Toward Fairness in Face Matching Algorithms

Jamal Alasadi
Electrical and Computer Engineering
Rutgers University
New Brunswick, NJ, USA
jamal.alasadi@rutgers.edu

Ahmed Al Hilli
Electrical and Computer Engineering
Rutgers University
New Brunswick, NJ, USA
ahmed.alhilli@rutgers.edu

Vivek K. Singh
School of Communication & Information
Rutgers University
New Brunswick, NJ, USA
v.singh@rutgers.edu

ABSTRACT

Automated face matching algorithms are used in a wide variety of societal applications ranging from access authentication, to criminal identification, to application customization. Hence, it is important for such algorithms to be equitable in their performance for different demographic groups. If the algorithms work well only for certain racial or gender identities, they would adversely affect others. Recent efforts in algorithmic fairness literature (typically not focused on multimedia or computer vision tasks such as face matching) have argued for designing algorithms and architectures to tackle such bias via trade-offs between accuracy and fairness. Here, we show that adopting an adversarial deep learning-based approach allows for the model to maintain the accuracy at face matching while also reducing demographic disparities compared to a baseline (non-adversarial deep learning) approach at face matching. The results motivate and pave way for more accurate and fair face matching algorithms.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence → Philosophical/theoretical foundations of artificial intelligence • Computing methodologies → Artificial intelligence → Computer vision • Human-centered computing → Accessibility → Empirical studies in accessibility

KEYWORDS

Algorithmic Fairness, Face Matching, Deep Learning.

ACM Reference format:

Jamal Alasadi, Ahmed Al Hilli and Vivek K. Singh. 2019. Toward Fairness in Face Matching Algorithms. In *Proceedings of ACM Workshop on Fairness, Accountability, and Transparency in MultiMedia (FAT/MM '19). ACM, New York, NY, USA, 7 pages.* https://doi.org/*10.1145/3347447.3356751*

FAT/MM '19, October 25, 2019, Nice, France

ACM ISBN 978-1-4503-6915-2/19/10…$15.00
https://doi.org/10.1145/3347447.3356751

1 Introduction

Algorithms now affect multiple aspects of human lives including parole decisions, search results, and loan applications. Researchers have shown multiple such algorithms to be biased based on age, race, and gender. Multiple multimedia algorithms have also been reported to biased; for instance, [29] has reported that face detection and gender detection algorithms perform worse for women and people of color as compared to white men. Similar questions have been raised about object detection based on the different background settings across socio-economic status [31].

However, the issue of fairness in multimedia related tasks remains severely understudied. While the emerging efforts that *audit* multimedia algorithms (mentioned above) are a welcome change, the efforts to *counter* bias in multimedia algorithms are even rarer. This is in contrast to multiple efforts in other research communities attempting to create fair models across different domains such as credit scoring and recidivism prediction.

Here, we focus our attention on a specific multimedia research relevant task, namely face matching. Face matching is the problem of matching a low-resolution face image of a person with a high-resolution face image of the same person. While long-term face image datasets (e.g., an organization's database of employees) are captured at high resolution, they often need to be matched with low-resolution face images (e.g., those captured by a security camera) in different practical tasks. Such face matching is now increasingly being used to decide who gets access to buildings (and prosecute trespassers), to search for criminals, and for customizing digital applications. Hence, fairness in this process is very important to ensure that certain sections of the society are not unfairly denied access to resources or criminalized simply because of their demographic characteristics. There have been no attempts till date to quantify and ameliorate the issue of bias in *face matching* algorithms and this paper aims to fill this research gap.

Specifically, we present a new framework for fair matching of low resolution (LR) and high resolution (HR) facial images. The GAN-based framework consists of two parts - one which tries to maximize the face matching quality and the other which tries to minimize the ability of the network to infer the demographic properties of the individual whose facial image is under

consideration. We first compare the LR images with the HR images to decide if they are a same or different image; the classification is done without constructing a high-resolution image. We also include an adversary network that tries to infer the demographic properties of the individual being considered by simply looking at the output of the matching network. The adversary network punishes the face matching network if a demographic property (gender) is predictable from the matching prediction because that would indicate that the prediction is not independent of the demography of the person being considered. To give an analogy, if just knowing the result of a college admission algorithm (accept/reject) is enough to infer the gender of the applicant, then the algorithm is likely biased towards a gender.

The results obtained using the proposed approach on two well-known datasets indicate comparable (or better) performance at face matching and reduced discrepancy in the true positive rate (TPR) and false positive rate (FPR) results across different demographic (gender) descriptors.

