Does lossy image compression affect racial bias within face recognition?

Seyma Yucer, Matt Poyser, Noura Al Moubayed, Toby P. Breckon
Department of Computer Science, Durham University
Durham, UK

Abstract

Yes *- This study investigates the impact of commonplace lossy image compression on face recognition algorithms with regard to the racial characteristics of the subject. We adopt a recently proposed racial phenotype-based bias analysis methodology to measure the effect of varying levels of lossy compression across racial phenotype categories. Additionally, we determine the relationship between chroma-subsampling and race-related phenotypes for recognition performance. Prior work investigates the impact of lossy JPEG compression algorithm on contemporary face recognition performance. However, there is a gap in how this impact varies with different race-related inter-sectional groups and the cause of this impact. Via an extensive experimental setup, we demonstrate that common lossy image compression approaches have a more pronounced negative impact on facial recognition performance for specific racial phenotype categories such as darker skin tones (by up to 34.55%). Furthermore, removing chroma-subsampling during compression improves the false matching rate (up to 15.95%) across all phenotype categories affected by the compression, including darker skin tones, wide noses, big lips, and monolid eye categories. In addition, we outline the characteristics that may be attributable as the underlying cause of such phenomenon for lossy compression algorithms such as JPEG.*

1. Introduction

A growing number of studies focus on racial bias within face recognition due to the prevalence of disparate real-world performance on inter-sectional racial groups [1]. Such attention has forced several organisations to withdraw their algorithms or datasets due to racial biases, and disparities [2, 3, 4]. Nevertheless, there are still many areas, such as employment, public security, criminal justice, and credit reporting, where face recognition applications are in use, meaning that we need fair, trustworthy, and bias-free face recognition [5, 6].

From image acquisition to evaluation, all phases of face recognition are prone to bias. However, most research focuses on the latter aspects of dataset collection and model evaluation to explore and mitigate such bias [7, 8, 9]. As such, many datasets and annotations have been released [10, 11], generative adversarial networks have been explored to enrich under-represented groups during training [12, 13] and regularisation methods have been proposed to minimise performance differences between subgroups [14]. Furthermore specific evaluation methodologies have been devised to tackle bias collaboratively [15, 16, 17]. Despite this plethora of research, no studies examine the potential impact of image acquisition decisions when addressing racial bias within face recognition. Any source of bias at this early stage is just propagated and exacerbated within contemporary face recognition approaches [18].

On the other hand, existing image acquisition standards for face recognition systems such as ISO/IEC 19794-5 [19] and ICAO 9303 [20] propose both image-based (i.e. illumination, occlusion) and subject-based (i.e. pose, expression, accessories) quality standards to ensure facial image quality. Accordingly, facial images should also be stored using lossy image compression standards such as JPEG [21] or JPEG2000 [22]; and identifiable for gender, eye colour, hair colour, expression, properties (i.e. glasses), pose angles (yaw, pitch, and roll), and landmark positions. However, common face recognition benchmarks do not conform to the ISO/IEC 19794-5 and ICAO 9303 standards. Moreover, in-the-wild samples are often obtained under the varying camera and environmental conditions to challenge the proposed solutions. Nevertheless, most facial image samples within such datasets are compressed via lossy JPEG compression [23].

Accordingly, some limited previous work [24, 25, 26] focuses on the impact of low-quality, blurred, noisy or distorted imagery on Convolutional Neural Network (CNN) based image recognition or classification. Dodge and Karam [27] highlight a significant decrease in contemporary neural network performance, whilst human examiners remain resilient to such factors. Particularly, Torfason [28] focuses on compression methods and bypasses the decoding phase of image compression. They point out that encoded representations are more advantageous than com-

pressed/decoded images for classification and semantic segmentation. Poyser [29] evaluates the impact of lossy compression algorithms on various CNN architectures, in which they measure the robustness and performance impact of compression for various computer vision tasks. They determine that, in general, CNN architectures can be resilient to the introduction of lossy JPEG compression artefacts if the initial training regime includes the use of compressed images [29]. These results align with the findings of Zanjani [30], who considers the impact of JPEG 2000 compression [22] on CNN for cancer diagnosis systems. Indeed, retraining the CNN architecture on lossily compressed images affords a 59% performance increase for tumour detection within compressed test imagery [30].

For face recognition approaches, the National Institute of Standards and Technology (NIST) provides a comprehensive assessment of compressed image influence on facial recognition algorithms [31]. It investigates the speed versus accuracy trade-off for early machine learning-based face recognition algorithms. Karahan [32] indicates that image blur, noise, and occlusion can cause significant degradation in face recognition accuracy. Another study [33] improves face detection and recognition performance on low-quality images by introducing a fusion quality prediction network. Moreover, Terhörst [34] shows quality assessment algorithms are skewed towards the subgroups which are also affected by face recognition bias.

Prior literature on image acquisition operations (compression, quality assessment) for face recognition [35] are limited with regard to racial bias and its race-based phenotypic influence, which is where this study is focused. The most related work to ours, [36] explores the test image distortion impact on pre-trained face recognition models using binary gender G1 (Male) and G2 (Female), and race R1 (light skin colour) and R2 (dark skin colour) subgroups. As a result, they find that the regions of interest used in the models shift towards less discriminatory regions in the presence of distortions, resulting in unequal performance degradation among subgroups.

In this study, we examine whether lossy image compression adversely impacts phenotype-based racial performance bias within face recognition during training and testing. We estimate such impact on phenotype attribute categories individually. Furthermore, we also investigate differing chroma-subsampling rates to assess how this common lossy compression colour-related trait directly impacts recognition performance across varying phenotype-based categories. More precisely, however, we determine the relationship between the level of compression and chroma-subsampling applied and recognition performance in order to allow us to build a better understanding.

To these ends, we adapt the recently established evaluation methodology [17] that introduces phenotype-based racial bias measurement for face recognition. Furthermore, we determine the effect of varying factors, including the compression levels of lossy JPEG [21] image encoding, chroma-subsampling, and compressed versus non-compressed training on different race-based phenotype categories in order to evaluate the racial bias across multiple face recognition datasets. In this paper, our key contributions are as follows:

Attribute	Categories
Skin Type	Type 1 / 2 / 3 / 4 / 5 / 6
Eyelid Type	Monolid / Other
Nose Shape	Wide / Narrow
Lip Shape	Full / Small
Hair Type	Straight / Wavy / Curly / Bald
Hair Colour	Red / Blonde / Brown / Black / Grey

Table 1. Adapted Facial phenotype attributes and their categorisation from [17].

- we evaluate the impact of lossy image compression on CNN-based facial recognition approaches across different racial characteristics using the phenotype-based methodology [17], extending the earlier studies of [17, 35, 29].
- we compare several variants of training strategies, including lossy compression, within the balanced/imbalanced training datasets and race-related facial phenotypes.
- we experimentally demonstrate that the use of lossy image compression during inference adversely affects the performance of contemporary face recognition approaches [37] on a subset of race-related facial phenotype grouping (i.e. darker skin tones, monolid eye shape) and that its effect is present regardless of whether compressed imagery is used for model training.
- we investigate the specific impact of chroma-subsampling on bias performance by comparing recognition performance with and without chroma-subsampling within lossy compressed facial imagery.

2. Experimental Methodology

In this section, we explain the phenotype-based racial bias evaluation methodology used (Section 2.1), the most widespread lossy image compression process (JPEG, Section 2.2), how we evaluate the influence of chroma subsampling (Section 2.3), our compression level selection (Section 2.4), and the training strategies used (Section 2.5) for the generation of our results (Section 3).

