingness to deploy them in many cases.

   One of the major sources of bias in FR systems is dataset imbalance. This work sheds light on the effects of different types of dataset imbalance on bias in face identification systems, which are widely deployed in sensitive applications. We hope that practitioners will take into account our findings when performing bias audits in face identification systems. Our findings demonstrate that each type of imbalance has a distinct effect on a model's performance on each sensitive group. Furthermore, we demonstrate how train and gallery set imbalance interact in a way that leads to systematic underestimation of the true bias. Finally, we compare class imbalance in machines with human bias on face recognition tasks and show that class imbalance does not capture all the inherent bias in face recognition.

2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?
   Answer: Our work reduces the potential for harm caused by neural network bias by (1) showing how to better audit systems and (2) showing that a proper examination of such systems must reach beyond dataset imbalances as the systems are inherently biased beyond their training data.

3. **Diffuse Effects.** If this work reduces existential risks indirectly or diffusely, what are the main contributing factors that it affects?
   Answer: The findings in our work suggest that researchers should examine architectures for their biases as even randomly initialized architectures exhibit bias.

4. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?
   Answer: Face Identification systems are widely used in critical applications, such as law enforcement or mass surveillance. It has been shown that these systems are more error-prone on images of female subjects and darker-skin subjects, and this bias has already been shown to have caused the wrongful incarceration of innocent black men in the US. Our findings can be used to audit models, which are under consideration for deployment, for bias to make sure that minority populations are not inadvertently incarcerated because of flawed ML systems.

5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters?
   Answer: Our experiments are conducted on realistic face recognition pipelines with large-scale pre-trained backbones and with relevant fairness metrics. Our findings are not sensitive to hyperparameters whatsoever. □

6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task?
   Answer: Machine face recognition systems can already outperform humans, both in terms of efficiency and accuracy. □

7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability?
   Answer: Auditing face recognition systems as we do requires data with sensitive attribute

annotations, for example gender labels. Fortunately, such audits can be conducted using existing datasets as we do and do not need to be conducted using the model's training set. ☒

8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility?
Answer: Under some definitions of fairness, there may be a trade-off between accuracy and fairness. However, there is no reason to believe that our current face recognition systems are on the Pareto frontier of this trade-off. ☐

A.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

1. **Overview.** How does this improve safety more than it improves general capabilities?
Answer: This work does not in fact improve general capabilities whatsoever and is only focused on mitigating bias. On the other hand, future work may focus on developing new architectures which are simultaneously accurate and fair.
2. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
Answer: The fairness auditing experiments included in this work do not improve general capabilities, but nonetheless, any non-exhaustive auditing of industrial or government systems may lead the organizations and the public into a false sense of security and bring on deployment. Thus, it is important that such audits be thorough as in this work.
3. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research?
Answer: No, our work only proposes techniques for auditing fairness in face recognition systems, not for improving their accuracy. ☐
4. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities?
Answer: Our work does not improve any of the above capabilities. ☐
5. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment?
Answer: Fairness is not known to be predicted by "general cognitive ability" and in fact can come with provable accuracy trade-offs for some definitions of fairness. For other definitions, the trade-off is not well understood. ☐
6. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI?
Answer: Our work does not advance other capabilities outside of fairness. ☐

A.3 Elaborations and Other Considerations

1. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?
Answer: Please, see our detailed responses to the questions above.

B Appendix

B.1 Broader impact and limitations

In this work we explore the effects of various forms of data imbalance on bias in face identification, and we hope that practitioners will take our findings into account when performing bias auditing. However, it is important to understand that biases can originate from various sources besides data imbalance and therefore models should be carefully evaluated for other bias factors.

The availability of high quality datasets, which are suitable for the identification task (as opposed to verification), have demographic metadata/annotation for both train and test sets, and contain enough identities and images belonging to each demographic group to allow for subsampling, is extremely limited. For this reason we focus solely on gender bias and leverage CelebA dataset, which meets these criteria. Also, using a binary demographic attribute (such as gender, when restricting only to male- and female-presenting identities) allows the proportion of data in each group to be conveniently tuned as a single parameter, which in turn makes interpreting results more straightforward.