To summarize, this paper makes the following contributions:

(1) It motivates and grounds the problem of *fairness* in face matching algorithms.

(2) It proposes a deep-learning adversarial approach for reducing bias in face-matching algorithms.

2 Related Work

For more than three decades, many advancements in face recognition and face matching have been introduced and the effectiveness of the developed systems has been widely demonstrated in practice. However, a theoretical assumption is generally made for the face region to be sufficient in its content quality to allow such a recognition [4]. With increasing application domains, most monitoring systems today depend on cameras with restricted definitions as opposed to high definition ones. In addition, the broad angle of the cameras is usually used in a way that would maximize the coverage area. This often results in very small face area with low quality content (e.g., in building security cameras). It becomes particularly challenging to recover intelligible high-resolution features from their corresponding low-resolution counterparts due to extreme information minimization that can be as small as 12x12 pixels. Hence, face matching i.e. matching the low resolution (LR) facial images with high resolution (HR) facial images is an area of active research interest [5]. Note that face matching problem is different from that face detection (finding a face within an image) and that of face recognition (matching a face image with a named identity).

Although there is no fixed boundary in the resolution ratio between LR and HR images, it is well documented that the recognition performance rapidly decreased when the image resolution is lower than 32 x 32 pixels. Moreover, Torralba et al., in [6] have demonstrated a certain degraded performance in recognition when face regions were less than 16 x16. Many studies have focused on the development of algorithms that recognize high-quality face images in normal conditions. However, according to the authors of [8], few have focused on the problem of low-resolution faces in low-quality image conditions such as in surveillance systems or resource constrained Internet-of-Things devices. A major challenge in such applications is the fact that high resolution images might not be accessible. Therefore, the performance of traditional face recognition procedures would not be satisfactory.

Multiple recent efforts have focused on low high-resolution face recognition. The learning method proposed in [3] maps the low and high resolution the high and low-resolution face images to a common area with nonlinear conversions using deep learning convolutional neural networks. Wang et al., in [7] have down sampled the original face images by a selected factor and then used nearest neighbor interpolation process to turn them back to the original resolution. Yang et al., [9] create the high-resolution image from the low-resolution image using a sparse representation for each patch of the low input then reconstruct by compressed sensing to find high resolution image. Wang et al., [10] use single neural network to learn more features that transfer from high resolution to low-resolution images.

Looking beyond the problem of face matching, multiple empirical studies have reported the issue of bias across machine learning applications in different domains ranging from school admission, to recruiting, getting loans, and insurances prices [24, 25, 26]. For example, the authors have reported that the algorithms used to decide on parole applications on inmates are systematically biased against non-White individuals [27], and the authors in [28, 29] have reported bias in automated face detection algorithms across genders and races.

On the positive side, there have also been multiple recent efforts aimed at increasing the fairness of machine learning algorithms. Zafar et al., [11] implemented a resilient tradeoff between fairness and accuracy. Also, they provide the relation to removing disparate misclassify data point solely on false positive and negative rates for sensitive attribute (i.e. the attribute which should not affect the classifier performance, such as gender) values. Bechavod et al., [12] showed fair classification using logistic regression by relying on the distance from the decision boundary rather than the real classification across values of the protected attributes. They evaluate this method to the ability to perform both fairness and accuracy using different dataset like criminal risk assessment, credit, and college admissions. Hardt et al., [13] provide a framework to remove discrimination against a particular sensitive attribute to predict the target based on available features. They used post-processing that selects different thresholds for different groups and adds randomization. Beutel et al., [14] used adversarial learning to input embedding on shared hidden layer in order to predict sensitive attributes and labels.

