On the Importance of Architectures and Hyperparameters for Fairness in Face Recognition

Rhea Sukthanker[1,*], **Samuel Dooley**[2,*], **John P. Dickerson**[2,3], **Colin White**[4], **Frank Hutter**[1,5]
Micah Goldblum[6]

[1]University of Freiburg, [2]University of Maryland, [3]ArthurAI, [4]Abacus.AI,
[5]Bosch Center for AI, [6]New York University

Abstract

Face recognition systems are used widely but are known to exhibit bias across a range of sociodemographic dimensions, such as gender and race. An array of works proposing pre-processing, training, and post-processing methods have failed to close these gaps. Here, we take a very different approach to this problem, identifying that both architectures and hyperparameters of neural networks are instrumental in reducing bias. We first run a large-scale analysis of the impact of architectures and training hyperparameters on several common fairness metrics and show that the implicit convention of choosing high-accuracy architectures may be suboptimal for fairness. Motivated by our findings, we run the first neural architecture search for fairness, jointly with a search for hyperparameters. We output a suite of models which Pareto-dominate all other competitive architectures in terms of accuracy and fairness. Furthermore, we show that these models transfer well to other face recognition datasets with similar and distinct protected attributes. We release our code and raw result files so that researchers and practitioners can replace our fairness metrics with a bias measure of their choice.

1 Introduction

Face recognition is regularly deployed across the world by government agencies for tasks including surveillance, employment, and housing decisions. However, recent studies have shown that face recognition systems exhibit disparity in accuracy based on race and gender [1, 2, 3, 4]. While existing methods for de-biasing face recognition systems use a fixed neural network architecture and training hyperparameters, we instead ask a fundamental question which has received little attention: *does model-bias stem from the architecture and hyperparameters?* We further exploit the extensive research in the fields of neural architecture search (NAS) [5] and hyperparameter optimization (HPO) [6] to search for models that achieve a desired trade-off between bias and accuracy.

In this work, we take the first step towards answering these questions. To this end, we conduct the first large-scale analysis of the relationship between hyperparameters, architectures, and bias. We train a diverse set of 29 architectures, ranging from ResNets [7] to vision transformers [8, 9] to Gluon Inception V3 [10] to MobileNetV3 [11] on CelebA [12], for a total of 88 493 GPU hours. We train each of these architectures across different head, optimizer, and learning rate combinations. Our results show that different architectures learn different inductive biases from the same dataset. We conclude that the implicit convention of choosing the highest-accuracy architectures can be detrimental to fairness, and suggest that architecture and hyperparameters play a significant role in determining the fairness-accuracy tradeoff.

Next, we exploit this observation in order to design architectures with a better fairness-accuracy tradeoff. We initiate the study of NAS for fairness; specifically, we run NAS+HPO to jointly

[*]Equal contribution. Email to: sukthank@cs.uni-freiburg.de, sdooley1@cs.umd.edu.

2022 Trustworthy and Socially Responsible Machine Learning (TSRML 2022) co-located with NeurIPS 2022.

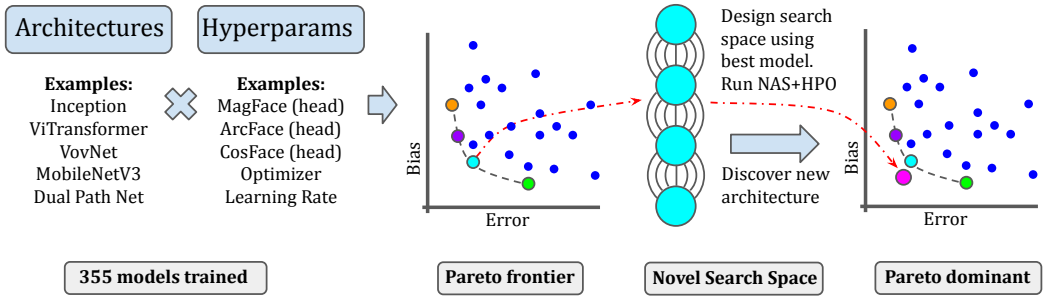

Figure 1: Overview of our methodology.

optimize fairness and accuracy. To tackle this problem, we construct a search space based on the highest-performing architecture from our analysis. We use the Sequential Model-based Algorithm Configuration method (SMAC [13]), for multi-objective architecture and hyperparameter search. We discover a Pareto frontier of face recognition models that outperform existing state-of-the-art models on both accuracy and multiple fairness metrics. An overview of our methodology can be found in Figure 1. We release all of our code and raw results at https://github.com/dooleys/FR-NAS so that users can adapt our work to any bias measure of their choice.

Our contributions. We summarize our main contributions below:

- We run a large-scale study of 29 architectures from ViT to Xception, each trained across a variety of hyperparameters, totalling 88 493 GPU hours. Our analysis shows that there is a distinct trade-off between accuracy and popular fairness metrics, such as disparity, and simply improving accuracy would not guarantee improvement on different fairness metrics.
- Motivated by the above observation, we conduct the first neural architecture search for fairness, jointly with hyperparameter optimization and optimizing for accuracy — culminating in a set of architectures which Pareto-dominate all models in a large set of modern architectures
- We show that the architectures discovered transfer across different datasets with the same (perceived gender) and different (ethnicities) protected attributes.

Background and related work. Face recognition tasks fall into two categories: verification and identification. *Verification* asks whether the person in a source image is the same person as in the target image; this is a one-to-one comparison. *Identification* instead asks whether a given person in a source image appears within a gallery composed of many target identities and their associated images; this is a one-to-many comparison. Novel techniques in face recognition tasks [14, 15, 16] use deep networks to extract feature representations of faces and then compare those to match individuals (with mechanisms called the *head*). We focus our analysis on identification and on examining how close images of similar identities are in the feature space of trained models.