2.1. Phenotype-based Bias Analysis Methodology

Previous work highlights the negative impacts of using standard geographically based racial grouping labels to evaluate cross-race face recognition performance [38, 39]. Accordingly, many studies [17, 39, 40] suggest avoiding using erroneous racial or binary skin tones grouping strategies or

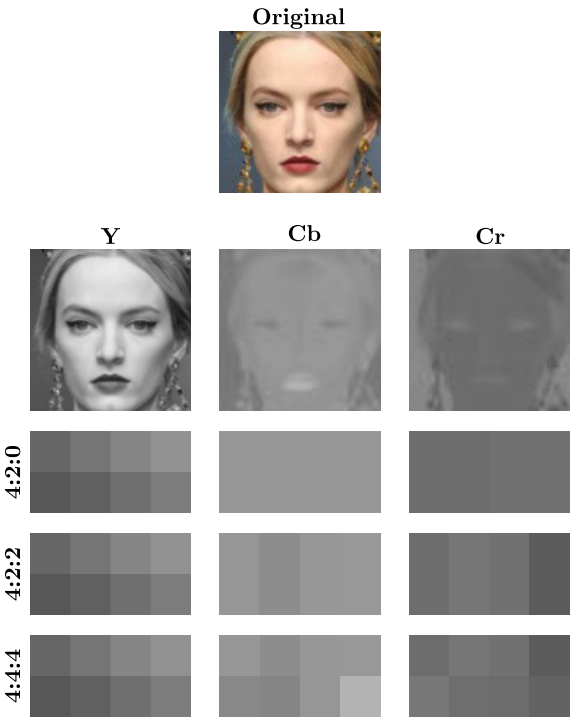

Figure 1. Chroma subsampling operation on different rates (4:2:0, 4:2:2, 4:4:4). Each rate differs according to how many pixels will be the same in the block.

exposing protected demographic attributes that can cause privacy and consent violations for individuals.

Alternatively, we adopt a racial bias analysis methodology that uses facial phenotype attributes for face verification (one-to-one facial matching) task [17]. The study categorises representative racial characteristics on the face and audits these attributes: skin types, eyelid type, nose shape, lips shape, hair colour and hair type for two different publicly available face datasets: VGGFace2 (test set) [41], and RFW [42]. We show each of the predefined phenotypes and their categories in Table 1. Moreover, this methodology provides different pairing strategies for face verification to draw attention to the importance of pairing for comprehensive evaluation. It introduces attribute-based pairings, which contain same-attribute grouping pair combinations to compare individual attribute performance for face verification. Additionally, the study shares cross-attribute pairing combinations between each grouping to measure false matching rates between all possible attribute category pair combinations.

On this basis, we use the set of observable characteristics of an individual face where race-related facial phenotype labels provide a relation between the task performance for a given face image sample under varying levels of lossy image compression and its underlying racial characteristics.

2.2. Lossy Image Compression

The Joint Photographic Experts Group (JPEG), an international image compression standard [21] for still images, operates within manageable algorithmic space and time complexity whilst offering good reconstruction image quality. The JPEG standard defines four operating modes *(1: Sequential Lossless Mode, 2: Sequential DCT-based Mode, 3: Progressive DCT-based Mode, 4: Hierarchical Mode)*, formed by an encoder and decoder which follow block-based transform coding. The image encoding strategy includes colour space transformation (from RGB to YCrCb), chroma channel subsampling, Discrete Cosine Transform (DCT), quantisation and entropy coding to compress the image [21].

In this study, we use ImageMagick Library (version 7.0.11.13) to perform JPEG compression (via libjpeg 8). The implementation switches the JPEG operational modes according to the compression level specified (i.e. quality level q, range: 0 - 100 for JPEG, higher = better image quality, less information loss + larger file sizes). Similar to the mode one operation, it does not down-sample the chroma channels if the compression level is higher than 90 (i.e. there is no colour-based information loss for compression, $q = 90$). It applies the baseline JPEG algorithm between compression levels 90 and 10, which is sequential DCT-based Mode (2). For compression levels, $(q = 90)$, lossy compression is applied to both the luminance channel, Y, and the colour containing chroma channels, Cr, Cb.

2.3. Chroma Subsampling

Standard lossy compression algorithms such as JPEG contain a colour space reduction step, as the human eye is less sensitive to chromatic (i.e. colour) changes than changes in illumination (i.e. brightness). In this step, the luminance channel (Y) remains unchanged, but the image colour space (Cr and Cb) is reduced. Subsequently, by default JPEG algorithm employs 4:2:0 chroma subsampling to reduce the colour information of the original image. It takes a 2-by-2-pixel block within each block and assigns the same colour (the colour of the top-left pixel) while the luminance component varies. Alternatively, for less colour information reduction, 4:2:2 with half sampling rate horizontally takes 2 pixels in each row and assigns the same colour. In Figure 1, we illustrate the three different sampling ratios (4:2:0, 4:2:2 and 4:4:4 no subsampling) on image pixels. In this first step of compression, chroma subsampling converts the image to YCbCr colour space and then reduces the chroma channels Cb, Cr information by assigning the top-left block pixel value to other pixels in the block. Block size and how many pixel values remain vary according to the sampling ratio.

This evaluation investigates the effect of sampling ratio on phenotype-based face recognition performance. We

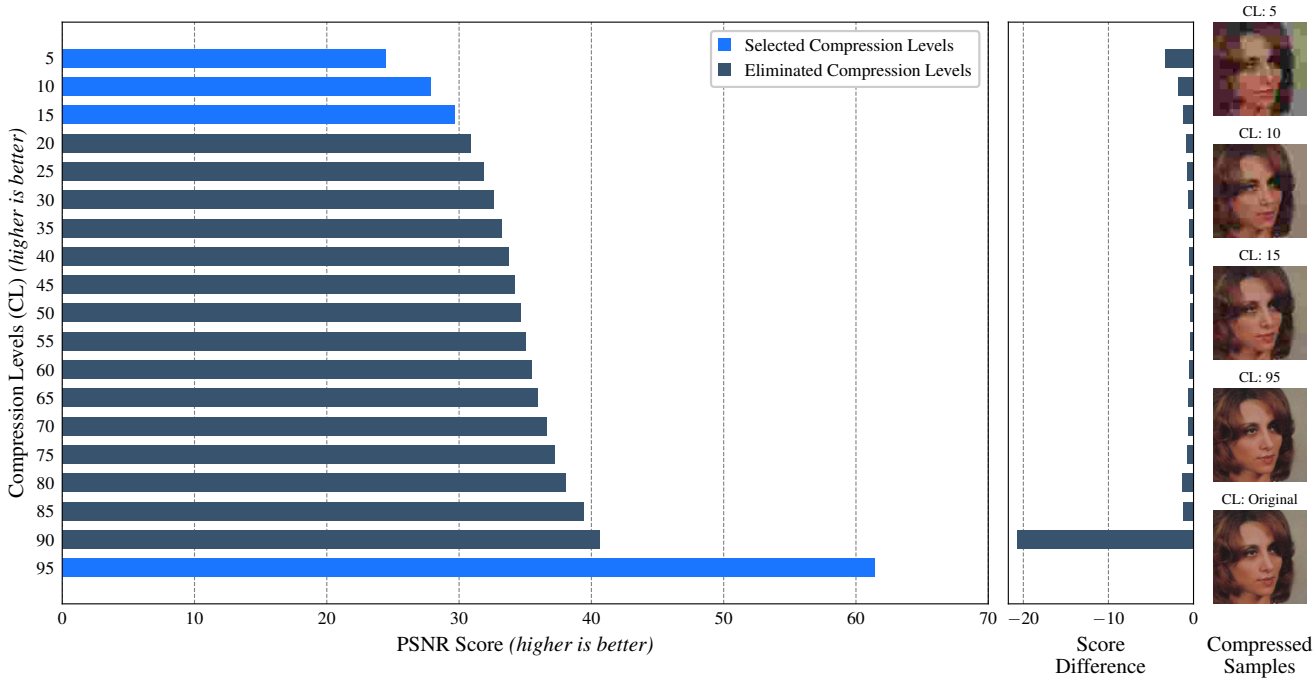

Figure 2. PSNR Scores of RFW dataset on different compression levels (CL). Relative score difference shows how much the quality changes at each level.

compare the default 4:2:0 subsampling with the 4:4:4 no chroma-subsampling factor, which keeps luminance and colour information in its entirety (i.e. unchanged). The rationale behind this evaluation is that if chroma subsampling has a profound impact on recognition performance, we can avoid this issue by recommending the use of 4:4:4 (no chroma-subsampling) with only a small impact on compression performance.

