F. Boutros, N. Damer, M. Fang, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and M. Todisco (Eds.): BIOSIG 2024,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2024

Detection and Mitigation of Bias in Under Exposure Estimation for Face Image Quality Assessment

André Dörsch[1] Christian Rathgeb[1] Marcel Grimmer[1] Christoph Busch[1]

Abstract: The increasing employment of large scale biometric systems such as the European "Entry-Exit System" and planned national initiatives such as the "Live Enrolment" procedure require quality assessment algorithms to ensure reliable recognition accuracy. Among other factors, facial image quality and hence face recognition accuracy can be negatively impacted by underexposure. Therefore, quality assessment algorithms analyse the exposure of live-captured facial images. To this end, mainly handcrafted measures have been proposed which are also referenced in current standards. However, this work shows that handcrafted measures, which use basic statistical approaches to analyse facial brightness patterns, exhibit racial bias. It is found that these algorithms disproportionately classify images of black people as underexposed as they do not take into account natural differences in skin color, particularly when relying on average pixel brightness values. To ensure fair biometric quality assessment, we have fine-tuned a data-efficient image transformer (DeiT) on synthetic data. The resulting underexposure estimation outperforms state-of-the-art algorithms in detection accuracy and biometric fairness. Precisely, an Equal Error Rate (EER) of approximately 7% is achieved. Our findings highlight the importance of developing robust and fair biometric classification methods to mitigate discrimination and ensure fair performance for all users, regardless of their skin color.

Keywords: Biometrics, Quality Measure, Demographic Differentials, Fairness, Bias, Vision Transformer

1 Introduction

Biometric systems are widely recognized for their reliable and efficient authentication capability in various applications. These systems utilize unique physiological or behavioral characteristics, such as fingerprints, faces, or irises, to automatically verify individuals [Wa05]. For a biometric system to function accurately and ensure interoperability, the quality of the captured biometric sample must be high. To address this need and establish a uniform face quality assessment framework, the German Federal Office for Information Security introduced the Open Source Face Image Quality (OFIQ) software[3]. This software measures the quality of facial images and serves as a reference for algorithms compliant with the international standard ISO/IEC DIS 29794-5 [IS24]. Biometric quality can be measured using two concepts: *Unified quality*, which directly evaluates the overall quality of the facial image considering all variation factors and the concept of *component quality*

[1] da/sec – Biometrics and Security Research Group, Hochschule Darmstadt, Germany,
firstname.lastname@h-da.de

[3] `https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Freie-Software/OFIQ/OFIQ_node.html`

which quantifies how distinct defects (associated with quality components) influence the recognition performance [Gr23].

A crucial quality component of facial image quality assessment is the underexposure measure, ensuring that facial images are suitable for reliable biometric recognition. In this context, two different error types can arise: Type 1, in which the quality component algorithm falsely claims that a normally exposed facial image is underexposed and type 2, in which an underexposed facial image is overlooked by the quality measure algorithm and does (falsely) not lead to an actionable feedback. In this context, error type 1 is more serious, as the data subject is not able to take any action. For example, if the captured facial image of a black subject is wrongly estimated to be underexposed and hence rejected, the person can not take any useful action and may feel discriminated.

In [Dr20], concerns about the demographic fairness of biometric systems have been raised, some of which were even labelled as racist, biased or unfair. In biometric recognition systems, this indicates that the likelihood of false positives and false negatives can vary between demographic groups, caused by demographic performance differentials [Ra22]. The performance of the underexposure quality component also shows considerable demographic differentials between various ethnic groups, hence resulting in different outcomes. We found that facial images of black individuals are more frequently classified as underexposed compared to those of other ethnicities. This discrepancy highlights inherent biases in current methods, underlining the need for developing more robust and fair classification techniques. To mitigate biases observed in multiple underexposure measures and ensuring fair performance across diverse ethnicities, this research proposes a DeiT-based model entirely trained on synthetic data. The proposed method is compared against several algorithms to highlight improvements in both accuracy and fairness.[4]

2 Related Work

To ensure the reliability of face recognition systems, prior assessment of the underlying exposure level in facial images is necessary to avoid poorly illuminated images being processed. Various approaches and algorithms have been proposed in the scientific literature to address this task. We found research on exposure assessment to be widely understudied and dominated by statistical algorithms, which are prone to fairness issues.