In this work, we focus on *measuring* sociodemographic disparities across neural architectures and hyperparameter settings, and finding the Pareto frontier of face recognition performance and bias for current and novel architectures. Our work searches for architectures and hyperparameters which improve undesired disparities. A few works have applied hyperparameter optimization to mitigate bias in models for tabular data. Perrone et al. [17] recently introduced a Bayesian optimization framework to optimize accuracy while satisfying a bias constraint. The concurrent works of Schmucker et al. [18] and Cruz et al. [19] extend Hyperband [20] to the multi-objective setting and apply it to fairness. To the best of our knowledge, no prior work uses any AutoML technique (NAS, HPO, or joint NAS and HPO) to design fair face recognition models, and no prior work uses NAS to design fair models for any application. For additional related work, see Appendix A.

2 A Large-Scale Analysis of Architectures and Fairness

Experimental Setup. We train and evaluate each model configuration on a gender-balanced subset of the CelebA dataset [12]. While this work analyzes phenotypal metadata (perceived gender), the reader should not interpret our findings as a social lens of what these demographic groups mean inside society. We guide the reader to Hamidi et al. [21] and Keyes [22]. We use the following training

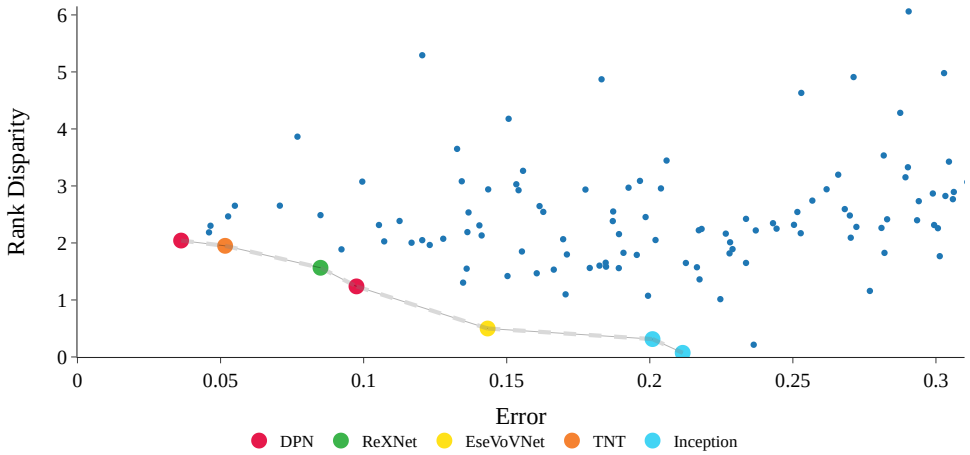

Figure 2: Error-Rank Disparity Pareto front of the architectures with lowest error (< 0.3). Models in the lower left corner are better. The Pareto front is notated with a dashed line. Other points are architecture and hyperparameter combinations which are not Pareto-optimal. DPN, ReXNet, EseVovNet, TNT, and Inception architectures are Pareto-optimal.

pipeline – ultimately conducting 355 training runs with different combinations of 29 architectures from the Pytorch Image Model (`timm`) database [23] and hyperparameters. For each model, we use the default learning rate and optimizer that was published with that model. We then conduct a training run with these hyperparameters and each of three heads, ArcFace [14], CosFace [15], and MagFace [16]. Next, we use that default learning rate with both AdamW [24] and SGD optimizers (again with each head). Finally, we also conduct training routines with AdamW and SGD with unifed learning rates. In total, we run a single architecture between 9 and 13 times. All other hyperparameters were the same for each model training run.

Evaluation procedure.

We evaluate performance via *Error* and use a common fairness metric in face recognition, *rank disparity*, which is explored in the NIST FRVT [25]. To compute the rank of a given sample, we ask how many images of a different identity are closer to it in feature space. We define this index as the *Rank* of a given image. Thus, *Rank(image)* $= 0$ if and only if *Error(image)* $= 0$; *Rank(image)* > 0 if and only if *Error(image)* $= 1$. We examine the **rank disparity** which is the absolute difference of the average ranks for each perceived gender in a dataset $\mathcal{D}$:

$$\text{Rank Disparity} = \left| \frac{1}{|\mathcal{D}_{\text{male}}|} \sum_{x \in \mathcal{D}_{\text{male}}} \text{Rank}\,(x) - \frac{1}{|\mathcal{D}_{\text{female}}|} \sum_{x \in \mathcal{D}_{\text{female}}} \text{Rank}(x) \right|.$$

Results and Discussion. By plotting the performance of each training run with the error on the x-axis and rank disparity on the y-axis in Figure 2, we can easily conclude two main points. First, optimizing for error does not also optimize for fairness, and second, different architectures have different fairness properties.

On the first point, a search for architectures and hyperparameters which have high performance on traditional metrics does not translate to high performance on fairness metrics. We see that within models with lowest error – those models which are most interesting to the community – there is low correlation between error and rank disparity ($\rho = -.113$ for models with error < 0.3). In Figure 2, we see that Pareto optimal models are versions of DPN, TNT, ReXNet, VovNet, and ResNets (in increasing error and decreasing fairness). We conclude that both architectures and hyperparameters play a significant role in determining the accuracy and fairness trade-off, motivating their joint optimization in Section 3.

Additionally, we observe that the Pareto curve is dependent upon what fairness metric we consider. For example, in Figure 3, we demonstrate that a very different set of architectures are Pareto optimal

Figure 3: Depending on the fairness metric, different architectures are Pareto-optimal. On the left, we plot the metric Ratio of Ranks which admit DPN, ReXNet, HRNet, MobileNet, EseVovNet, and Inceptions as Pareto-optimal. On the right, we plot the metric Ratio of Errors where DPN, ReXNet, EseVovNet, ResNet-RS, and VGG19 are Pareto-optimal.

if instead of rank disparity (rank difference between perceived genders) we consider the ratio of ranks between the two perceived genders or the ratio of the errors. Specifically, on the ratio of ranks metric, the Pareto frontier contains versions of HRNet, MobileNet, VovNet, and ResNet whereas the Pareto frontier under the ratio of errors metric includes versions of NesT, ResNet-RS, and VGG19.

Further, different architectures exhibit different optimal hyperparameters. For example, the Xception65 architecture finds SGD with ArcFace and AdamW with ArcFace are Pareto-optimal whereas the Inception-ResNet architecture finds MagFace and CosFace optimal with SGD. This illustrates the care that needs to be taken when choosing a model – optimizing architectures and hyperparameters for error alone will not lead to fair models.