2.4. Compression Level Selection

In order to ascertain the impact of lossy compression on face recognition performance, we are interested in the resulting reduction in image quality at varying levels of JPEG compression. Consequently, we analyse uniformly distributed compression levels on the RFW benchmark face recognition dataset [42] using PSNR; Peak signal-to-noise ratio [43]. PSNR score is correlated with the quality of reconstruction of lossy JPEG compression. In Figure 2, we show the relation between the PSNR score versus the JPEG compression level, q. Firstly, we uniformly select levels $q = \{5...95\}$ in intervals of 5 and compress the whole dataset to each of these JPEG compression levels. Secondly, we measure the PSNR score on all levels and highlight the relative score difference. Based upon this analysis, we downselect the set of JPEG compression levels ($q = \{5, 10, 15\}$), in which quality decrease is most apparent (PSNR score decreases harshly). In addition, we select $q = 95$ as it represents the case where there is no chroma down-sampling used within the lossy compression scheme.

2.5. Training Strategies

We design different test scenarios to measure the impact of image compression on face verification performance.

Racially Imbalanced Dataset: Firstly, we train ArcFace [37] with ResNet101v2 [44] on the original aligned VGGFace2 benchmark dataset [41], containing 3.3 million images with 8631 subjects where subject distribution is racially imbalanced. Subsequently, we test using the RFW benchmark dataset [42] with the original (aligned) images and compressed images to each of the previously down-selected JPEG compression levels. We then repeat the training on the VGGFace2 benchmark dataset [41] four times, having first compressed the entire dataset to each of the down-selected JPEG compression levels. This results in four ArcFace models, each trained on image samples at a different JPEG compression level. Subsequently, we measure the performance of each of these four trained ArcFace models using the RFW benchmark dataset [42] that has been compressed to the corresponding JPEG compression level upon which each of the models was trained.

Racially Balanced Dataset: Similar to the imbalanced train set strategy, we train ArcFace [37] with ResNet50 on the original aligned BUPT-Balanced benchmark dataset [11] that contains 28000 face subjects containing balanced racial distributions among four groups {*African, Asian, Indian, Caucasian*} with 7000 subjects each. Subsequently,

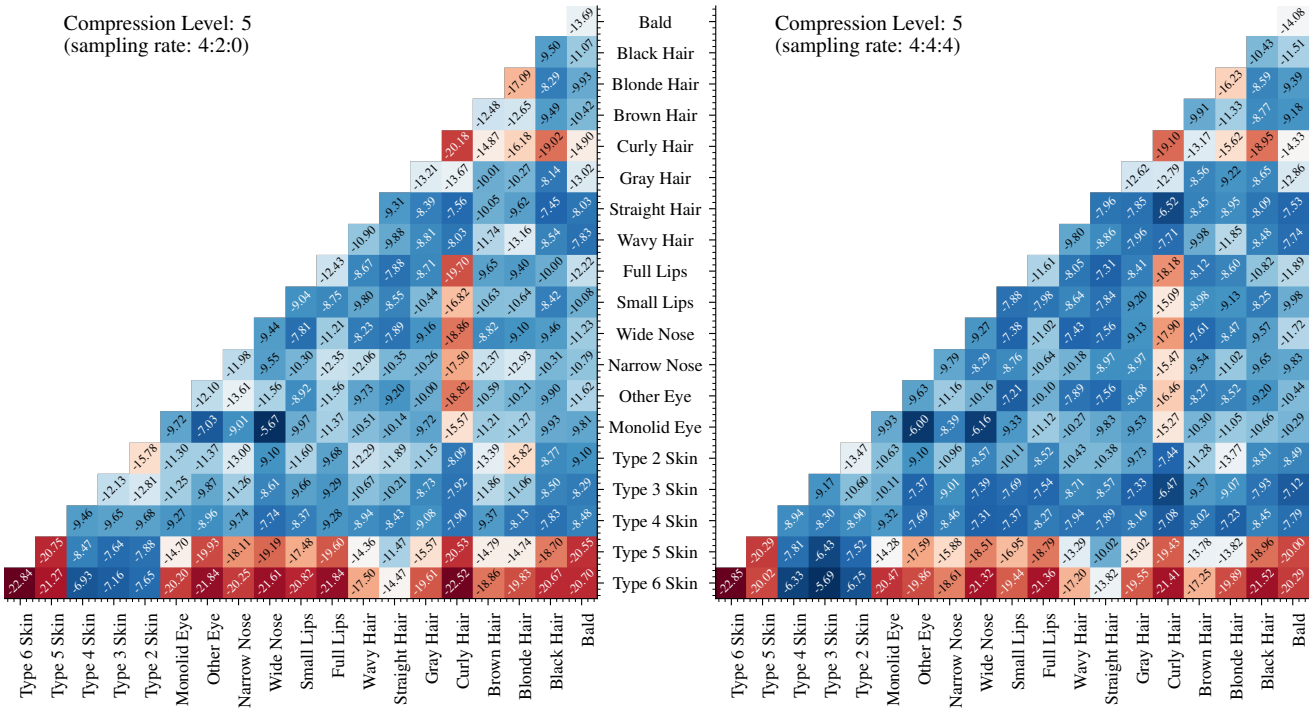

Figure 3. BUPT-Balanced non-compressed, compressed RFW test imagery (q=5); FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

we repeat the training on the BUPT-Balanced benchmark dataset [11] four times, having first compressed the entire dataset to each of the same down-selected JPEG compression levels. This way, another four ArcFace models are trained on image samples at a different JPEG compression level. Additionally, we replicate non-compressed and compressed training at level 5 ($q = 5$) by removing chroma subsampling (4:4:4) to measure the impact of the colour reduction step in lossy compression on face verification performance.

3. Results and Discussion

This section provides extensive experimental results to understand the impact of chroma subsampling and compressed training imagery using two different dataset training datasets and different compression levels. Additionally, we place extended results in Supplementary Material.

3.1. False Verification Matching Rates

In this section, we present False Matching Rate (FMR) differences for each of the proposed training strategies in Section 2.5 and the down-selected compression levels (Figure 2). FMR is a critical metric, such that any change in performance may result in false facial verification and the associated consequences [5].

Figures 3, 4, 5 show the FMR changes under the varying sampling rates of lossy image compression and how this varies across the racial phenotype labels associated with the dataset. Using the cross attribute pairings provided by [17], we evaluate $FMR_{original} - FMR_q$ where $FMR_{original}$ is FMR of non-compressed training and test imagery. FMR_q is the FMR of compressed or non-compressed training but compressed test imagery at down-selected level q. Smaller (and negative) values indicate a more considerable decline from the original level of performance.

Compression Levels: We observe that for all down-selected compression levels $q = \{5, 10, 15, 95\}$, the FMR increases when additional lossy compression is applied, demonstrating that compression level 5 (the highest compression rate) results in the most significant decrease in FMR performance, whilst compression level 95 (the lowest compression rate) does not result in any noticeable FMR performance differences. We compare compression levels 95, 15, 10 and 5 with baseline results to show how FMR rise at higher compression levels. For additional performance results on different levels, see Supplementary Materials.