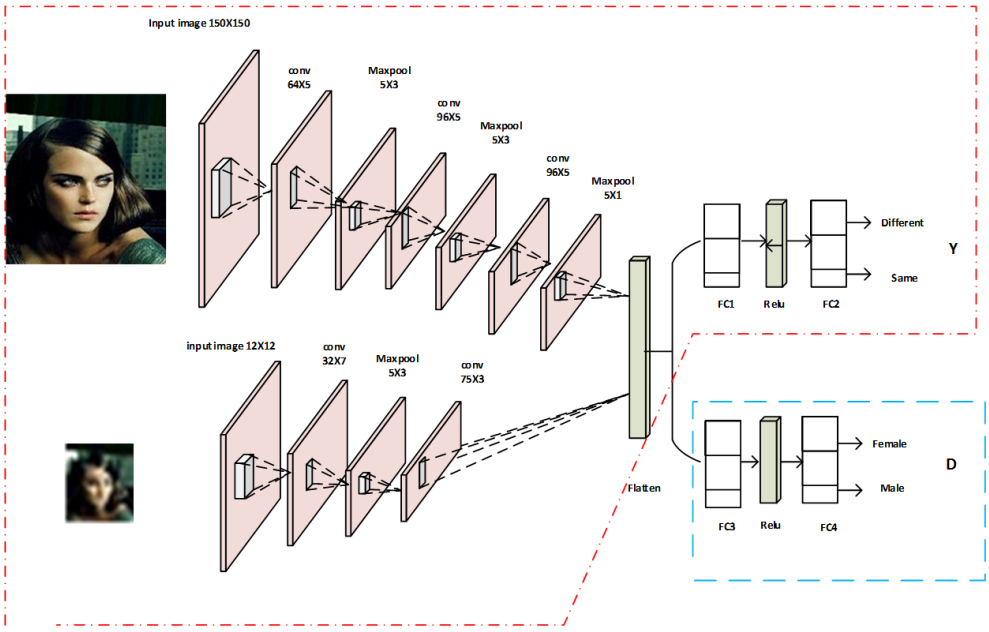

Figure 1: Overview of the proposed architecture for good face matching *but poor demographic identification.*

However, the literature on developing fair machine learning models in computer vision or multimedia tasks is severely lacking behind. One notable recent exception is [22]. However, while Amini et al. [22], focus on the problem of "face detection" i.e. the presence and absence of a face, we focus on the problem of face matching. Further, while Amini et al., focus on the "pre-processing" step via the inclusion of samples of different kinds in the training process, our key idea is based on the use of adversarial learning to ensure that the output vales do not inadvertently encode demographic values.

To quantify the "fairness" of algorithms, we will focus on the comparisons based on three different metrics: equality of accuracy, equality of opportunity, and equal odds based on the emerging literature on Fairness in Machine Learning [12], [15], [16], [17].

Equality of accuracy metric is based on the idea that the accuracy of the results needs to be as equal as possible for the different demographic groups considered. Equality of opportunity (EoO) metric mandates an equal true positive rate for the two classes [17]. Hence, a ground truth "true" match should have equal odds of being labeled as "true" by the developed algorithms irrespective of the demographic properties (here, gender) of the person in the low-resolution image. Considering a toy example of (output = college admissions) where deservingness is defined on SAT score and sensitive class is defined on binary gender, EoO mandates that if 90% of the ground truth-based *deserving* candidates are given admission amongst males, then 90% of the *deserving* female candidates should also get admission. Conversely, in the considered scenario of face matching based building admission, if 90% of the deserving men are allowed entry, then 90% of the deserving women should also be allowed access to the building.

Equal odds metric is an extension of the above idea to include both the true positive rate and the false positive rate [17]. Hence, the difference in the false positive rates for different demographic groups (e.g. male/female) is also considered in this work. The goal is to minimize the discrepancy of these metrics across different demographic groups.

3 The Proposed Approach

Following Zhang et al., [18] we create a network that undertakes two competing tasks- one focused on accuracy and one on fairness and adapt it to the task of face matching as shown in Figure 1. The baseline approach is dedicated to the matching between high resolution image and low-resolution images. The network takes two face images, HR image of size 150X150, and LR image of size 12X12. The output of the network is a decision indicating whether the input images belong to the same person or not.

The path corresponding to the high-resolution image consists of a set of convolutional layers and MaxPooling to map face features followed by a flattening concatenation and a set of fully connected nodes. The path corresponding to the low-resolution images consists of two convolutional layers followed by flattening concatenation layers. The output of both paths is then concatenated by a flattening layer, which is followed by two branches of a fully connected layer.

As shown in Figure 1, the first branch is used to predict Y [same/different face] accurately and the second branch is used to predict D (e.g. male/female) poorly. We consider the binary cross-

entropy loss for [same/different face], which we refer to as L_y and the binary cross entropy loss for demographic (male/female), which we refer to as L_d. In the baseline architecture, to train the network, we back-propagate L_y through the network marked in red color in Figure 1 and L_d through the network marked in blue color. But in the proposed architecture, we need to train the network [same/different] to be good at predicting Y and bad at predicting D. If we subtract L_d from L_y, the network will be encouraged to produce a logit that cannot be used to predict demographic descriptor and yield Y values that are better at achieving parity.