In [Wa17], the Absolute Central Moment (ACM) is used to quantify the exposure level of an image by analyzing its histogram and mean intensity value. This algorithm was referenced as exposure measure in the deprecated ISO/IEC 29794-5 Technical Report from 2010 [IS10]. The current draft international standard ISO/IEC 29794-5 [IS24], which is the basis for the OFIQ implementation, considers a face image underexposed if it contains a high proportion of pixels in the low luminance range [0, 25], and overexposed if it contains a high proportion of pixels in the high luminance range [247, 255]. In addition, the work by Hernandez-Ortega et al. [He22] uses a similar statistical approach as proposed in [Wa17]

[4] The source code and data of this work is made available at: `https://github.com/dasec/FIQA-Underexposure-Bias`

to analyse general illumination conditions in face images by detecting whether an image is too dark or too bright based on the mean pixel value of the face. Furthermore, an automated quality assessment framework proposed by Kim et al. [KLR14] addresses potential defects in face images. Their method quantifies brightness by comparing the histogram of the face image to a uniform reference histogram using relative entropy. Images with a relative entropy above a certain threshold are considered to have poor brightness quality and are discarded to improve recognition performance.

In addition to the presented methods for assessing underexposure in face images, there are already studies in the field of face image quality assessment that show that quality algorithms do not always treat individuals equally and consequently exhibit demographic biases (see e.g. [Te20]). In addition, in the recently added section "Quality Measures by Demographic Group" the expanded NIST Face Analysis Technology Evaluation (FATE) Quality Specific Image Defect Detection (SIDD) report [Ya24] examines demographic performance differences for various face quality assessment algorithms, including underexposure quality algorithms. The authors found that several algorithms for assessing underexposure exhibited demographic biases, which underlines this important field and calls for further research.

3 Fairness Evaluation

3.1 Testing Database Preparation

There are several causes for the occurrence of bias in biometric systems. One widely discussed reason can be statistically unbalanced class distributions of training data, which consequently shows a different performance for certain subgroups [Te22]. This could include the underrepresentation of specific classes in the training set (e.g. the underrepresentation of a certain ethnicity), mislabelled data, poor data quality of a certain class and many other reasons that distort the overall data distribution. Another cause might be the implementation of an algorithm itself being biased by poor design, or other disruptive factors [Dr20]. To detect potential bias across demographic groups, a suitable database is required. It is important to mention at this point that the testing database for our purposes should only consist of real image data and should not contain any synthetically generated facial images. To fulfill the requirements for a diverse and balanced testing database of real exposure variations in facial images, we decided to merge several subsets of existing face databases.

In order to create a more representative exposure database for evaluation purposes, we first used a subset of the RFW database [Wa19], which provides a diverse collection of facial images specifically designed for the study of racial bias in face recognition. This database provides a balanced distribution of images across four ethnic groups: Caucasian, Asian, Indian and African, making it particularly suitable for our experiments. Given the primary purpose of the RFW database is to evaluate fairness in face recognition, it includes images with various image quality such as strong compression artifacts, blurriness, excessive padding, unnatural coloring, and cropped or partially occluded faces. These low-quality

images do not represent scenarios where a facial quality assessment, e.g. OFIQ, would be mainly used and could adversely affect the accuracy of our results. Therefore, we thoroughly sorted out low-quality images by manual inspections to ensure that our RFW subset contains facial images of higher quality. We also labelled images that we considered to be underexposed as such. To acquire additional real under- and normal exposed face images for the fused testing database, a custom subset of the CAS-PEAL-R1 database [Ga08] was further used. This face database provides grayscale images of Asian subjects with different exposure variants. Furthermore, we decided to use a custom subset of the FEI-Face database [TG10], a Brazilian face database that contains a set of face images with different exposure variants. Facial images from the FEI-Face database were selected in such a way that for each subject a frontal normal-exposed image and a frontal underexposed image (if available) were used.