Chroma subsampling vs No-chroma subsampling We compress all the imagery in the BUPT-Balanced training dataset under two different sampling rates, 4:2:0 (JPEG default) and 4:4:4 on compression level 5 ($q = 5$). The FMR cross-attribute category results are compared in Figures 3, 4, 5. For non-compressed and compressed training,

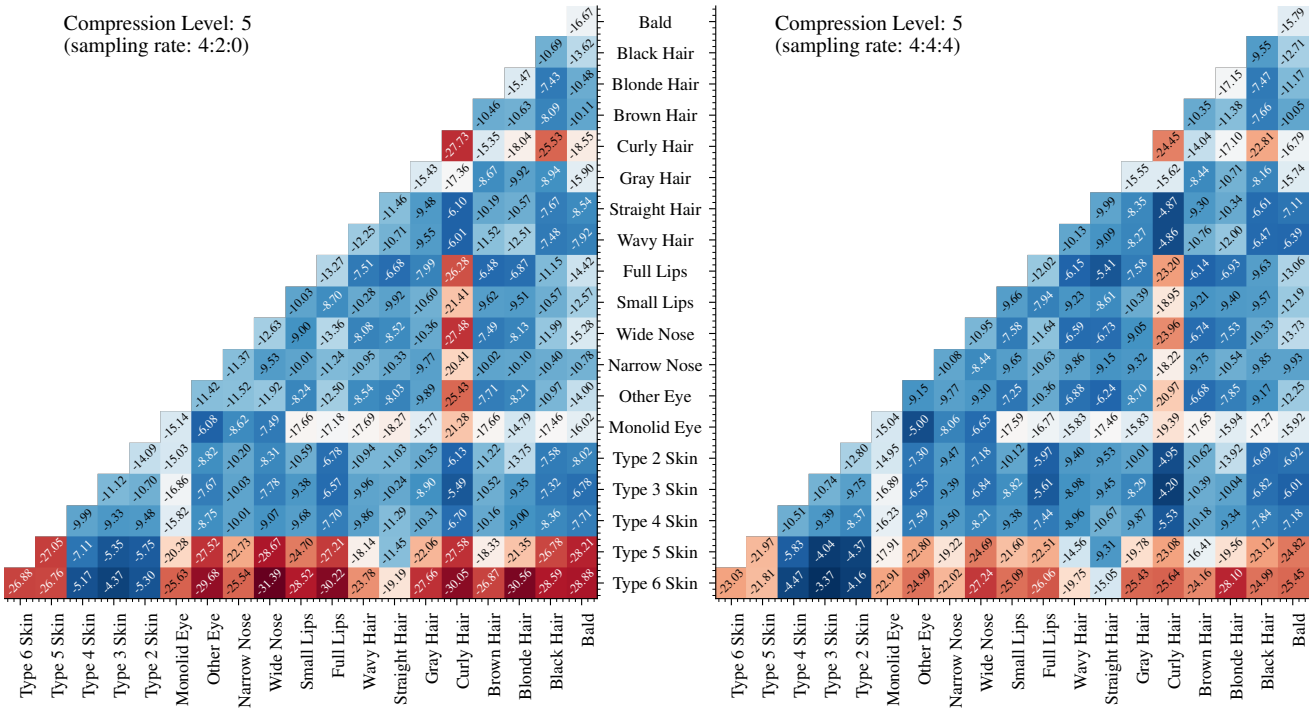

Figure 4. VGGFace2 non-compressed, compressed RFW test imagery ($q = 5$); FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

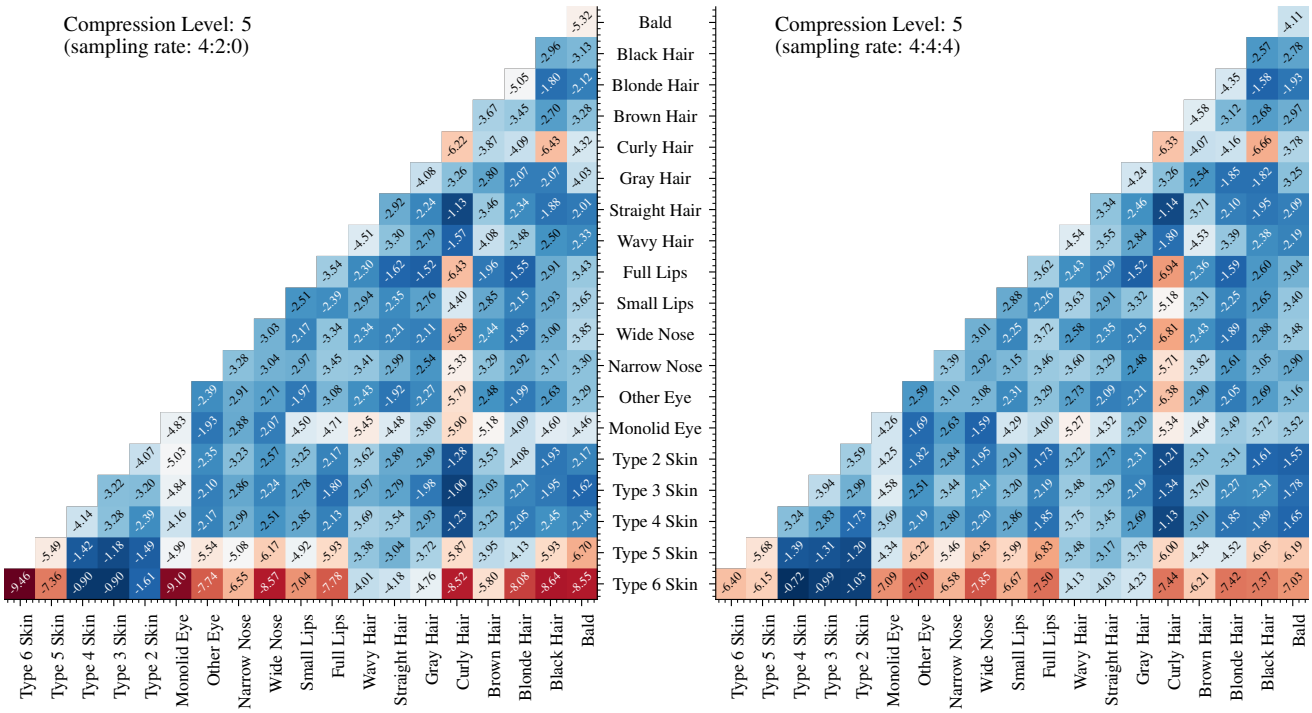

Figure 5. BUPT-Balanced compressed ($q = 5$), compressed RFW test imagery ($q = 5$); FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

	None-Compressed Training Set				Compressed Training Set				
Attribute Name	95	15	10	5	95	15	10	5	Original
Curly Hair	93.10	82.37	75.80	59.53	92.77	87.20	82.90	73.27	93.15
Full Lips	93.37	83.55	77.03	61.37	92.80	87.97	83.62	75.30	93.38
Monolid Eye	93.25	83.43	77.28	63.18	93.48	87.62	85.10	76.95	93.30
Type 5	94.87	85.98	80.17	60.32	94.53	90.22	87.03	76.97	94.85
Type 6	94.85	86.55	79.35	61.75	94.43	90.02	86.20	77.72	94.82
Black Hair	93.70	85.13	79.97	65.83	93.50	89.55	86.87	77.92	93.73
Wide Nose	93.95	85.53	79.97	63.15	93.42	89.57	86.78	78.33	93.98
Other Eye	94.32	86.65	81.10	65.28	93.70	89.57	87.43	78.55	94.38
Type 4	94.05	87.72	83.47	67.28	93.72	89.67	87.45	79.23	94.07
Type 1	92.86	86.88	84.72	72.43	94.19	89.87	88.21	79.57	92.86
Straight Hair	94.18	86.70	81.98	66.15	93.92	89.43	86.28	79.65	94.12
Narrow Nose	94.35	86.30	80.07	66.73	94.60	89.63	87.20	79.77	94.43
Type 3	94.05	86.07	81.03	67.05	94.32	89.48	86.80	79.93	93.98
Small Lips	94.35	87.28	82.03	67.53	95.00	90.63	87.97	81.22	94.37
Wavy Hair	95.87	89.05	84.63	69.53	95.52	92.17	89.33	82.73	95.83
Brown Hair	95.12	88.40	83.33	67.32	95.23	91.85	89.03	82.80	95.15
Bald Hair	96.55	90.43	85.93	67.62	95.88	93.07	90.37	83.13	96.55
Red Hair	96.91	90.57	84.97	71.20	96.33	92.49	89.98	84.89	96.91
Type 2	96.27	89.98	85.98	68.45	96.57	94.27	91.58	85.93	96.33
Gray Hair	96.53	92.47	88.83	72.60	96.42	94.35	91.93	86.75	96.55
Blonde Hair	97.15	92.50	88.52	71.55	97.15	94.83	93.40	87.85	97.15
Mean Accuracy	94.74	87.31	82.20	66.47	94.64	90.64	87.88	80.40	94.76
STD	1.31	2.76	3.58	3.85	1.27	2.18	2.61	3.81	1.31

