Morphing Attack Detection (MAD) based on the location of the corneal specular highlights

LIA STREAPCO, University of Twente, The Netherlands

Generative Adversarial Networks (GAN) are tools that allow for the creation of realistic synthetic images. Their development in the past years has seen significant growth. New methodologies are emerging, including the one examined in this research, face morphing. However, the spread of GAN-generated morphed images has also made the world vulnerable, due to their realism it is a challenge to distinguish them from the real ones. This research paper proposes a way of detecting GAN-generated morph of portrait pictures, by utilizing the specular corneal highlights. The brightest point from the highlight is extracted and this data is then processed through a logistic regression algorithm to determine the authenticity of the portrait picture. The research investigates the effectiveness of the proposed approach by utilizing two different methods of iris extraction, thus creating two different datasets. The results demonstrate promising outcomes and offer a path for further development.

Additional Key Words and Phrases: GAN-generated images, face morphing detection MAD, position, corneal highlights.

1 INTRODUCTION

There has been a massive growth in the quality of AI-generated images. Deep learning methods have paved a path to a high level of realism [6]. However, despite the high level of realism, those are not perfect. The generation process leaves some identifiable artefacts. Some of those can be seen by the naked eye, for example, anomalies in human faces or asymmetries like different earrings, others are not so easily observable, like the corneal highlights or minor unobservable colour differences. More generally, these images present invisible artefacts, closely linked to the architecture of the generative network, which can be extracted through appropriate processing steps [6].

Detection using the eyes and the region around the eye is a current research topic. The objective of this research is to investigate a way to detect GAN-generated images in portrait settings. Such a detection mechanism is proper when detecting photos used for example in passports or other official documents. In document pictures, the portrait setting is rigorously followed and required. The person should look directly into the camera, the light source should be further away and in front of the person. In those circumstances, due to the anatomy of the human eye, the light reflects in the eyes of the person, forming highlights on the eye's cornea.

Due to the setting of the image, the formed highlights should be similar in shape and location, since the light source is directly in front of the person, representing the main hypothesis of the research paper. Thus, the location of the brightest point of the highlights is expected to have the same coordinates in the left and right eyes, minor deviations are possible. However, due to the software limitations GAN-generated images do not reflect those similarities, thus opening an opportunity for detection of the morphing attacks.

The research utilizes GAN-generated morph images. GAN-generated morph images refer to the style of generating which involves blending two or more input images together to create an entirely new one. The systems assure that the transition between the blended parts is done smoothly. This ensures a seamless blend and produces a coherent image. It is interesting to observe that the StyleGAN-based morph generation did not create any perceptual noise[12]. Thus face morphing attacks aim at creating face images that are verifiable to be the face of multiple identities, which can lead to building faulty identity links in operations like border checks [4]. GAN-generated morphs have their limitations, precisely due to the blending process which is especially visible in the highlights. The specific limitations are exploited in this research paper.

The proposed research aims to answer the following research question:

(1) What is the accuracy of detecting GAN-generated portrait settings images based on the location of corneal specular highlights using a logistic regression algorithm?

To provide an answer to the question experiments using diverse datasets were conducted. The datasets contained GAN-generated morph images as well as real images. The hardest part lay in the extraction of the iris itself, thus two different methods of extraction are proposed. One of the proposed methods has higher accuracy, while the other is better when extracting the iris from the generated images. The first method utilizes Hugh circles and the second is based on landmark detection and creating an iris mask. Both those methods are described in detail in the Methods section.

The performance of the approach is analyzed using the accuracy score, F1-score, precision, and recall. Furthermore, the metrics are compared to state-of-the-art technology to assess the efficacy of the proposed method.

The results of this research contribute to the development of robust methods of GAN-generated image detection. Moreover, the analysis of the specular corneal highlights for detecting GAN-generated images has received limited attention from the research community. The advancements have potential implications in such domains as cybersecurity, where the identification of the authenticity of an image plays a primordial role.

TScIT 37, July 8, 2022, Enschede, The Netherlands

2 BACKGROUND AND REVIEW OF LITERATURE

There are three categories of detection of GAN-generated images. The first category is pixel-level analysis, which focuses on statistical techniques for analyzing the pixel distribution in the image. One of the most recent methods involves generating the image from the noise vectors, instead of using the convolutional neural network. The second category includes model-based analysis, which involves the training of an AI model, those work well, but are hard to generalize. This phenomenon has been shown both in [3] and in [10], where some interesting experiments are carried out that highlight the inability of both handcrafted and data-driven features to support cross-dataset generalization [6]. Methods of the third category look for physical/physiological inconsistencies by GAN models [7]. Such physiological/physical-based detection methods are more robust to adversarial attacks and afford intuitive interpretations [7].

The eye region has been used in the forensics analysis. Works describing how to estimate the 3-D direction of a light source from specular highlights on the eyes [8] have been done. Works like [11] show that reflections in the eyes are either missing or appear simplified as a white blob. However, that work was done on the earlier version of generators. Currently, the generation techniques have been improving and are becoming more evolved [9] shows how the generative models are being improved by analyzing the artefacts and making changes in the architecture of the GAN model itself.

3 METHODOLOGY

3.1 Data

All the data was taken from the FRLL-MORPHS dataset, created from the publicly available Face Research London Lab dataset [5]. StyleGan2 generated the images used. The dataset also contained real images, which were divided into neutral and smiling faces.