Ethnicities of the FEI-Face database were labeled in a way that they match with the four ethnicities from the RFW database. All images were then aligned and scaled to 256×256 pixels. As subsequent post-processing, the images were cropped vertically by 16 pixels and from the upper edge of the image by 32 pixels to 224×224 pixels. Statistics about the testing database can be found in Table 1 and Table 2.

Tab. 1: Distribution of facial images by exposure variations (Normal-, Underexposed) of our testing database

Exposure	**Subset-RFW**	**Subset-CAS-PEAL-R1**	**Subset-FEI-Face**	**Total**
Normal-Exposed	4,071	1,040	200	5,311
Under-Exposed	66	293	128	487

Tab. 2: Distribution of facial images by ethnicity of our testing database

Exposure	**African**	**Asian**	**Caucasian**	**Indian**	**Total**
Normal-Exposed	1,105	1,752	1,103	1,351	5,311
Under-Exposed	35	311	120	21	487

Fig. 1: Sample subjects of Normal- and Underexposed images from our testing database

3.2 Fairness Evaluation Results

To evaluate the performance of the handcrafted exposure algorithms mentioned in section 2 across different demographic groups, we decided to consider OFIQ as it is the most

relevant candidate and should be established as the de-facto standard for facial quality assessment. Following ISO/IEC DIS 29794-5 [IS24] the reference implementation OFIQ provides a mapping from native quality measures (i.e. real numbers) to quality component value (i.e. integer number in the range of 0 to 100), where higher values imply a higher biometric utility. This means that facial images with a higher component value assigned by the OFIQ algorithm are considered normally exposed, while images with a lower component value are considered underexposed.

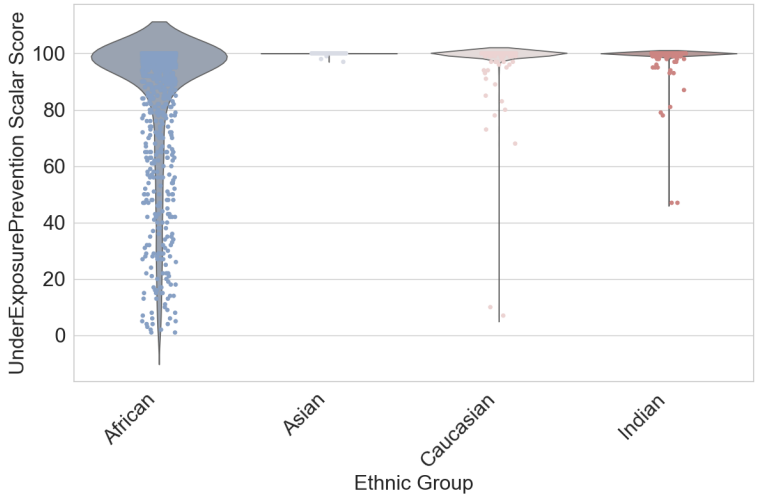

Fig. 2: Distribution of OFIQ's UnderExposurePrevention quality measure on normally exposed facial images by Ethnic Group

Figure 2 shows the distribution of the quality measures of the OFIQ algorithm on exclusively normal exposed facial images from our testing database. A clear trend is observable: dark-skinned subjects (left) often tend to receive low quality measures, even though they are normally exposed. This result points to a significant problem, namely its inherent disadvantage for people with darker skin color. There is currently no metric for measuring fairness in the context of quality components. This finding emphasises the high demand for a standardized quality component-based fairness metric in facial image quality assessment. Corresponding descriptive statistics are shown in Table 3. There is an underlying

Tab. 3: Quality measures statistics from OFIQ component of normally exposed facial images by ethnic group

Ethnic Group	Images	Mean	Std Dev	Min	Max
African	1,105	87.44	22.92	1.0	100
Asian	1,752	100	0.09	97.0	100
Caucasian	1,103	99.64	4.28	7.0	100
Indian	1,351	99.81	2.33	47.0	100

assumption that a high luminance value in a facial image corresponds to poor image qual-

ity without taking into account the differences in skin color. However, this assumption also indicates that faces with dark skin color are disproportionately often incorrectly classified as underexposed, as they tend to have pixels in the low luminance range[5].