Table 2. Verification performance on RFW test set using uncompressed (left) and compressed (right) training imagery. Attribute-based pairings are those from the study of [17].

the 4:4:4 sampling rate decreases the FMR for all phenotype categories meaning that removing chroma sampling within the image encoding strategy of the lossy compression technique improves the performance difference and reduces the prevalence of the bias. Accordingly, we evaluate the average FMR for each phenotype category and calculate the standard deviation across all categories. Indeed, for both training strategies in Figure 4 and 5, using no chroma-sampling improves FMR variation across all categories. For VGGFace2 non-compressed training (Figure 4), standard deviation drops from 3.91 to 3.28 $(15.95\% \downarrow)$, whilst BUPT compressed training (Figure 5) standard deviation drops from from 0.91 to 0.81 $(10.88\% \downarrow)$.

Non-compressed vs compressed training sets: When the model is trained on original/non-compressed training imagery (Figures 3 and 4), FMR on darker skin tone (Type 5-6) increases considerably compared to other phenotypes such as lighter skin tones (Types 2-4) with the introduction of lossy compression at test time. At the highest level of compression $(q = 5)$, the increase in FMR is greater when both phenotype categories in the pair are correlated with the stereotypically African/Afro-Caribbean racial features [45]. For instance, the Full Lips $\leftrightarrow$ Type 6 pair has the highest FMR among all other pairs higher than Type 2 $\leftrightarrow$ Type 6 skin tone pairings. For compressed training imagery (Figures 5 and Supplementary S3), we observe improved results for both imbalanced and balanced dataset training. However, darker skin tone and related categories still maintain FMR higher than the other phenotype categories.

Racially balanced vs imbalanced training sets: Using the racially balanced dataset for training does not ameliorate FMR differences among such pairings. For example, at the highest level of compression $(q = 5)$, the average performance decrease of all skin tone Type 5 pairings (Type 5-Bald, Type 5-Black Hair etc.) is 16.06% for imbalanced dataset training (Figure 3). At the same time, it is decreases by 17.69% (Figure 4) from balanced dataset training. However, in racially imbalanced training, the FMR results for pairings with monolid eyes degrade more compared to racially balanced training. As there are significantly fewer monolid eye face samples than other phenotypes in the imbalanced VGGFace2 dataset, we assume that their representation degrades more than other phenotypes as the lossy compression level increases.

3.2. Attribute-based Verification vs. Compression Levels

We additionally present attribute-based verification accuracy for the down-selected compression levels applied at training and test time for the BUPT-Balanced benchmark dataset [11]. Moreover, we provide supporting evidence of compressed *vs.* uncompressed training set face verification performance in Table 2. We use the same 6000 (3000 positive 3000 negatives) attribute-based image pairings provided by [17]. For both non-compressed and compressed training setups, we show that as the compression increases, the standard deviation across all phenotype categories increases (as a measure of non-uniform performance and bias). Similarly, accuracy decreases for all phenotype categories. However, using uncompressed training imagery (Table 2, left) results in a further decline in performance for darker skin tones Type 5-6, curly hair, full lips and monolid eye, when compared to other facial phenotypes, as the level of lossy compression within the test set is increased. Skin Type 5 attribute pairings accuracy drops from 94.87% to 60.32% (34.55% $\downarrow$), while Skin Type 2 attribute accuracy drops from 96.33% to 68.45% (27.88% $\downarrow$). Similar to the non-compressed training set, we do observe non-uniform disparate changes in accuracy when the model is trained on compressed imagery (Table 2, right). Furthermore, the compressed training set produces a smaller standard deviation in accuracy between phenotype categories.

Lastly, we summarise the relationship between all factors (dataset distribution, compression, chroma subsampling) in Figure 6. We evaluate attribute-based pairings accuracy for all phenotype categories and compare different training strategies mean accuracy and standard deviations. We change one factor during training in each strategy and provide corresponding performance results. We use a compressed RFW test set in level 75 ($q = 75$) for all training strategies. Firstly, we show racially imbalanced VGGFace2 datasets training performance, which is lowest in accuracy and highest in standard deviation. A balanced BUPT-Balance dataset provides the most significant improvement in accuracy and standard deviation. Furthermore, while compressed training imagery causes a minor decrease in standard deviation, no-chroma subsampling improves bias performance more significantly. Therefore, removing chroma sampling during compression becomes viable for reducing racial performance bias. We conclude from the abovementioned results that while compressed imagery or racially balanced training data during training improves the overall performance for all race-related categories, disparate results remain for specific phenotype characteristics. Furthermore, we highlight that the reduced retention of the chroma (colour) information affects, due to the use of chroma subsampling in lossy JPEG compression, on darker skin tones to a greater degree than on lighter skin tones. Furthermore, it is likely that the lossy image quantisation disproportionately affects finer image details on the facial region, such as those associated with monolid eye characteristics. Both areas are for further future work.

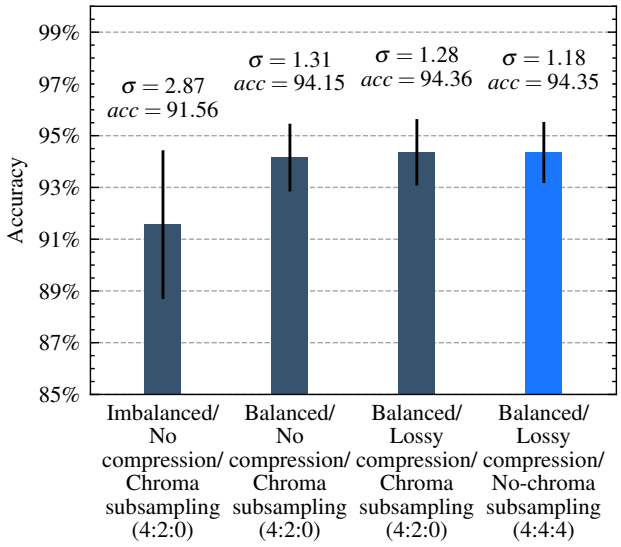

Figure 6. Mean Accuracy and standard deviation of all attribute categories and their comparison on different training strategies using compressed ($q = 75$) RFW test set.

4. Conclusion

This study examines the relationship between face verification performance for a given race-related phenotypic group under varying levels of lossy compressed sets. Overall, our evaluation finds that using lossy compressed facial image samples at inference time decreases performance more significantly on specific phenotypes, including dark skin tone, wide nose, curly hair, and monolid eye across all other phenotypic features. However, the use of compressed imagery during training does make the resulting models more resilient and limits the performance degradation encountered: lower performance amongst specific racially-aligned subgroups remains. Additionally, removing chroma subsampling improves FMR for specific phenotype categories more affected by lossy compression. Future work will explore the impact of lossy image quantisation across various face recognition architectures and propose corresponding results to have fair face recognition algorithms.

Ethical Considerations: This work aims to investigate the impact of lossy compression algorithms on phenotype-based racial groups from [17] to provide additional insight and understanding to guide the mitigation of bias in the development of future face recognition algorithms and systems. We conduct our experiments on three different face datasets publicly available for research use only. The reader is directed to the source publication and the associated research organisation for access to these datasets.