While the training for the blue component remains same in the baseline and proposed architectures, the key difference is in the red component. While only L_y is back-propagated in the baseline approach, a combination of L_y and L_d is back-propagated in the proposed approach. Specifically, the loss is minimized through backpropagation algorithm by gradient computation.

$$L = L_y - \gamma L_d \qquad (1)$$

Where γ is a weighting scalar, L_y is same/ different face cross entropy loss function and L_d represents male/female cross entropy loss. Note that while Zhang et al., [18] also propose the inclusion of an additional term that prevents the predictor from moving in a direction that helps the adversary decrease its loss, we do not include it based on the findings of Wadsworth et al., [23] who report the results to be more stable without such an additional term.

3.1 Implementation Details

The feature maps obtained from the layers are processed by the CNN network that consists of five convolutional layers with ReLU activation function after each layer except the last convolutional layer for the second path. The first path consists of three convolutional layers with ReLU activation function after every layer. The first convolutional layer has 64 feature maps with a filter size of 5X5, the second convolutional layers have 96 feature maps with a filter size of 5X5 and the third layer has 128 feature maps with a filter size of 3X3 the output of this path is combined with the last convolutional layer of low-resolution images. So, the second path for the low-resolution image is consists of two convolutional layers the first convolutional layers have 32 feature maps with a filter size of 7X7 and the second convolutional layer has 75 feature maps with a filter size of 3X3. These layers are followed by flattening concatenation, which represented a features layer, and two branches of fully connected layers. The feature layer is used as input to two fully connected classification networks, both with two fully connected layers of 128 and 2 neurons with ReLU layer between them to reduce the error and the range of the losing thereby increase the accuracy. The first is used for same/different classification while the second is used for demographic (e.g. male/female) classification.

Also, we used a max pooling layer that basically down sampled the size of our activation maps. So, it is going to be the region at which we pool over. Then, finally we take our last convolutional layer output connected with fully connected layers and use that to get a final score function. We implement our models using the CUDA package. The batch size of the training is 50. Also, the batch size of validation is 30 and it is done every 50 training. the learning rate is 0.001 with decay of $5*10^{-4}$. In each training trial, an image is taken from the training set, and with 50% is a random uniform scaler (α) is generated if $\alpha >= 0.5$ then low-resolution version of the same person face image is constructed else, different face images is uses to construct the low-resolution version. Both high and low-resolution images are fed to the network, and the output label is set to 0 or 1 if the low-resolution image is same or else then the high-resolution image respectably. The next training trial, an image is taken from the training set, and low-resolution image is constructed from different image. The label is set to 1. We train the network for 5 epochs.

4 Datasets Considered

Our model is implemented using the PyTorch open source framework [19]. We validate our method using the Celeb A [20] and UMD faces [21] datasets, which both contain facial images. Each of the two datasets contain more than 100,000 images to allow for testing and training. They also have different strengths and challenges with regards to this work's goal of validating fairness in face matching across different demographic groups. Celeb A dataset includes demographic information in terms of gender (male/female), age (young, not young), and attractiveness (attractive, not attractive) but does not includes identity information to allow for comparison of multiple poses of the same person. Hence, we use a low-resolution of the same pose image for undertaking the comparisons. UMD faces dataset contains demographic labels in terms of only gender (male, female) but has multiple poses for the same person for comparison. This allows for the low-resolution image to come from a different pose than the high-resolution image. Hence, we have decided to use both the datasets to consider the same pose scenario (Celeb A) as well as the multiple poses scenario (UMD Faces) to analyze the trends. Also, while the proposed approach is generic, in this paper we focus on gender as the sensitive attribute and consider bias and the debiasing results over binary genders (male, female) as marked in the dataset. Please note that we recognize gender to include multiple non-binary variants but focus on the two genders as a proof of concept based on the currently available annotations in the datasets.

Table 1. Baseline Approach Results: Celeb-A dataset

Baseline (same/different) for face images				
Attributes	No. of Images	Accuracy	TPR	FPR
Male	33,282	95%	95.22%	04.53%
Female	47,317	94%	94.22%	04.73%
Baseline architecture demography prediction				
Male &Female		97%	96.02%	01.89%

Table 2. Proposed Approach Results: Celeb-A dataset

Proposed (same/different) for face images				
Attributes	No. of Images	Accuracy	TPR	FPR
Male	33,282	96%	98.51%	05.64%
Female	47,317	96%	98.34%	05.71%
Adversarial architecture demography prediction				
Male &Female		77%	75.2%	20.82%

The Celeb A dataset consists of 202,599 images. The number of training images is 60% i.e. approximately 121,000 images, which includes 20% of the images which are used for validation. The test set includes roughly 81,000 images. UMD faces dataset consists of multiple "batches" as defined by its creators. We used "batch 1", which has 175,534 images for training phase and use "batch 2" that has 115,100 images for the testing phase. During the test phase we split the sample images based on the gender of the person (male, female) and select an equal number (10,000 male and 10,000 females) for the test set.