Finally, existing architectures and hyperparameters do not yield models which simultaneously exhibit both low error and low disparity. For example, in Figure 2 there is a significant area under the Pareto curve. While there are models with very low error, in order to improve the disparity metric, one must sacrifice significant performance. However, in Section 3, we will see that our joint NAS+HPO experiments for rank disparity ultimately find a model convincingly in the area to the left of this Pareto curve – that is, we find a model with low error *and* disparity.

3 Joint NAS+HPO for Fairness

In this section, we employ joint NAS+HPO to find better architectures. Inspired by our findings on the importance of architectures and hyperparameters for fairness in Section 2, we initiate the first joint NAS+HPO study for fairness in face recognition. We start by describing our search space and search strategy. We then present a comparison between the architectures obtained from multi-objective joint NAS+HPO and the handcrafted image classification models studied in Section 2. We conclude that our joint NAS+HPO indeed discovers simultaneously accurate and fair architectures.

3.1 Search Space Design

We design our search space based on our analysis in Section 2 namely Dual Path Networks [26] due to its strong trade-off between rank disparity and accuracy as seen in Figure 2. We choose three categories of hyperparameters for NAS+HPO: architecture head/loss, optimizer, and learning rate.

Architecture Search Space Design. Dual Path Networks [26] for image classification share common features (ResNets [27]) while possessing the flexibility to explore new features [28] through a dual path architecture. We replace the repeating `1x1_conv-3x3_conv-1x1_conv` block with a

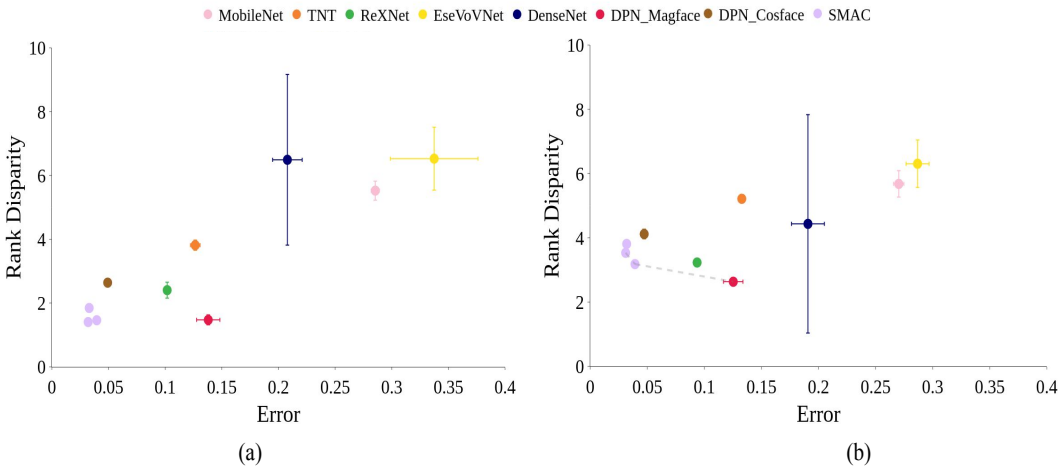

Figure 4: Pareto front of models discovered by SMAC and rank-1 models from `timm` for *(a)* validation and *(b)* test sets on CelebA averaged over 4 seeds.SMAC models Pareto-dominate top performing `timm` models ($Error < 0.1$).

simple recurring searchable block. Furthermore, we stack multiple such searched blocks to closely follow the architecture of Dual Path Networks. We have nine possible choices for each of the three operations in the DPN block. The choices include a vanilla convolution, a convolution with pre-normalization and a convolution with post-normalization. To summarize, our search space consists of a choice among 81 different architecture types, 3 different head types, 3 different optimizers (discrete hyperparameters) and a possibly infinite number of choices for the continous learning rate.

3.2 Search Strategy

We navigate the search space defined in Section 3.1 using Black-Box-Optimization (BBO). We want our BBO algorithm to support the following important techniques:

Multi-fidelity optimization. A single function evaluation for our use-case corresponds to training a deep neural network on a given dataset. This is generally quite expensive for traditional deep neural networks on moderately large datasets. Hence we would like to use cheaper approximations to speed up the black-box algorithm with multi-fidelity optimization techniques [29, 20, 30], e.g., by evaluating many configurations based on short runs with few epochs and only investing more resources into the better-performing ones.

Multi-objective optimization. We want to observe a trade-off between the accuracy of the face recognition system and the fairness objective of choice (rank disparity). Hence, our joint NAS+HPO algorithm needs to support multi-objective optimization [31, 32, 33]. The SMAC3 package [13] offers a robust and flexible framework for Bayesian Optimization with few evaluations. SMAC3 offers a SMAC4MF facade for *multi-fidelity optimization* to use cheaper approximations for expensive deep learning tasks like ours. We choose ASHA [29] for cheaper approximations with the initial and maximum fidelities set to 25 and 100 epochs, respectively, and $\eta = 2$. Every architecture-hyperparameter configuration evaluation is trained using the same training pipeline as in Section 2. For the sake of simplicity, we use ParEGO [32] for *multi-objective optimization* with ρ set to 0.05.