References

[1] P. Grother, M. Ngan, and K. Hanaoka, Face recognition vendor test (frvt) part 3: Demographic effects 2019. 1

[2] S. Shiaeles, Facebook will drop its facial recognition system–but here's why we should be sceptical *The Conversation*, 2021. 1

[3] J. Menn, Microsoft turned down facial-recognition sales on human rights concerns *UK Reuters*, 2019. 1

[4] D. Castelvecchi, Is facial recognition too biased to be let loose? *Nature*, vol. 587, no. 7834, pages 347–350, 2020. 1

[5] P. Grother, M. Ngan, and K. Hanaoka, Ongoing face recognition vendor test (frvt) part 1: Verification 2019. 1, 5

[6] P. Grother, M. Ngan, and K. Hanaoka, Face recognition vendor test (frvt) part 2: Identification 2019. 1

[7] I. Serna, A. Morales, J. Fierrez, and N. Obradovich, Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning *Artificial Intelligence*, vol. 305, page 103682, 2022. 1

[8] M. Wang, Y. Zhang, and W. Deng, Meta balanced network for fair face recognition *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[9] S. Gong, X. Liu, and A. K. Jain, Mitigating face recognition bias via group adaptive classifier in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3414–3424, 2021. 1

[10] T. Sixta, J. Jacques Junior, P. Buch-Cardona, E. Vazquez, and S. Escalera, Fairface challenge at eccv 2020: Analyzing bias in face recognition in *Proceedings of the European Conference on Computer Vision*, pages 463–481, Springer, 2020. 1

[11] M. Wang and W. Deng, Mitigating bias in face recognition using skewness-aware reinforcement learning in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020. 1, 4, 5, 8

[12] M. Georgopoulos, J. Oldfield, M. A. Nicolaou, Y. Panagakis, and M. Pantic, Mitigating demographic bias in facial datasets with style-based multi-attribute transfer *International Journal of Computer Vision*, pages 2288–2307, 2021. 1

[13] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon, Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020. 1

[14] E. Tartaglione, C. A. Barbano, and M. Grangetto, End: Entangling and disentangling deep representations for bias correction in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13508–13517, 2021. 1

[15] A. R. Joshi, X. Suau Cuadros, N. Sivakumar, L. Zappella, and N. Apostoloff, Fair SA: Sensitivity analysis for fairness in face recognition in *Proceedings of the Algorithmic Fairness through the Lens of Causality and Robustness*, Proceedings of Machine Learning Research, pages 40–58, PMLR, 2022. 1

[16] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pages 101–111, 2020. 1

[17] S. Yucer, F. Tektas, N. Al Moubayed, and T. P. Breckon, Measuring hidden bias within face recognition via racial phenotypes in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 995–1004, 2022. 1, 2, 3, 5, 7, 8,

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, A survey on bias and fairness in machine learning *ACM Computing Surveys*, vol. 54, no. 6, pages 1–35, 2021. 1

[19] B. D. I. Formats-Part, 5: Face image data *ISO/IEC JTC1/SC37 N506, ISO/IEC IS 19794*, vol. 5, 2004. 1

[20] S. Vaudenay and M. Vuagnoux, About machine-readable travel documents *Journal of Physics: Conference Series*, 2007. 1

[21] G. K. Wallace, The JPEG still picture compression standard *Commun. ACM*, 1991. 1, 2, 3

[22] A. Skodras, C. Christopoulos, and T. Ebrahimi, The JPEG 2000 still image compression standard *IEEE Signal Processing Magazine*, 2001. 1, 2

[23] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still image data compression standard*. Springer Science & Business Media, 1992. 1

[24] S. Dodge and L. Karam, Understanding how image quality affects deep neural networks in *Proceedings of the International Conference on Quality of Multimedia Experience*, pages 1–6, IEEE, 2016. 1

[25] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, Examining the impact of blur on recognition by convolutional networks *ArXiv*, vol. abs/1611.05760, 2016. 1

[26] M. Koziarski and B. Cyganek, Impact of low resolution on image recognition with deep neural networks: An experimental study *International Journal of Applied Mathematics and Computer Science*, 2018. 1

[27] S. Dodge and L. Karam, A study and comparison of human and deep learning recognition performance under visual distortions in *Proceedings of the International Conference on Computer Communations and Networks*, pages 1–7, Institute of Electrical and Electronics Engineers Inc., 2017. 1

[28] R. Torfason, F. Mentzer, E. Ágústsson, M. Tschannen, R. Timofte, and L. V. Gool, Towards image understanding from deep compression without decoding in *Proceedings of the International Conference on Learning Representations*, 2018. 1

[29] M. Poyser, A. Atapour-Abarghouei, and T. P. Breckon, On the impact of lossy image and video compression on the performance of deep convolutional neural network architectures in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 2830–2837, IEEE, 2021. 2

[30] F. G. Zanjani, S. Zinger, B. Piepers, S. Mahmoudpour, P. Schelkens, and P. H. N. de With, Impact of jpeg 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images *Journal of Medical Imaging*, vol. 6, no. 2, page 027501, 2019. 2

[31] G. Quinn and P. Grother, Performance of face recognition algorithms on compressed images 2011. 2

[32] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, How image degradations affect deep cnn-based face recognition? in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, IEEE, 2016. 2

[33] F. Yang, Q. Zhang, M. Wang, and G. Qiu, Quality classified image analysis with application to face detection and recognition in *Proceedings of the International Conference on Pattern Recognition*, pages 2863–2868, 2018. 2

[34] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, Face quality estimation and its correlation to demographic and non-demographic bias in face recognition in *Proceedings of the IEEE International Joint Conference on Biometrics*, pages 1–11, IEEE, 2020. 2

[35] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay, Biometric quality: Review and application to face recognition with faceqnet *ArXiv*, vol. abs/2006.03298, 2020. 2

[36] P. Majumdar, S. Mittal, R. Singh, and M. Vatsa, Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3786–3795, 2021. 2

[37] J. Deng, J. Guo, N. Xue, and S. 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. 2, 4

[38] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, Towards a critical race methodology in algorithmic fairness in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, Association for Computing Machinery, 2020. 2

[39] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. 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, Association for Computing Machinery, 2020. 2

[40] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker, How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis *Proc. ACM Human-Computer Interaction*, vol. 4, no. CSCW1, pages 1–35, 2020. 2

[41] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, Vggface2: A dataset for recognising faces across pose and age in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018. 3, 4

[42] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, Racial faces in the wild: Reducing racial bias by information maximization adaptation network in *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019. 3, 4

[43] A. Mittal, A. K. Moorthy, and A. C. Bovik, No-reference image quality assessment in the spatial domain *IEEE Transactions on image processing*, vol. 21, no. 12, pages 4695–4708, 2012. 4

[44] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[45] Z. Zhuang, D. Landsittel, S. Benson, R. Roberge, and R. Shaffer, Facial anthropometric differences among gender, ethnicity, and age groups *Annals of occupational hygiene*, vol. 54, no. 4, pages 391–402, 2010. 7

Does lossy image compression affect racial bias within face recognition?

Seyma Yucer, Matt Poyser, Noura Al Moubayed, Toby P. Breckon
Department of Computer Science, Durham University
Durham, UK

5. FMRs on Selected Compression Levels

We provide down-selected compression levels differences (additional compression levels $(q = 10, 15, 95)$) for each of the proposed training strategies using cross attribute pairings provided by [17].