4 Proposed Model

4.1 Methodology

A custom synthetic dataset was used to fine-tune a data-efficient image transformer (DeiT) as described by Touvron et al. [To21]. We opted for the DeiT architecture because it offers robust performance even with a limited amount of training data. The specific model for this process was the DeiT base model, which contains 86 million parameters. For the interested reader, the DeiT implementation details and code can be found on the DeiT GitHub repository[6].

The fine-tuned model was then compared with the "UnderExposurePrevention" quality measure from OFIQ, a commercial off-the-shelf (COTS) system, and the exposure algorithm from the Technical Report ISO/IEC TR 29794-5:2010 [IS10] referenced in Section 2.

To make our model comparable to the aformentioned algorithms, we modified the original DeiT classification head, resulting in a binary classification head (Normal-Exposed vs. Under-Exposed). The basis for the UnderExposure quality score was then obtained by the confidence of the trained model for the UnderExposure class.

For the synthetic dataset, a subset of the HDA-SynChildFaces database [Fa24] was used, ensuring a demographically balanced representation of different genders and ethnicities at scale. Since the facial images were sampled from the latent space of StyleGAN3 [Ka21], which typically generates high-quality images with minimal environmental distortions, we also ensure few to no instances of underexposed images in the dataset. Our custom subset was limited to post-filtered images of adults aged 20 and older, as this research does not evaluate younger age groups. From the original set of 17 images per subject, only the first 11 were used. These images exhibit slight variations in pose and expression while maintaining consistent illumination.

Furthermore, we augment our subset by adding synthetic facial images from the Syn-Multi-PIE [CdFPM21], which emulates the Multi-Pie database [Gr08] with entirely synthetic images. For each identity, we used the frontal neutral face image (reference image) as normal exposed image and selected two images with the most significant illumination variation (images 0 and 5) as the basis for creating further underexposed variants. Underexposure was simulated using OpenCV. We applied gamma correction and contrast scaling to the normal exposed images from our HDA-SynChildFaces subset, creating a

[5] We found that all algorithms analysed in this work (including the COTS exposure algorithm) showed a similar trend to OFIQ, namely an inherent disadvantage for people with dark skin color

[6] `https://github.com/facebookresearch/deit`

underexposed version of the images and the selected illumination-varied images from Syn-Multi-PIE. Thus, our training dataset for normal exposed facial images comprised the unmodified subset of HDA-SynChildFaces and the frontal neutral face images from Syn-Multi-PIE. For underexposed images, we included the underexposed modified HDA-SynChildFaces images and the underexposed modified variants of the illumination-varied images from Syn-Multi-PIE. To ensure a comprehensive range of exposure conditions and enhance the robustness of our training, we varied the hyperparameters for exposure adjustments, such as the exposure factor and contrast scaling factor, using values drawn from a uniform distribution. Figure 3 shows a series of facial images of individuals with different exposure variations from our custom training dataset. The top row displays normally exposed images and the bottom row presents the synthetically underexposed images.

Fig. 3: Comparison of Normally-Exposed and Under-Exposed Images from our Training Database

Resulting in a total amount of 59,062 images, divided into 24,531 normal exposed and 34,531 underexposed facial images.

5 Results

As mentioned earlier, we also evaluated our fine-tuned transformer on the normally exposed images from our testing database.

Figure 4 illustrates that, in comparison to the OFIQ ”UnderExposurePrevention” quality measures, the ”African” demographic group no longer faces a noticeable disadvantage. However, there is a slight trend of more values being distributed towards the lower end across all ethnic groups compared to OFIQ quality component. Corresponding descriptive statistics are shown in Table 4.