B.2 Related Work

B.2.1 Imbalance in verification

Even before the advent of neural network-based face recognition systems, researchers have studied how the composition of training data affects verification performance. Phillips et al. [28] compared algorithms from the Face Recognition Vendor Test [27] and found that those developed in East Asia performed better on East Asian Faces, and those developed in Western countries performed better on Caucasian faces. Klare et al. [19] expanded on these results by comparing performance across race, gender presentation, and age cohorts, observing that training exclusively on images of one demographic group improved performance on that group and decreased performance on the others. They further conclude that training on data that is “well distributed across all demographics" helps prevent extreme bias.

Multiple verification datasets have been proposed in the interest of eliminating imbalance as a source of bias in face verification. The *BUPT-BalancedFace* dataset [34] contains an approximately equal number of identities and images of four racial groups[2]. *Balanced Faces in the Wild* [30] goes a step further, balancing identities and images across eight categories of race-gender presentation combinations. Also of note is the *BUPT-CBFace* dataset [36], which is class-balanced (each identity possesses the same number of images), rather than demographically balanced.

Some recent work in verification has questioned whether perfectly balanced training data is in fact an optimal setting for reducing bias. Albiero et al. [1] studied sources of bias along gender presentation; among their findings, they observe that balancing the amount of male and female training images and identities in the training data reduces, but does not eliminate, the performance gap between gender presentations. Similarly, Gwilliam et al. [15] trained models on data with different racial makeups, finding that models which were trained with more images of African subjects had lower variance in performance on each race than those which were trained on balanced data.

B.2.2 Bias in Identification

Although the effect of imbalance on bias has only been explicitly studied in face verification, there is some research on identification which is relevant. The National Institutes of Standards and Technology performed large-scale testing of commercial identification algorithms, finding that many (though not all) exhibit gender presentation or racial bias [14]. The evaluators speculate that the training data or procedures contribute to this bias, but could not study this hypothesis due to the proprietary nature of the models. Dooley et al. [12] evaluated commercial and academic models on a variant of identification in which each probe image is compared to 9 gallery images of distinct identities, but belonging to the same skin type and gender presentation. They find that academic models (and some, but not all, commercial models) exhibit skin type and gender presentation bias despite a testing regime which makes imbalance effectively irrelevant.

[2]This work also introduces *BUPT-GlobalFace*, which instead approximately matches the distribution across races to that of the world population.

B.2.3 Imbalance in Deep Learning

Outside the realm of facial recognition, there is much study about the impacts of class imbalance in deep learning. In standard machine learning techniques, i.e., non-deep learning, there are many well-studied and proven techniques for handling class imbalances like data-level techniques [33, 5, 7], algorithm-level methods [13, 23, 20], and hybrid approaches [6, 31, 24]. In deep learning, some take the approach of random over or under sampling [16, 21, 29]. Other methods adjust the learning procedure by changing the loss function [35] or learning cost-sensitive functions for imbalanced data [18]. We refer the reader to Buda et al. [3], Johnson and Khoshgoftaar [17], for a thorough review of deep learning-based imbalance literature. Much of the class-imbalance work has been on computer vision tasks, though generally has not examined specific analyses like we present in this work like network initialization, face identification, or intersectional demographic imbalances.

B.2.4 Other sources of bias in facial recognition

Face recognition is a complex, sociotechnical system where biases can originate from the algorithms [10], preprocessing steps [11], and human interpretations [8]. While we do not explicitly examine these sources, we refer the reader to Mehrabi et al. [26], Suresh and Guttag [32] for a broader overview of sources of bias in machine learning.

B.3 Training details

We pre-process CelebA images by aligning them using the provided facial landmarks and cropping to 112x112 size. All face recognition models are trained with Focal loss [22] using SGD for 100 epochs with learning rate of 0.1, momentum of 0.9 and weight decay of 5e-4. The learning rate is reduced by 10 times at epochs 35, 65 and 95. Horizontal flip data augmentation is used during training. For the model architectures, we use implementation from publicly available github repository `face.evoLVe.PyTorch`[3]. We run our experiments on NVIDIA GeForce RTX 2080 Ti machines and each experiment takes from 6 to 12 hours of compute time on one GPU.

Table 1: Details on the number of identities, total number of images and average number of images per identity used in experiments with train and test data balance. We also report statistics for the default train and test sets. M denotes male, F denotes female.

