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

Rhea Sukthanker1; Samuel Dooley2; John P. Dickerson2,3, Colin White4, Frank Hutter1,5
Micah Goldblum6

1University of Freiburg, 2University of Maryland, 3ArthurAI, 4Abacus.AI,
5Bosch Center for AI, 6New 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...
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 Appendix A. 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 Section 2.
• 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 Section 3.
• We show that the architectures discovered transfer across different datasets with the same (perceived gender) and different (ethnicities) protected attributes Section 3.3.1.

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 B.

### 2 A Large-Scale Analysis of Architectures and Fairness

Experimental Setup. We use the following training pipeline – ultimately conducting 355 training runs with different combinations of 29 architectures from the Pytorch Image Model (timm) database [21] 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 and optimizer that was published with that model. We then conduct a training run with these hyperparameters and each of the three heads, ArcFace [14], CosFace [15], and MagFace [16]. Next, we use that default learning rate and optimizer that was published with that model. We then conduct a training run with these hyperparameters and each of the three heads, ArcFace [14], CosFace [15], and MagFace [16]. Next, we use that default learning rate and optimizer that was published with that model. We then conduct a training run with these hyperparameters and each of the three heads, ArcFace [14], CosFace [15], and MagFace [16].

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 [23]. 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 D:

Results and Discussion. By plotting the performance of each run with the error on the x-axis and rank disparity on the y-axis in Figure 1, we conclude that optimizing for error does not optimize for fairness. 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 for models with
lowest error there is low correlation between error and rank disparity ($\rho = -0.113$ for models with error $< 0.3$). In Figure 1, we see that Pareto optimal models include DPN, TNT, ReXNet, VovNet, and ResNets. 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.

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 architecture 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 [24] due to its strong trade-off between rank disparity and accuracy as seen in Figure 1. We choose three categories of hyperparameters for NAS+HPO: architecture head/loss, optimizer, and learning rate.

Architecture Search Space Design. Dual Path Networks [24] for image classification share common features (ResNets [25]) while possessing the flexibility to explore new features [26] through a dual path architecture. We replace the repeating $1x1_{\text{conv}}-3x3_{\text{conv}}-1x1_{\text{conv}}$ block with a 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 continuous 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 [27, 20, 28], 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 [29, 30, 31]. 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 [27] 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 [30] for multi-objective optimization with \( \rho \) 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 2, 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 2, 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 \( \text{Error} < 0.1 \)), 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 2 it is also apparent that some models such as VoVNet and DenseNet show 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.

![Figure 2: 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 (\( \text{Error} < 0.1 \)).](image)

#### 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 [32], 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) [33], 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 [21] 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 3, 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. In the case of RFW, we see that the transferred models generally exhibit low error, which is in contrast to the majority of the models on VGGFace2. We also call attention to the fact that, on both datasets, the models found by SMAC have considerably lower 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.

Acknowledgments

This research was partially supported by the following sources: NSF CAREER Award IIS-1846237, NSF D-ISON 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


A Schematic for Finding Fair Architectures

```
Architectures
Examples:
   Inception
   ViTransformer
   VovNet
   MobileNetV3
   Dual Path Net

Hyperparams
Examples:
   MagFace (head)
   ArcFace (head)
   CosFace (head)
   Optimizer
   Learning Rate

Design search space using best model.
Run NAS+HPO

Bias Error

355 models trained
Pareto frontier
Novel Search Space
Pareto dominant

Figure 4: Overview of our methodology.
```

B 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 [34]) focus areas: preprocessing [e.g., 35, 36, 37, 38], inprocessing [e.g., 39, 40, 41, 42, 43, 38, 44, 45, 46, 47], and post-processing [e.g., 48, 49].

**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 [27]. Lin et al. [50] 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.

B.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 [24] due to its strong trade-off between rank disparity and accuracy as seen in Figure 1.

**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 [24] for image classification share common features (ResNets [25]) while possessing the flexibility to explore new features [26] 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 5. 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 as depicted in Table 2. The choices include a vanilla convolution, a convolution with pre-normalization and a convolution with post-normalization.

### C Further Details on Experimental Design and Results

#### C.1 Experimental Setup

The list of the models we study from timm are: coat_lite_small [51], convit_base [52], cspdarknet53 [53], dial102x2 [54], dpn107 [24], ese_vovnet39b [55], fbnetv3_g [56], ghostnet_100 [57], gluon_inception_v3 [10], gluon_xception65 [58], hrnet_w64 [59], ig_resnext101_32x8d [60], inception_resnet_v2 [61], inception_v4 [61], jx_nest_base [62], legacy_senet154 [63], mobilenetv3_large_100 [11], resnetrs101 [64], resnet_200 [65], selecsls60b [66], swin_base_patch4_window7_224 [9], tf_efficientnet_b7_ns’ [67], ’tnt_s_patch16_224[68], twins_svt_large [69], vgg19 [70], vgg19_bn [70], visformer_small [71], xception and xception65 [58].
Table 3: Fairness Metrics Overview

<table>
<thead>
<tr>
<th>Fairness Metric</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disparity</td>
<td>\left</td>
</tr>
<tr>
<td>Rank Disparity</td>
<td>\left</td>
</tr>
<tr>
<td>Ratio</td>
<td>\left</td>
</tr>
<tr>
<td>Rank Ratio</td>
<td>\left</td>
</tr>
<tr>
<td>Error Ratio</td>
<td>\left</td>
</tr>
</tbody>
</table>

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.

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

In Figure 6 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.

![Figure 6: 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](image)

C.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 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 1 in Figure 8; Figure 7 in Figure 9; Figure 3 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 7: 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.

Figure 8: Replication of Figure 1 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 7 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 3 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 3 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 \texttt{timm} models.

Figure 12: Extension of Figure 1 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 1 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.