3.3 Results

We follow the evaluation scheme of Section 2 to compare models discovered by joint NAS+HPO with the handcrafted models. In Figure 4, we compare the set of models discovered by joint NAS+HPO vs. the models on the Pareto front from Section 2. We train each model for 4 seeds to study the robustness of error and disparity. As seen in Figure 4, we Pareto-dominate all other models with above random accuracy on the validation set. On the test set, we still Pareto-dominate all competitive models (with $Error < 0.1$), but due to differences between the two dataset splits, one of the original configurations (DPN with Magface) also becomes Pareto-optimal. However, the error of this architecture is 0.13, which is significantly higher than the the best original model (0.05) and the SMAC models (0.03-0.04). Furthermore, from Figure 4 it is also apparent that some models such as VoVNet and DenseNet show

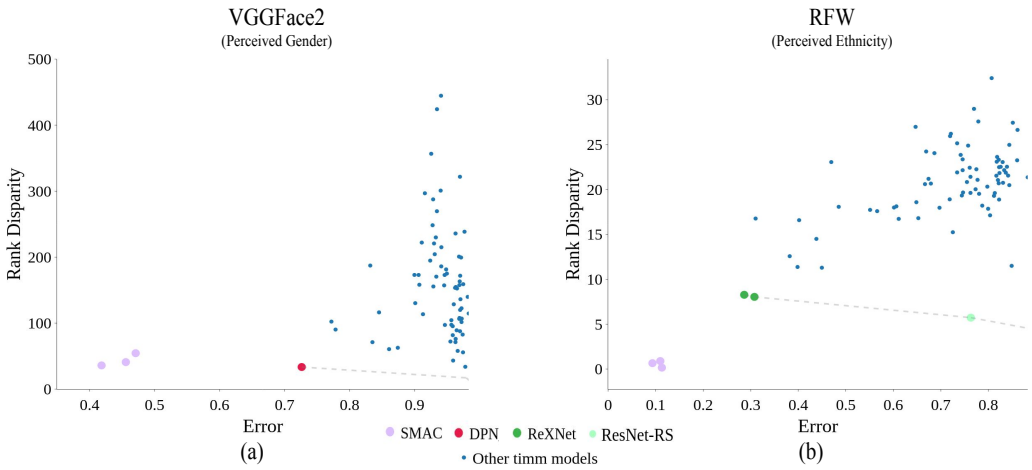

Figure 5: Pareto front of the models on *(a)* the VGGFace2 test set with perceived gender as the protected attribute and *(b)* the RFW test set with perceived ethnicity as the protected attribute. The SMAC models discovered by joint NAS+HPO Pareto-dominate the `timm` models.

very large standard errors across seeds. Hence, it becomes very important to also study the robustness of the models across seeds along with the accuracy and disparity Pareto front.

3.3.1 Transfer across Face Recognition Datasets

Inspired by our findings on the CelebA dataset, we now study the accuracy-disparity trade-off of the models studied in Section 2 and the searched models from Section 3 on two different datasets. The first face recognition dataset we use is VGGFace2 [34], which is based on the same protected attribute (perceived gender) that has served as the focus of our study. The second dataset, Racial Faces in the Wild (RFW) [35], consists of four different racial identities: Caucasian, Indian, Asian, and African. We compute the rank disparity within different *ethnicities*, i.e., a different attribute than the *perceived gender* studied in previous sections. With this dataset, we aim to study the generalization of the fair representations learned by the models across a different protected attribute. However, we caution the reader that the labels of these datasets rely on socially constructed concepts of gender presentation and ethnicity. The intention here is to study how the models discovered by SMAC generalize to these datasets and compare to the other handcrafted `timm` [23] architectures.

To evaluate our models on these datasets, we directly transfer our models to the two test sets. That is, we use the models trained on CelebA, without re-training or fine-tuning the models on the new datasets. As observed in Figure 5, the models discovered using joint NAS+HPO still remain Pareto-optimal on both datasets. In the case of VGGFace2, the models found by SMAC are the only ones to have an error below 0.5, where the next-best model has an error above 0.7. In the case of RFW, the models found by SMAC have considerably lower rank disparity *and* error than the standard models studied in Section 2. This might be due to the optimized architectures learning representations that are intrinsically fairer than those of standard architectures, but it requires further study to test this hypothesis and determine in precisely which characteristics these architectures differ.

4 Conclusion and Future Work

We conducted the first large-scale analysis of the relationship among hyperparameters and architectural properties, and accuracy, bias, and disparity in predictions. We expect the future work in this direction to focus on studying different multi-objective algorithms [36, 37] and NAS techniques [38, 39, 40] to search for inherently fairer models. Further, it would be interesting to study how the properties of the architectures discovered translate across different demographics and populations. Another potential direction of future work is including priors and beliefs about fairness in the society from experts to further improve and aid NAS+HPO methods for fairness by integrating expert knowledge. Finally, given the societal importance of fairness, it would be interesting to study how our findings translate to real-life face recognition systems under deployment.

Acknowledgments

This research was partially supported by the following sources: NSF CAREER Award IIS-1846237, NSF D-ISN Award #2039862, NSF Award CCF-1852352, NIH R01 Award NLM-013039-01, NIST MSE Award #20126334, DARPA GARD #HR00112020007, DoD WHS Award #HQ003420F0035, ARPA-E Award #4334192; TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215; the German Federal Ministry of Education and Research (BMBF, grant RenormalizedFlows 01IS19077C); the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 417962828; the European Research Council (ERC) Consolidator Grant "Deep Learning 2.0" (grant no. 101045765). Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the ERC. Neither the European Union nor the ERC can be held responsible for them.

References

[1] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology, 2019.

[2] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[3] Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. *arXiv preprint arXiv:2102.00813*, 2021.

[4] Erik Learned-Miller, Vicente Ordóñez, Jamie Morgenstern, and Joy Buolamwini. Facial recognition technologies in the wild, 2020.

[5] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[6] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[13] Marius Lindauer, Katharina Eggensperger, Matthias Feurer, André Biedenkapp, Difan Deng, Carolin Benjamins, Tim Ruhkopf, René Sass, and Frank Hutter. Smac3: A versatile bayesian

optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23(54):1–9, 2022. URL http://jmlr.org/papers/v23/21-0888.html.

[14] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[16] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.

[17] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 854–863, 2021.

[18] Robin Schmucker, Michele Donini, Valerio Perrone, Muhammad Bilal Zafar, and Cédric Archambeau. Multi-objective multi-fidelity hyperparameter optimization with application to fairness. In *NeurIPS Workshop on Meta-Learning*, volume 2, 2020.

[19] André F Cruz, Pedro Saleiro, Catarina Belém, Carlos Soares, and Pedro Bizarro. A bandit-based algorithm for fairness-aware hyperparameter optimization. *arXiv preprint arXiv:2010.03665*, 2020.