Setting	M ids	F ids	Total M imgs	Total F imgs	M imgs/id	F imgs/id	Total ids	Total imgs
Train default	3967	3967	70k	70k	17.65	17.65	7934	140k
Train id balance	0 - 3967	0 - 3967	0 - 70k	0 - 70k	17.65	17.65	3967	70k
Train img balance	3967	3967	14k - 56k	14k - 56k	3.53 - 14.11	3.53 - 14.11	7934	70k
Test default	406	406	7k	7k	17.24	17.24	812	14k
Test id balance	0 - 406	0 - 406	0 - 7k	0 - 7k	17.24	17.24	406	7k
Test img balance	406	406	1.4k - 5.6k	1.4k - 5.6k	3.45 - 13.80	3.45 - 13.80	812	7k

C Bias comparisons

C.1 Are models biased like humans?

Numerous psychological and sociological studies have identified gender, racial, and other biases in human performance on face recognition tasks. Dooley et al. [12] studied whether humans and FR models exhibit similar biases. They evaluated human and machine performance on the curated InterRace test questions, and found models indeed tend to perform with comparable gender presentation bias ratios to humans.

In this section, we use their human survey data to explore two related questions: how correlated are model and human performance *at the question level*, and how does this change with different levels of imbalance in training data? To answer these questions, we define a metric which allows us to distinguish how well a model performs on each InterRace identification question. Let

$$L_2 \text{ ratio} = \frac{\|v_{probe} - v_{false}\|_2}{\|v_{probe} - v_{true}\|_2 + \|v_{probe} - v_{false}\|_2},$$

[3] `https://github.com/ZhaoJ9014/face.evoLVe`

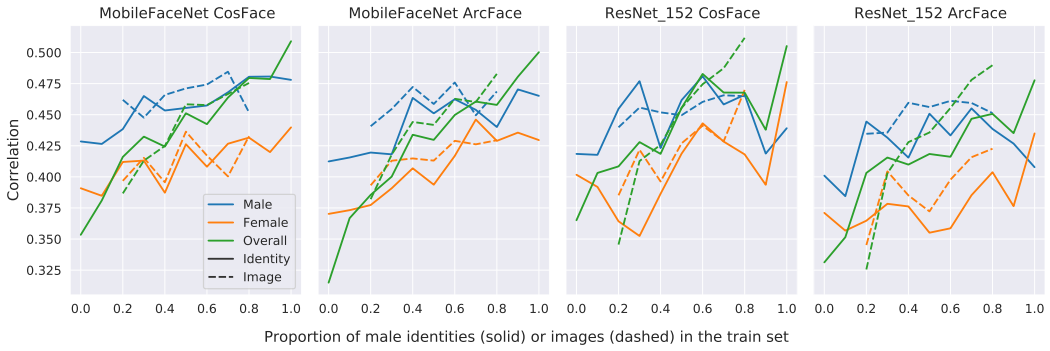

Figure 4: Pearson correlation of $L2$ ratio vs. human accuracy for various models as proportion of male training data varies.

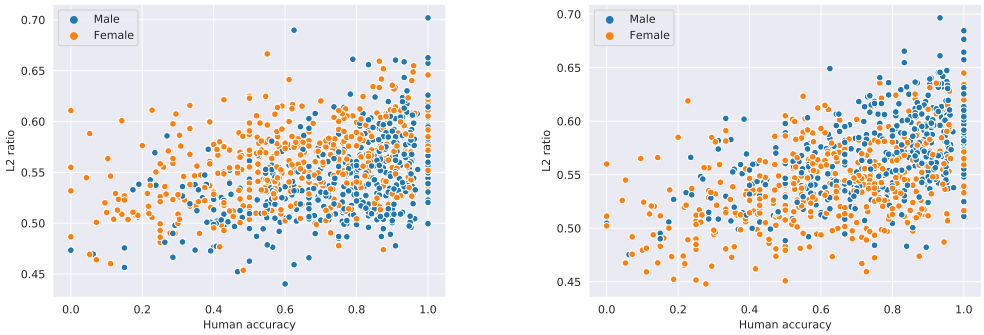

Figure 5: Scatterplots of model L2 ratio vs. human accuracy on each question in the InterRace identification dataset. Both models are MobileFaceNets trained with CosFace loss. (Left) a model trained on exclusively female images. (Right) a model trained on exclusively male images.

where v_{probe}, v_{true}, v_{false} are the feature representations of the probe image, the correct gallery image, and the nearest incorrect gallery image, respectively.[4] Figure 5 depicts examples of scatterplots comparing model confidence to human accuracy on each InterRace question.