5 Results

For Celeb A, our training set is around 121,000 images and our test set is around 80,000 images as shown in Table 1. Table 1 shows the results for accuracy as well as TPR (True Positive Rate) and FPR (False Positive Rate) for the task of matching LR and HR face images. The results have been split based on the test images selected as those belonging to specific demographic groups (male, female). Table 2 shows the same results for the proposed architecture that penalizes the algorithm if it can identify the demographic properties of the individuals. The comparison between the two approaches are summarized in Table 3. The results indicate that the proposed approach resulted in lower discrepancy between demographic groups in terms of accuracy, TPR, and FPR. Further, comparing Tables 1 and 2, we note that the average accuracy for the primary face matching task has improved. As further supporting evidence, we notice that the demography prediction rate goes down between the baseline approach and the proposed approach, suggesting that the logit parameter obtained from the first network encode lesser gender information in the proposed approach.

In the second set of experiments we focus on the UMD faces dataset. In UMD faces dataset, there are around 20 images for each person in different poses. We pick one HR image for the person and create a LR version of another pose by the same person. These LR images are matched with LR versions of the images of other people (of same gender) and there is a 50:50 chance of the LR image being the same person as the HR image. The two images (LR and HR) are passed on to the network to identify if they belong to the same person.

Table 3: Comparison between Baseline and Proposed Approaches for Celeb-A dataset

	BASELINE METHOD			PROPOSED METHOD		
Attributes	Delta Acc	Delta TPR	Delta FPR	Delta Acc	Delta TPR	Delta FPR
Gender	1.00	1.00	0.20	0.00	0.17	0.07

We take a random but equal sized sample of 10,000 males and females to analyze the results. The results presented in Tables 4 - 6 are based on an average across 10 runs to obtain confidence in the findings and understand the distribution to allow for statistical comparisons across them.

Table 4. Baseline Approach Results: UMD faces dataset

Baseline (same/different) for face images				
Attributes	No. of Images	Accuracy	TPR	FPR
Male	10,000	97.78%	98.02%	2.46%
Female	10,000	98.34%	98.13 %	1.44 %

As shown in Table 4, the accuracy, TPR, and FPR were different for the two considered genders. A statistical t-test undertaken over the distribution of these scores for accuracy and FPR was found to be significant at $\alpha = 0.05$ threshold. The difference in the scores for the two genders in terms of TPR gave α value of 0.094, indicating that it was not significant at $\alpha=0.05$ threshold. Hence, there exists a significant disparity across gender in the performance of the baseline face matching algorithm in terms of accuracy and FPR.

The results based on the proposed architecture are shown in Table 5 and the comparison between baseline and proposed architectures is shown in Table 6.

Table 5. Proposed Approach Results: UMD faces dataset

Proposed (same/different) for face images				
Attributes	No. of Images	Accuracy	TPR	FPR
Male	10,000	98.23%	98.68 %	2.21%
Female	10,000	98.46%	98.72 %	1.80%

Table 6. Comparison between Baseline and Proposed Approaches for UMD faces dataset

	BASELINE METHOD			PROPOSED METHOD		
Attributes	Delta Acc	Delta TPR	Delta FPR	Delta Acc	Delta TPR	Delta FPR
Gender	0.56%	0.11%	1.02%	0.23%	0.04%	0.41%

As can be seen in Tables 4, 5, and 6, the proposed architecture reduced the difference in the performance of the algorithm in terms of accuracy as well as TPR and FPR. We evaluated the changes in the discrepancy scores between the proposed and baseline architectures and found that the reductions in the disparity were significant for all three considered features i.e. accuracy, TPR, and FPR at α=0.05 level.

Based on the trends in the two datasets, we report that the proposed GAN based approach is useful at reducing the disparity in the performance of a deep-learning based face matching algorithms across gender as measured through the metrics of deltas in accuracy, TPR, and FPR. Also, notably, while most proposed approaches for fairness result in loss in accuracy, here we see that the average accuracy levels remained same or even increased slightly with the introduction of the adversarial network. One optimistic interpretation is that the architecture forced the network to focus on features that were really useful for face matching without taking any gender-based shortcuts. However, the real causal analysis is left part of the future work.