[20] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[21] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018.

[22] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

[23] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[25] Patrick J. Grother, George W. Quinn, and P J. Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency/Internal Report (NISTIR)*, 2010. URL https://doi.org/10.6028/NIST.IR.7709.

[26] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. *Advances in neural information processing systems*, 30, 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

[29] Robin Schmucker, Michele Donini, Muhammad Bilal Zafar, David Salinas, and Cédric Archambeau. Multi-objective asynchronous successive halving. *arXiv preprint arXiv:2106.12639*, 2021.

[30] Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning*, pages 1437–1446. PMLR, 2018.

[31] Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence*, pages 766–776. PMLR, 2020.

[32] Joan Davins-Valldaura, Saïd Moussaoui, Guillermo Pita-Gil, and Franck Plestan. Parego extensions for multi-objective optimization of expensive evaluation functions. *Journal of Global Optimization*, 67(1):79–96, 2017.

[33] Gong Mao-Guo, Jiao Li-Cheng, Yang Dong-Dong, and Ma Wen-Ping. Evolutionary multi-objective optimization algorithms. *Journal of Software*, 20(2), 2009.

[34] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[35] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[36] HC Fu and P Liu. A multi-objective optimization model based on non-dominated sorting genetic algorithm. *International Journal of Simulation Modelling*, 18(3):510–520, 2019.

[37] Marco Laumanns and Jiri Ocenasek. Bayesian optimization algorithms for multi-objective optimization. In *International Conference on Parallel Problem Solving from Nature*, pages 298–307. Springer, 2002.

[38] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[39] Arber Zela, Thomas Elsken, Tonmoy Saikia, Yassine Marrakchi, Thomas Brox, and Frank Hutter. Understanding and robustifying differentiable architecture search. *arXiv preprint arXiv:1909.09656*, 2019.

[40] Colin White, Willie Neiswanger, and Yash Savani. Bananas: Bayesian optimization with neural architectures for neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10293–10301, 2021.

[41] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. Intra-processing methods for debiasing neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[42] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the Annual Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268, 2015.

[43] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2018.

[44] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8227–8236. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00842. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Quadrianto_Discovering_Fair_Representations_in_the_Data_Domain_CVPR_2019_paper.html.

[45] Mei Wang and Weihong 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.

[46] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 2017. URL http://proceedings.mlr.press/v54/zafar17a.html.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL http://jmlr.org/papers/v20/18-262.html.

[48] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[49] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11662.

[50] Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2277–2283. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/315. URL https://doi.org/10.24963/ijcai.2020/315.

[51] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6755–6764, 2020. URL http://proceedings.mlr.press/v119/martinez20a.html.

[52] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

[53] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 2020.

[54] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

[55] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[56] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation, 2020.

[57] Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification. *arXiv preprint arXiv:2207.10888*, 2022.

[58] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9981–9990, 2021.

[59] Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021.

[60] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

[61] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.

[62] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020.

[63] Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Bichen Wu, Zijian He, Zhen Wei, Kan Chen, Yuandong Tian, Matthew Yu, Peter Vajda, et al. Fbnetv3: Joint architecture-recipe search using predictor pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16276–16285, 2021.

[64] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.

[65] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[66] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[67] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016. URL http://arxiv.org/abs/1611.05431.

[68] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[69] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. *arXiv preprint arXiv:2105.12723*, 2021.

[70] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[71] Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 34:22614–22627, 2021.

[72] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rexnet: Diminishing representational bottleneck on convolutional neural network. *arXiv preprint arXiv:2007.00992*, 6, 2020.

[73] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837*, 2019.

[74] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[75] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.

[76] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021.

[77] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[78] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 589–598, 2021.

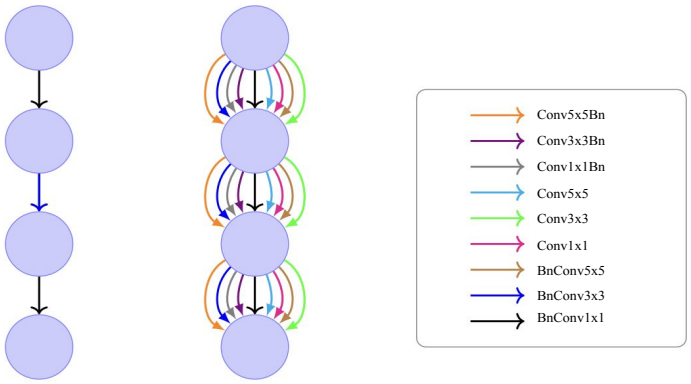

Figure 6: DPN block (left) vs. our searchable block (right).

A Additional Related Work

Face Recognition. Face recognition tasks fall into two categories: verification and identification. *Verification* asks whether the person in a source image is the same person as in the target image; this is a one-to-one comparison. *Identification* instead asks whether a given person in a source image appears within a gallery composed of many target identities and their associated images; this is a one-to-many comparison. Novel techniques in face recognition tasks, such as ArcFace [14], CosFace [15], and MagFace [16], use deep networks to extract feature representations of faces and then compare those to match individuals (with mechanisms called the *head*). We focus our analysis on identification and on examining how close images of similar identities are in the feature space of trained models.

Sociodemographic Disparities in Face Recognition. In this work, we focus on *measuring* the sociodemographic disparities across neural architectures and hyperparameter settings, and finding the Pareto frontier of face recognition performance and bias for current and novel architectures. Our work searches for architectures and hyperparameters which improve the undesired disparities. Previous work focuses on "fixing" unfair systems and can be split into three (or arguably four [41]) focus areas: preprocessing [e.g., 42, 43, 44, 45], inprocessing [e.g., 46, 47, 48, 49, 50, 45, 51, 52, 53, 54], and post-processing [e.g., 55, 56].