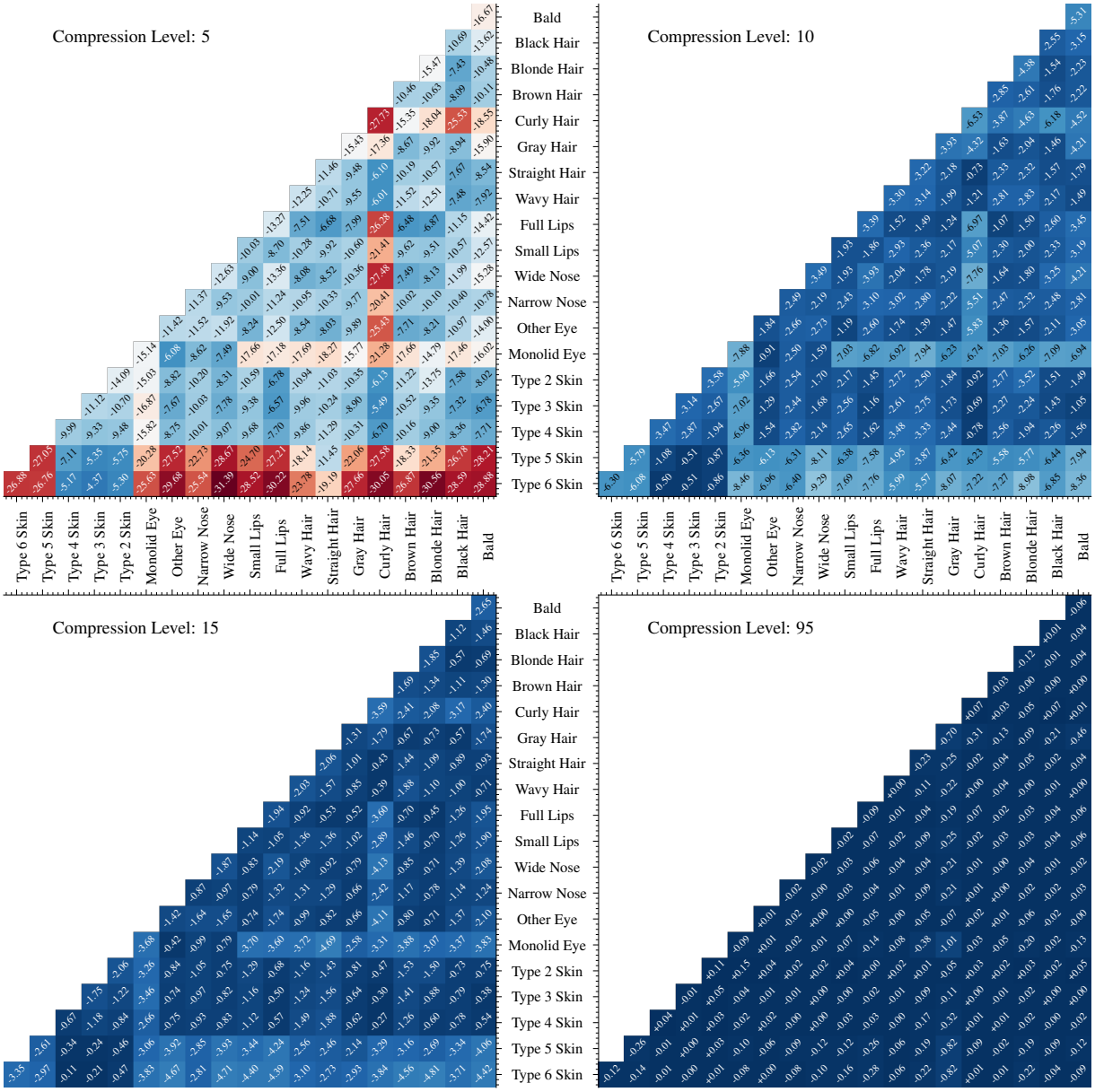

Figure S1. VGGFace2 original/non-compressed training imagery and compressed RFW test imagery; FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

As described in the paper, smaller (and negative) values indicate a larger decline from the original level of performance.The FMR increases when the lossy compression increases. In Figure S1, S2, S3 and S4, we demonstrate that compression level 5 (the highest compression rate) results in the most significant decrease in FMR performance for all different training strategies. In contrast, compression level 95 (the lowest compression rate) does not result in any noticeable FMR performance differences.

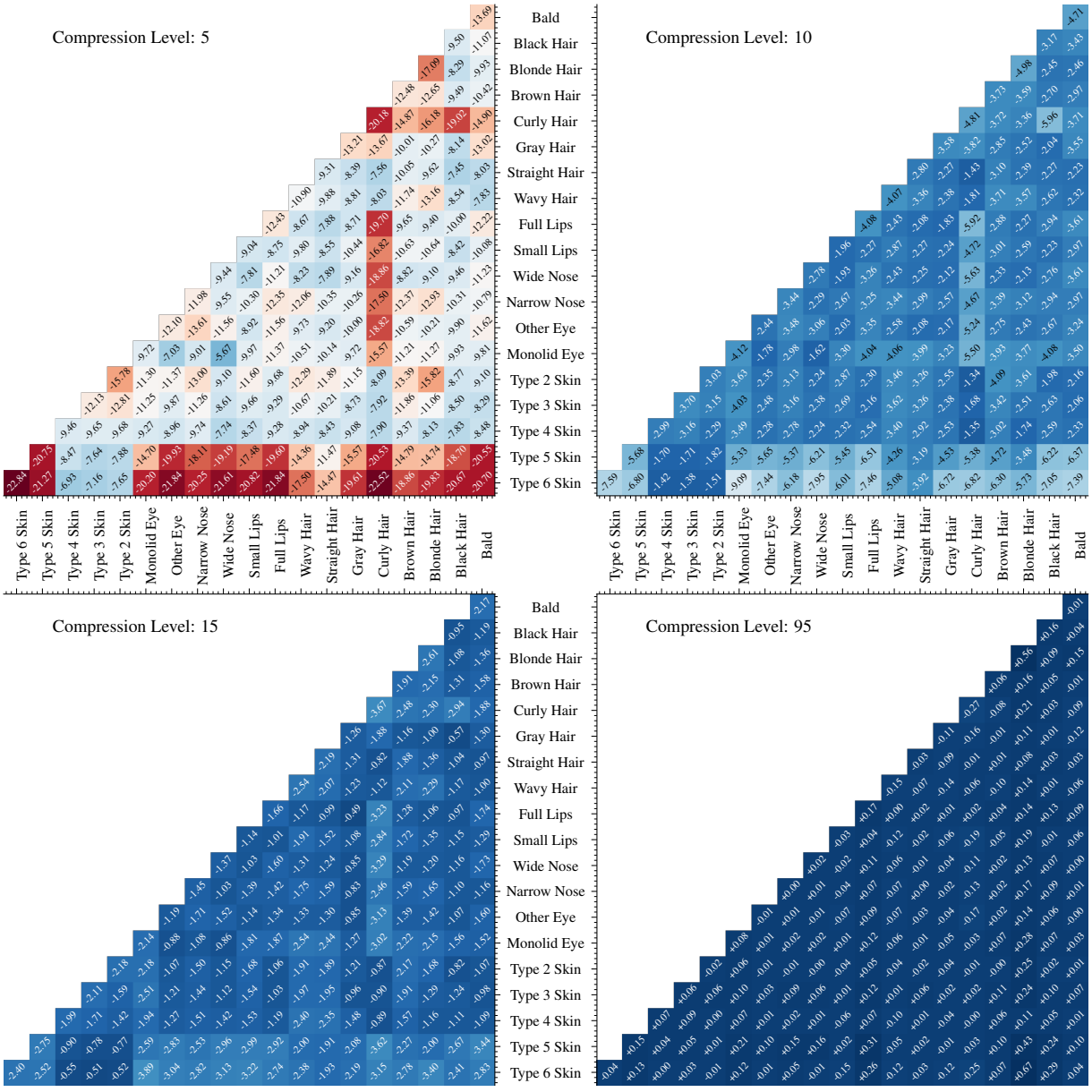

Figure S2. BUPT-Balanced original/non-compressed training imagery and compressed RFW test imagery FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