Figure 4 shows the correlation between $L2$ ratio and human performance for various models at each of the training imbalance settings that we have considered in earlier experiments. We see that the correlation between these values over *all* questions tends to rise as the proportion of male training data increases. However, the correlation when separately considering male and female questions does not rise as monotonically, or as much, from left to right as the overall correlation does. This suggests that the correlation between human and machine performance is largely driven by the fact that models and humans both find identifying females more difficult than identifying males, and that this disparity is exacerbated when the model in question is trained on male-dominated data. On the other hand, the *particular* males and females that are easier or harder to identify appear to differ between models and humans, which suggests the *reasons* for bias in humans and machines are different.

D Additional Plots

D.1 Plots for train set and gallery set imbalance.

Figure 6 illustrates performance of the models trained with different ratios of gender presentations in the train set and Figure 7 depicts performance of the models trained on perfectly balanced train set, but tested on a gallery set with different ratios of gender presentations.

[4]We note that other measures of confidence in a k-nearest neighbors setting, such as those discussed in [9], are inappropriate for this application.

Figure 6: **Train Set Imbalance.** Results of experiments that change the train set gender presentation balance. Male (blue) and female (orange) accuracy are plotted against the proportion of male data in the train set. All models are tested on the default balanced test set.

Figure 7: **Test Set Imbalance.** Results of experiments that change the test set gender presentation balance. Male (blue) and female (orange) accuracy are plotted against the proportion of male data in the test set. All models are trained on the default balanced train set.

D.2 Results for models trained without class-balanced sampling.

To explore the effect of class-balanced sampling on the results of our experiments, we train additional models without any oversampling strategies. Figures 8 - 10 show results of our experiments for models trained without oversampling. We find that most trends are similar to ones observed in the models trained with class-balanced sampling, however models trained without oversampling are more robust to balance in the number of images per identity, see Figure 8. In particular, the effect of balancing the number of images (dashed lines) is similar to the effect of balancing the number of identities (solid lines) for all models, but ResNet-152 trained with ArcFace head. This leads us to a conclusion that using class-balanced sampling strategy is not beneficial in scenarios of severe imbalance in number of images per identity in face recognition models.

Figure 8: **Train Set Imbalance** Results of experiments that change the train set gender presentation balance for models trained **without class-balanced sampling**. Male (blue) and female (orange) accuracy are plotted against the proportion of male data in the train set. All models are tested on the default balanced test set.

Figure 9: **Test Set Imbalance.** Results of experiments that change the test set gender presentation balance for models trained **without class-balanced sampling**. Male (blue) and female (orange) accuracy are plotted against the proportion of male data in the test set. All models are trained on the default balanced train set.

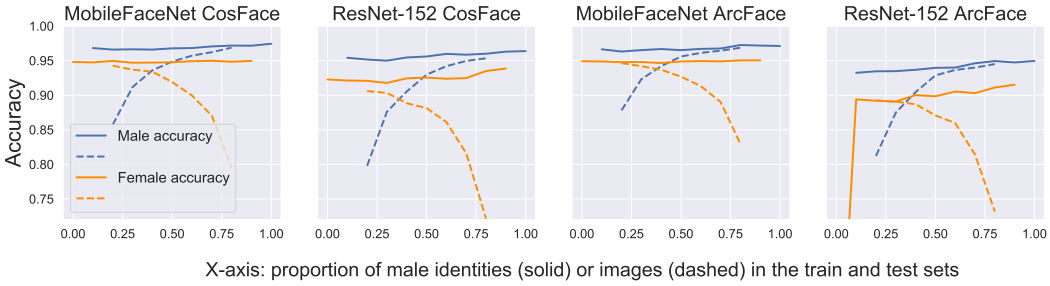

Figure 10: **Train & Test Set Imbalance.** Results of experiments that adjust the gender presentation balance in both the train and test set for models trained **without class-balanced sampling**. Male (blue) and female (orange) accuracy are plotted against the proportion of male data in the train and test set.