Neural Architecture Search (NAS) and Hyperparameter Optimization (HPO). A few works have applied hyperparameter optimization to mitigate bias in models for tabular datasets. Perrone et al. [17] recently introduced a Bayesian optimization framework to optimize accuracy of models while satisfying a bias constraint. The concurrent works of Schmucker et al. [18] and Cruz et al. [19] extend Hyperband [20] to the multi-objective setting and show its applications to fairness. The former was later extended to the asynchronous setting [29]. Lin et al. [57] proposes de-biasing face recognition models through model pruning. However, they consider just two architectures and just one set of hyperparameters. To the best of our knowledge, no prior work uses any AutoML technique (NAS, HPO, or joint NAS and HPO) to design fair face recognition models, and no prior work uses NAS to design fair models for any application.

A.1 Search Space Design

We design our search space based on our analysis in Section 2. In particular, our search space is inspired by Dual Path Networks [26] due to its strong trade-off between rank disparity and accuracy as seen in Figure 2.

Hyperparameter Search Space Design. We choose three categories of hyperparameters for NAS+HPO: the architecture head/loss, the optimizer, and the learning rate, depicted in Table 1.

Architecture Search Space Design. Dual Path Networks [26] for image classification share common features (ResNets [27]) while possessing the flexibility to explore new features [28] through a dual path architecture. We replace the repeating `1x1_conv-3x3_conv-1x1_conv` block with a simple recurring searchable block depicted in Figure 6. Furthermore, we stack multiple such

Table 1: Searchable hyperparameter choices.

Hyperparameter	Choices
Architecture Head/Loss	MagFace, ArcFace, CosFace
Optimizer Type	Adam, AdamW, SGD
Learning rate (conditional)	Adam/AdamW $\rightarrow [1e-4, 1e-2]$, SGD $\rightarrow [0.09, 0.8]$

Table 2: Operation choices and definitions.

Operation	Definition
BnConv1x1	Batch Normalization $\rightarrow$ Convolution with 1x1 kernel
Conv1x1Bn	Convolution with 1x1 kernel $\rightarrow$ Batch Normalization
Conv1x1	Convolution with 1x1 kernel
BnConv3x3	Batch Normalization $\rightarrow$ Convolution with 3x3 kernel
Conv3x3Bn	Convolution with 3x3 kernel $\rightarrow$ Batch Normalization
Conv3x3	Convolution with 3x3 kernel
BnConv5x5	Batch Normalization $\rightarrow$ Convolution with 5x5 kernel
Conv5x5Bn	Convolution with 5x5 kernel $\rightarrow$ Batch Normalization
Conv5x5	Convolution with 5x5 kernel

searched blocks to closely follow the architecture of Dual Path Networks. We have nine possible choices for each of the three operations in the DPN block as depicted in Table 2. The choices include a vanilla convolution, a convolution with pre-normalization and a convolution with post-normalization.

B Further Details on Experimental Design and Results

B.1 Experimental Setup

The list of the models we study from `timm` are: `coat_lite_small` [58], `convit_base` [59], `cspdarknet53` [60], `dla102x2` [61], `dpn107` [26], `ese_vovnet39b` [62], `fbnetv3_g` [63], `ghostnet_100` [64], `gluon_inception_v3` [10], `gluon_xception65` [65], `hrnet_w64` [66], `ig_resnext101_32x8d` [67], `inception_resnet_v2` [68], `inception_v4` [68], `jx_nest_base` [69], `legacy_senet154` [70], `mobilenetv3_large_100` [11], `resnetrs101` [71], `rexnet_200` [72], `selecsls60b` [73], `swin_base_patch4_window7_224` [9], `tf_efficientnet_b7_ns'` [74], `'tnt_s_patch16_224`[75], `twins_svt_large` [76] , `vgg19` [77], `vgg19_bn` [77], `visformer_small` [78], `xception` and `xception65` [65].

We study at most 13 configurations per model ie 1 default configuration corresponding to the original model hyperparameters with CosFace as head. Further, we have at most 12 configs consisting of the 3 heads (CosFace, ArcFace, MagFace) $\times$ 2 learning rates(0.1,0.001) $\times$ 2 optimizers (SGD, AdamW). All the other hyperparameters are held constant for training all the models. All model configurations are trained with a total batch size of 64 on 8 RTX2080 GPUS for 100 epochs each.

B.2 Obtained architectures and hyperparameter configurations from Black-Box-Optimization

In Figure 7 we present the architectures and hyperparameters discovered by SMAC. Particularly we observe that `conv 3x3` followed `batch norm` is a preferred operation and CosFace is the preferred head/loss choice.

B.3 Analysis of the Pareto-Front of different Fairness Metrics

In this section, we include additional plots that support and expand on the main paper. Primarily, we provide further context of the Figures in the main body in two ways. First, we provide replication

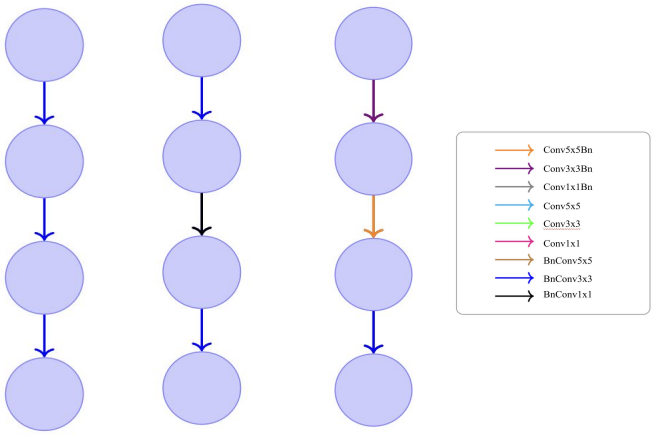

Figure 7: SMAC discovers the above building blocks with (a) corresponding to architecture with CosFace, with SGD optimizer and learning rate of 0.2813 as hyperparamters (b) corresponding to CosFace, with SGD as optimizer and learning rate of 0.32348 and (c) corresponding to CosFace, with AdamW as optimizer and learning rate of 0.0006

Table 3: Fairness Metrics Overview

Fairness Metric	**Equation**
Disparity	$\lvert Accuracy(male) - Accuracy(female)\rvert$
Rank Disparity	$\lvert Rank(male) - Rank(female)\rvert$
Ratio	$\lvert 1 - \frac{Accuracy(male)}{Accuracy(female)}\rvert$
Rank Ratio	$\lvert 1 - \frac{Rank(male)}{Rank(female)}\rvert$
Error Ratio	$\lvert 1 - \frac{Error(male)}{Error(female)}\rvert$

plots of the figures in the main body but for all models. Recall, the plots in the main body only show models with Error<0.3, since high performing models are the most of interest to the community. Second, we also show figures which depict other fairness metrics used in facial recognition. The formulas for these additional fairness metrics can be found in Table 3.