Tab. 4: Quality measures statistics from fine-tuned DeiT of normally exposed facial images by Ethnic Group

Ethnic Group	Images	Mean	Std. Dev.	Min	Max
African	1,105	99.61	4.28	16	100
Asian	1,752	99.90	1.91	42	100
Caucasian	1,103	99.40	5.15	11	100
Indian	1,351	99.76	3.76	9	100

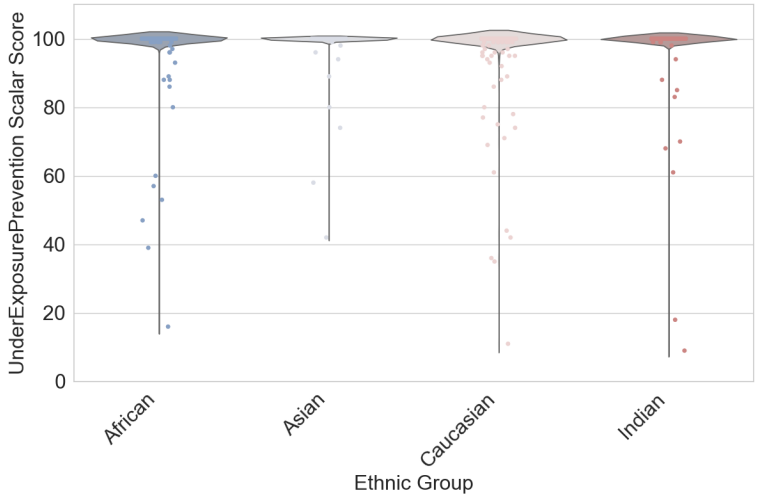

Fig. 4: Distribution of UnderExposurePrevention quality measures from our model on normally exposed facial images by Ethnic Group

The Detection error trade-off (DET) curve in Figure 5 compares the classification performance for classifying the images as normally exposed or underexposed of observed algorithms[7], showing the trade-off between false positive rate and false negative rate. Table 5 summarizes the Equal Error Rate (EER) and Accuracy (1-EER) for each algorithm, highlighting that our method achieves the best performance with the lowest EER of 0.07 and highest accuracy of 0.93.

Tab. 5: EER and Accuracy of the algorithms evaluated

Algorithm	**EER**	**Accuracy**
ISO/IEC TR 29794-5:2010	0.37	0.63
OFIQ quality component	0.11	0.89
COTS	0.09	0.91
DeiT (ours)	**0.07**	**0.93**

6 Conclusion

In this work, we have demonstrated that the transition from statistical to deep neural network-based exposure quality assessment models significantly improve demographic fairness and classification accuracy. Our fine-tuned DeiT model not only mitigates the disadvantages previously observed in certain demographic groups, as evidenced by our evaluations, but also achieves the lowest Equal Error Rate (EER) of 0.07 and the highest accuracy of 0.93 among the tested algorithms.

[7] For DET/Accuracy comparison, exposure scores for COTS and ISO/IEC TR 29794-5:2010 were min-max normalized, while raw scores were used for OFIQ

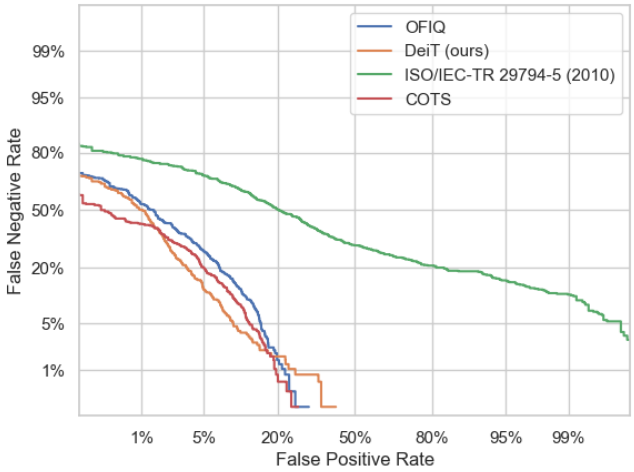

Fig. 5: DET classification plot of observed algorithms

Acknowledgement

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[CdFPM21] Colbois, L.; de Freitas Pereira, T.; Marcel, S.: On the use of automatically generated synthetic image datasets for benchmarking face recognition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). S. 1–8, 2021.