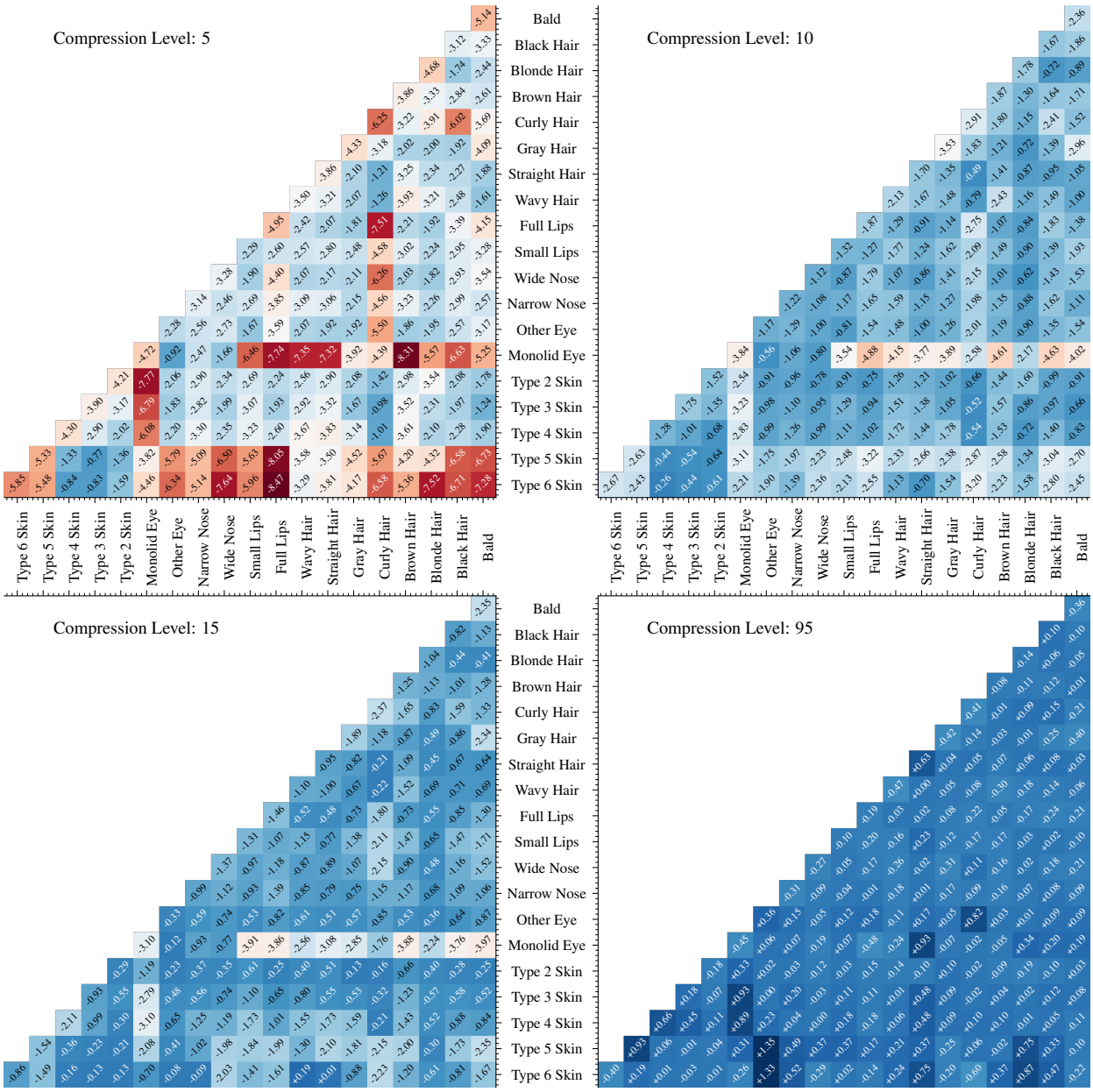

Figure S3. VGGFace2 compressed training imagery and compressed RFW test imagery; FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.

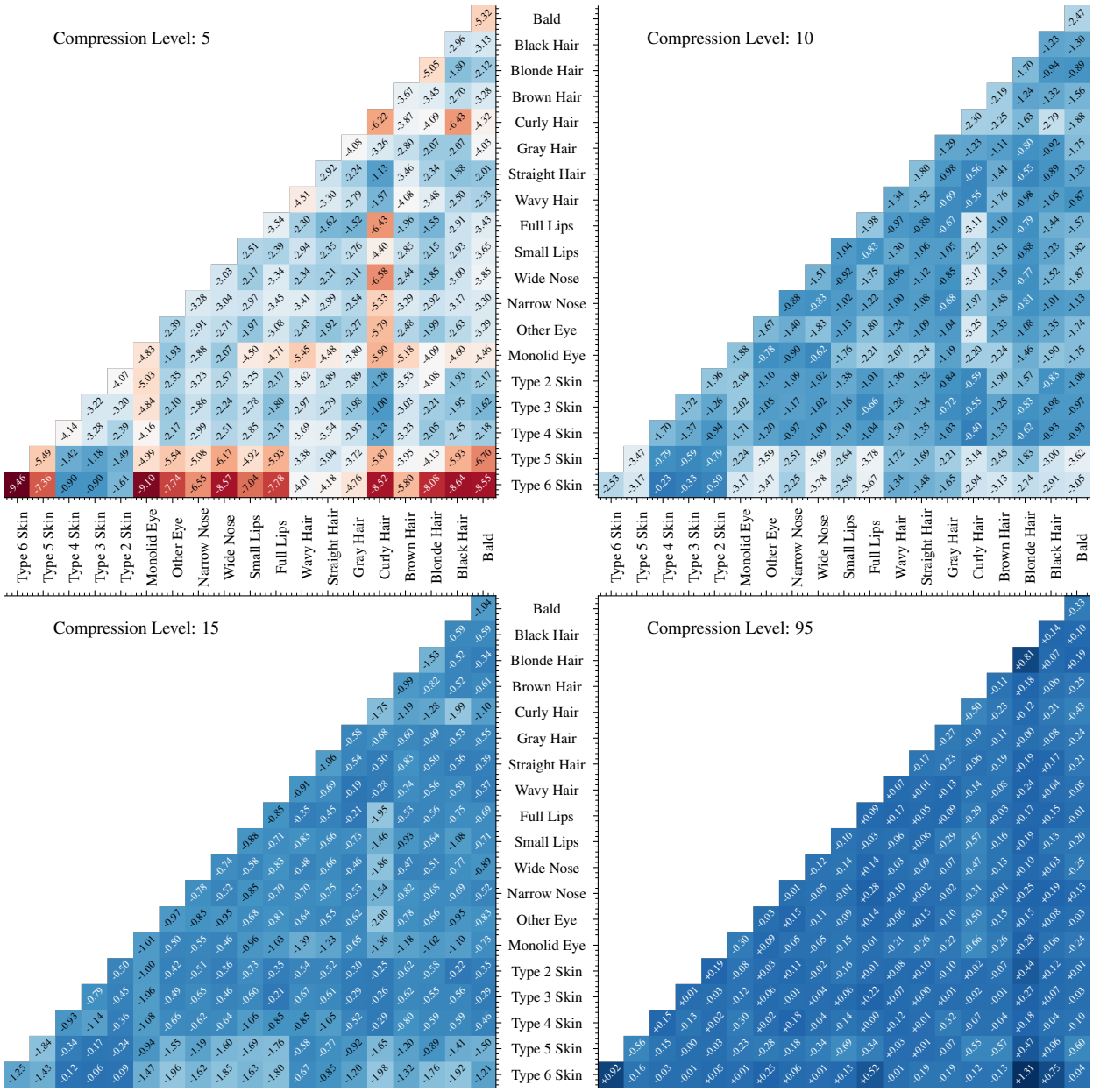

Figure S4. BUPT-Balanced compressed training imagery and compressed RFW test imagery; FMR performance differences of cross-attribute based pairings. Each cell depicts $FMR_{original} - FMR_q$.