We replicate Figure 2 in Figure 8; Figure 3 in Figure 9; Figure 5 in Figure 10 and Figure 11. We add additional metrics with Disparity being plotted in Figure 12 and Ratio being plotted in Figure 13.

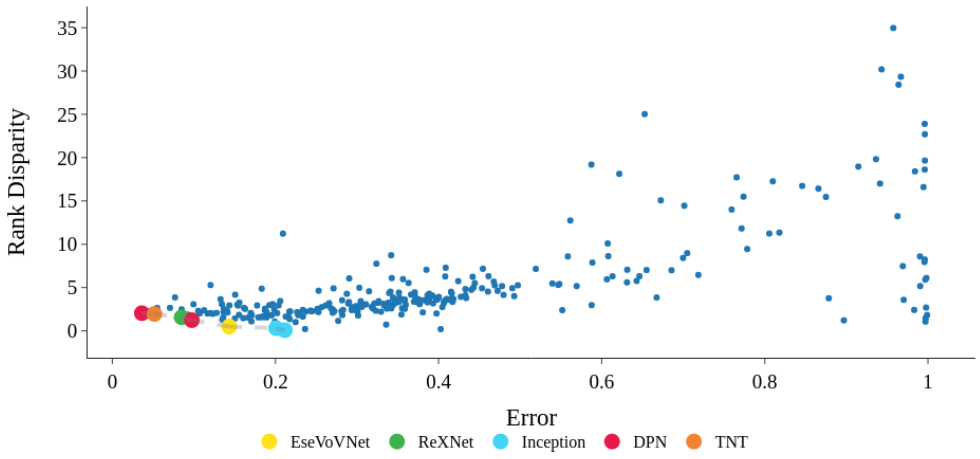

Figure 8: Replication of Figure 2 with all data points. Error-Rank Disparity Pareto front of the architectures with any non-trivial error. Models in the lower left corner are better. The Pareto front is notated with a dashed line. Other points are architecture and hyperparameter combinations which are not Pareto-dominant. DPN, ReXNet, EseVovNet, TNT, and Inception architectures are Pareto-dominant.

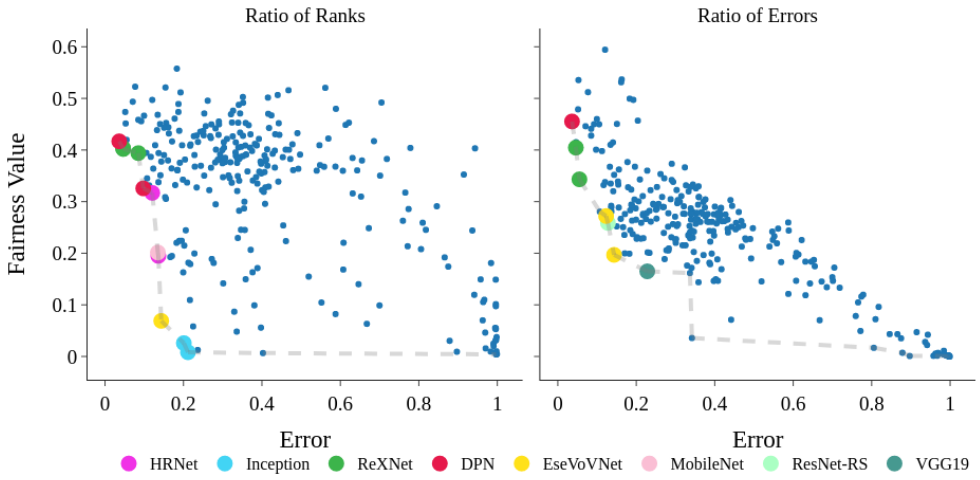

Figure 9: Replication of Figure 3 with all data points. Depending on the fairness metric, different architectures are Pareto-optimal. On the left, we plot the metric Ratio of Ranks which admit DPN, ReXNet, HRNet, MobileNet, EseVovNet, and Inceptions as Pareto-optimal. On the right, we plot the metric Ratio of Errors where DPN, ReXNet, EseVovNet, ResNet-RS, and VGG19 are architectures which are Parto-optimal

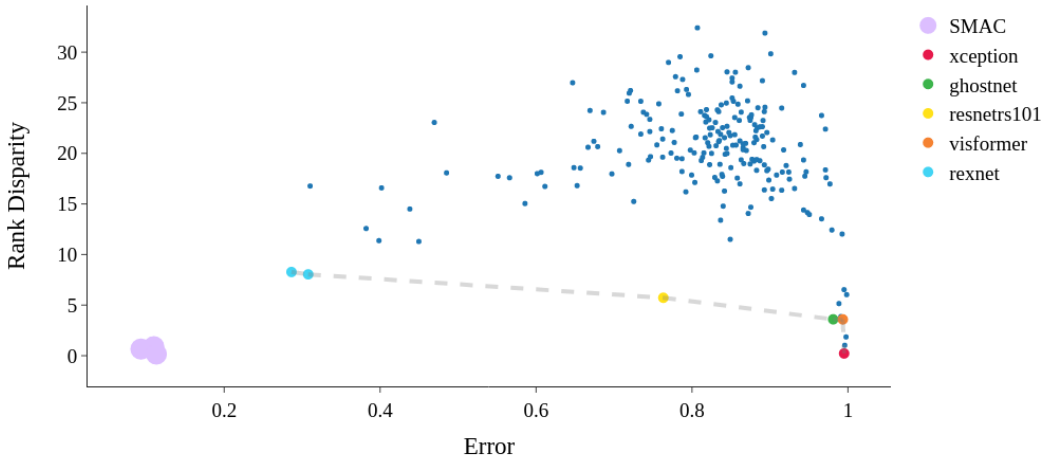

Figure 10: Replication of Figure 5 for VGGFace2 with all data points. Pareto front of the models on VGGFace2 test set with perceived gender as the protected attribute. The SMAC models discovered by joint NAS and HPO Pareto-dominate the `timm` models