This work motivates the problem of fairness in face matching algorithms and presents one method to tackle the issue. It also has a number of limitations. First it uses a fairly standard machine deep learning architecture (convolutional neural networks) for the task at hand. The obtained results however are promising, and seem to be adequate for the task. More importantly, they demonstrate the variance in performance across genders and hence motivate debiasing architectures. Future efforts in this direction should consider more advanced architectures for the baseline as well as the adversarial approaches. The notion of sensitive attributes in this work is limited to binary genders (as available in the dataset). This should be expanded in future work to include non-binary identities as well as other notions of demography such as race, age, nationality etc.

6 Conclusion

This paper marks the first effort at designing face matching algorithms that are accurate yet fair. The proposed GAN-based architecture yielded high accuracy but also reduced the disparities in the performance of the algorithm for different genders across accuracy, true positive rate and false positive rate. The results pave way for more accurate and fair face matching algorithms, which would provide equitable opportunities to different groups in security as well as personalization applications. While the presented work also has some limitations, we hope that it will help make the research agenda around fairness in multimedia algorithms more mainstream in the ACM Multimedia community.

ACKNOWLEDGMENTS

The authors would like to thank Hamed Pirsiavash (UMBC) for useful feedback on a previous version of this paper.

REFERENCES

[1] Wang, Z., Chang, S., Yang, Y., Liu, D., & Huang, T. S. (2016). Studying very low resolution recognition using deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4792-4800).

[2] Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 34-42).

[3] Zangeneh, E., Rahmati, M., & Mohsenzadeh, Y. (2017). Low Resolution Face Recognition Using a Two-Branch Deep Convolutional Neural Network Architecture. arXiv preprint arXiv:1706.06247.

[4] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Convex optimization with sparsity-inducing norms. Optimization for Machine Learning, 2011. 6

[5] Zou, W. W., & Yuen, P. C. (2012). Very low resolution face recognition problem. IEEE Transactions on Image Processing, 21(1), 327-340

[6] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. TPAMI, 2008. 2.

[7] Wang, Z., Chang, S., Yang, Y., Liu, D., & Huang, T. S. (2016). Studying very low resolution recognition using deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4792-4800).

[8] Zou, W. W., & Yuen, P. C. (2012). Very low resolution face recognition problem. IEEE Transactions on Image Processing, 21(1), 327-340

[9] Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. IEEE transactions on image processing, 19(11), 2861-2873.

[10] Wang, Z., Chang, S., Yang, Y., Liu, D., & Huang, T. S. (2016). Studying very low resolution recognition using deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4792-4800).

[11] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017, April). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web (pp. 1171-1180). International World Wide Web Conferences Steering Committee.

[12] Bechavod, Y., & Ligett, K. (2017). Penalizing unfairness in binary classification. arXiv preprint arXiv:1707.00044.

[13] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

[14] Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075.

[15] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-268). ACM.

[16] Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and information systems, 35(3), 613-644.

[17] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

[18] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. arXiv preprint arXiv:1801.07593.

[19] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... & Lerer, A. (2017). Automatic differentiation in pytorch.

[20] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[21] Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R., & Chellappa, R. (2017, October). Umdfaces: An annotated face dataset for training deep networks. In 2017 IEEE International Joint Conference on Biometrics (IJCB) (pp. 464-473). IEEE.

[22] Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., & Rus, D. (2019). Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. AIES conference, 2019.

[23] Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*.

[24] Hajian, S., Bonchi, F., & Castillo, C. (2016, August). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126). ACM.

[25] Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and information technology, 15(3), 209-227.

[26] O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

[27] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica, May, 23.

[28] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77-91).

[29] 'A white mask worked better': why algorithms are not colour blind, https://www.theguardian.com/technology/2017/may/28/joy-buolamwini-when-algorithms-are-racist-facial-recognition-bias

[30] Editorial Board, Washington Post (2019). Facial recognition poses serious risks. Congress should do something about it. https://www.washingtonpost.com/opinions/facial-recognition-poses-serious-risks-congress-should-do-something-about-it/2018/07/18/c4c8973c-89c1-11e8-a345-a1bf7847b375_story.html?utm_term=.8a991491dd14

[31] de Vries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does Object Recognition Work for Everyone?. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 52-59).