[Dr20] Drozdowski, P.; Rathgeb, C.; Dantcheva, A.; Damer, N.; Busch, C.: Demographic Bias in Biometrics: A Survey on an Emerging Challenge. Trans. on Technology and Society (TTS), 1(2):89–103, June 2020.

[Fa24] Falkenberg, M.; Ottsen, A. B.; Ibsen, M.; Rathgeb, C.: Child face recognition at scale: synthetic data generation and performance benchmark. Frontiers in Signal Processing, 4, 2024.

[Ga08] Gao, W.; Cao, B.; Shan, S.; Chen, X.; Zhou, D. et al.: The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans (TSMCA), 38(1):149–161, January 2008.

[Gr08] Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S.: Multi-PIE. In: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. S. 1–8, 2008.

[Gr23] Grimmer, M.; Rathgeb, C.; Veldhuis, R.; Busch, C.: NeutrEx: A 3D Quality Component Measure on Facial Expression Neutrality. In: Proc. Intl. Joint Conf. on Biometrics (IJCB). IEEE, S. 1–8, 2023.

[He22] Hernandez-Ortega, J.; Fierrez, J.; Gomez, L. F.; Morales, A.; de Suso, J. L. Gonzalez; Zamora-Martinez, F.: FaceQvec: Vector Quality Assessment for Face Biometrics Based on ISO Compliance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. S. 84–92, January 2022.

[IS10] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC TR 29794-5 Information Technology - Biometric Sample Quality - Part 5: Face Image Data. International Organization for Standardization, 2010.

[IS24] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC DIS 29794-5 Information Technology - Biometric Sample Quality - Part 5: Face Image Data. International Organization for Standardization, 2024.

[Ka21] Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T.: Alias-Free Generative Adversarial Networks. In: Proc. NeurIPS. 2021.

[KLR14] Kim, H. I.; Lee, S. H.; Ro, Y. M.: Investigating Cascaded Face Quality Assessment for Practical Face Recognition System. In: 2014 IEEE International Symposium on Multimedia. S. 399–400, 2014.

[Ra22] Rathgeb, C.; Drozdowski, P.; Frings, D. C.; Damer, N.; Busch, C.: Demographic fairness in biometric systems: What do the experts say? IEEE Technology and Society Magazine, 41:71–82, December 2022.

[Te20] Terhörst, P.; Kolf, J.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: Face Quality Estimation and Its Correlation to Demographic and Non-Demographic Bias in Face Recognition. In: Intl. Joint Conf. on Biometrics (IJCB), 2020. September 2020.

[Te22] Terhörst, P.; Kolf, J.; Huber, M.; Kirchbuchner, F.; Damer, N.; Morales, A.; Fierrez, J.; Kuijper, A.: A Comprehensive Study on Face Recognition Biases Beyond Demographics. IEEE Transactions on Technology and Society, 3(1):16–30, 2022.

[TG10] Thomaz, C. E.; Giraldi, G. A.: A new ranking method for principal components analysis and its application to face image analysis. Image and Vision Computing, 28(6):902–913, 2010.

[To21] Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning. Jgg. 139. PMLR, July 2021.

[Wa05] Wayman, J. L.; Jain, A. K.; Maltoni, D.; Maio, D.: Biometric Systems: Technology, Design and Performance Evaluation. Springer London, 2005.

[Wa17] Wasnik, P.; Raja, K. B.; Ramachandra, R.; Busch, C.: Assessing face image quality for smartphone based face recognition system. In: 2017 5th International Workshop on Biometrics and Forensics (IWBF). S. 1–6, 2017.

[Wa19] Wang, M.; Deng, W.; Hu, J.; Tao, X.; Huang, Y.: Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network. In: The IEEE International Conference on Computer Vision (ICCV). October 2019.

[Ya24] Yang, J.; Grother, P. J.; Ngan, M. L.; Hanaoka, K.; Hom, A.: , Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment: Specific Image Defect Detection, July 2024.