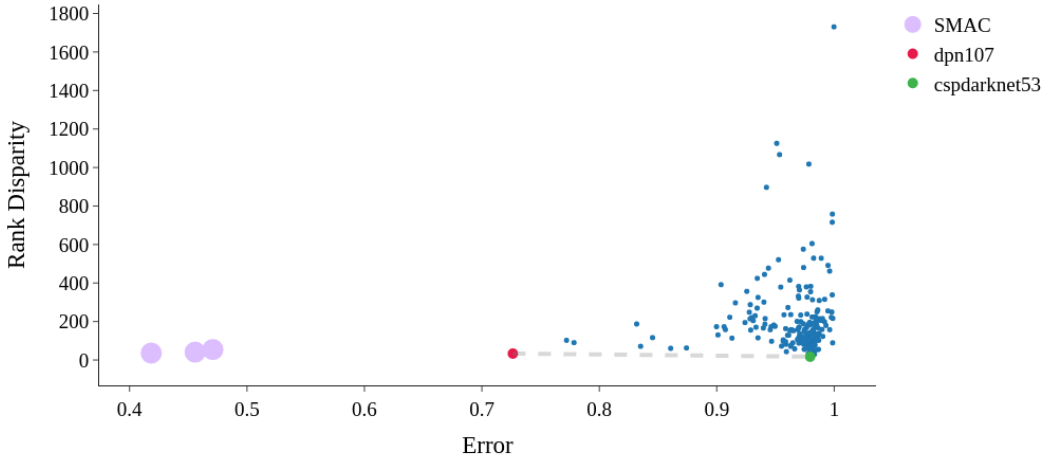

Figure 11: Replication of Figure 5 for RFW with all data points. Pareto front of the models on RFW test set with perceived ethnicity as the protected attribute. The SMAC models discovered by joint NAS and HPO Pareto-dominate the `timm` models

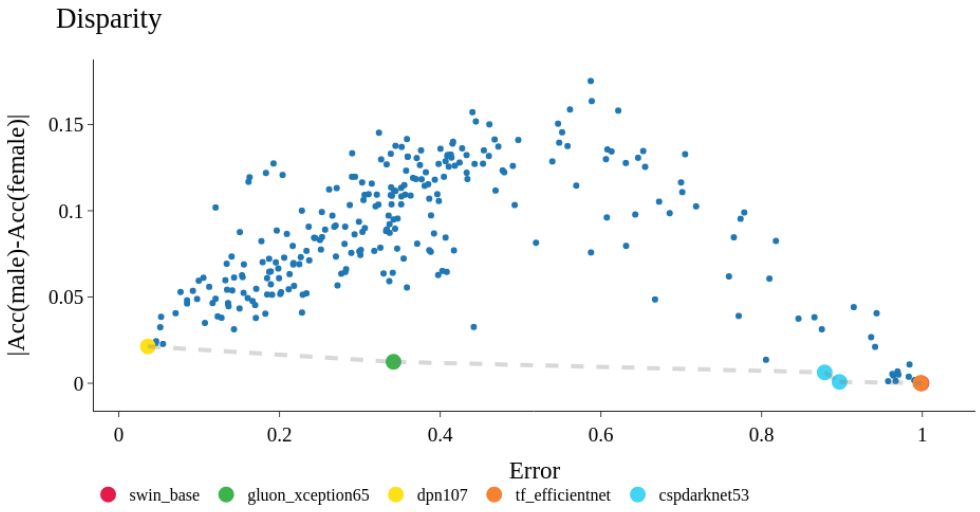

Figure 12: Extension of Figure 2 with all data points with the Disparity in accuracy metric. Error-Disparity Pareto front of the architectures with any non-trivial error. Models in the lower left corner are better. The Pareto front is notated with a dashed line. Other points are architecture and hyperparameter combinations which are not Pareto-dominant.

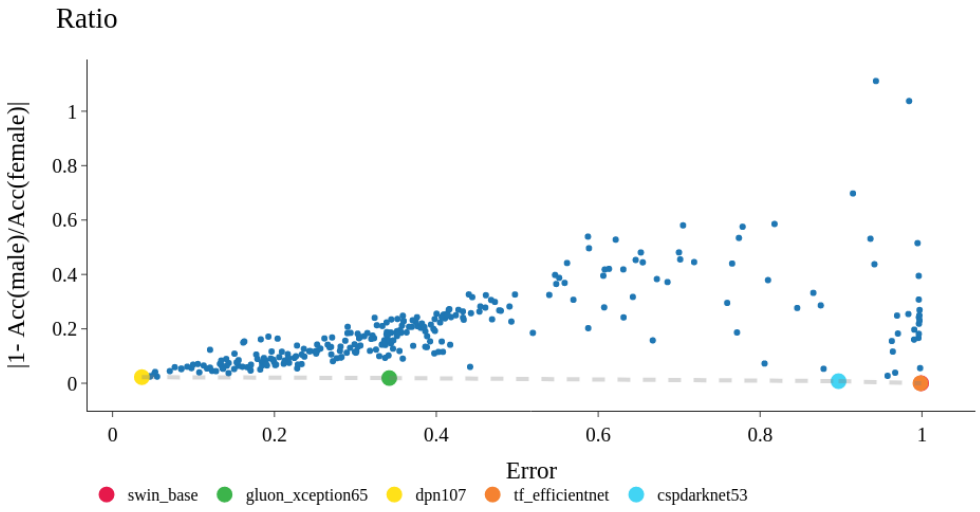

Figure 13: Extension of Figure 2 with all data points with the Ratio in accuracy metric. Error-Ratio Pareto front of the architectures with any non-trivial error. Models in the lower left corner are better. The Pareto front is notated with a dashed line. Other points are architecture and hyperparameter combinations which are not Pareto-dominant.