All the pictures are in portrait settings, which means that they correspond to the following criteria:

(1) The face of the person is directly facing the camera.
(2) The light source is directly in front of the person and within an adequate distance.
(3) The light sources are visible to both eyes.

All the images are in jpg format and have the dimensions of 1024x1024.

3.2 Data Preprocessing

The first step taken in the preprocessing process was to identify and crop out the iris. This was done to ensure the correct extraction of the highlights. First, the iris is cropped, this is done to ensure that no noise is impeding the detection of specular corneal highlights. Afterwards, the coordinates of the brightest pixel were found. It was enough to find the brightest pixel without needing extra steps since the highlights represent the brightest spots on an image.

To find the iris, two separate methods were used, thus creating two separate datasets. Using the two separate methods is justified by the fact that one of the methods works best on GAN-generated images, while the second works best on bonafide images. So to ensure, the integrity of data, we decided to create two separate datasets, one by employing the first method and the second one with method two.

Below the extraction methods, as well as finding the brightest pixel steps are described in detail. The used library is OpenCv and the analysis was done in Python programming language.

The final step of data preprocessing is to split the data into training and testing sets. The training set is represented by 80% of the data, while the testing one is the remaining 20%. The total number of images varies per extraction method and can be found in 3.3.1 and 3.3.2 respectively. To make the dataset more uniform and to reduce the gap between the number of generated and bonafide images, only half of the generated images was used. The half was chosen at random. This ensured that the model was not overfitted.

3.3 Iris Extraction

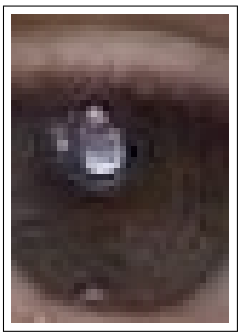

Fig. 1. Left Eye

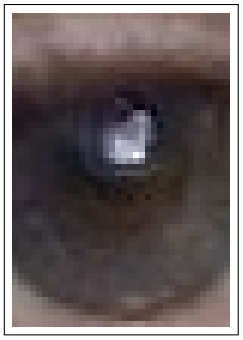

Fig. 2. Right Eye

3.3.1 Method 1 of Iris Extraction. The first method of iris extraction involves the usage of Hough Circles [2].

First, the face is detected, and after that the eyes. Both actions were done through the cascade classifier of the cv2 library. Because sometimes only one eye is detected, a preliminary check is done to determine whether both eyes were detected. The eyes from the images are then cropped and transformed to grayscale. After that, the Hough Circle function from the cv2 library is called. This function finds the circles in an image. the minimum circle radius given to the function is 20 and the maximum is 30. In general, all the parameters were derived by trial and error method, in which the images were analyzed and seen which parameters need to be passed to the function to achieve the best outcome.

The function outputs the centres of the circles which correspond to the iris. Then the iris was cropped and saved as left 1 and right 2. To better align the coordinates the right images were flipped. All the saved images are numbered and then sorted.

This method outputs 895 eyes of GAN-generated images and 81 real images.

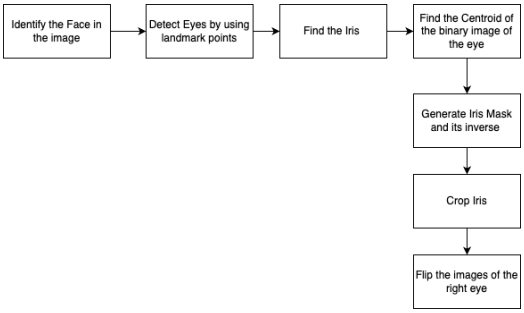

Fig. 3. The workflow expressed in a block diagram

3.3.2 Method 2 of Iris Extraction. The second method revolves around using 64 landmark points from the dlib [1] library. The dlib landmark points are 64 points which correspond to the eyes, nose, mouth, and face.

Firstly, the face is found, and then by using the landmark points the eyes are detected. After that, the eye mask is created. The iris is found by thresholding the red channel of the images within the boundaries of the eye mask. The centroid of the binary image of the eye is found. This centroid is then used to generate the iris mask and its inverse. Afterwards, the iris is cropped. The process can be visualized in diagram 3

The procedure is done for left 4 and right 5 eyes separately and the results are saved in the same manner as in the first method.

To better align the coordinates the right images were flipped. This helped since usually the highlights are closer to the nose, due to the light source being in the front of the eyes.

This method outputs 1050 GAN-generated images and 115 bonafide images.

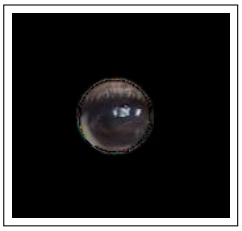

Fig. 4. Left Eye

Fig. 5. Right Eye

3.4 Brightest Pixel

By definition, highlights represent the brightest point in the image, since it is where the light directly reflects in a picture. This is why it is called a search for the brightest pixel. Finding the brightest pixel is done using the cv2.minMaxLoc function, which detects the brightest pixel in the image and returns its coordinates. First, the image needs to be turned to a grey image and then blurred with Gaussian Blur. Those 2 steps ensure that any noise is eliminated. After that, a circle is drawn around it to visualize where the pixel was found.

When the first method is employed to find the iris, in the GAN-generated images, the minMaxLoc function might give some outliers. For example, the highlight might not represent the brightest spot, but it is in the sclera. This is how the outlier looks 6 as opposed to standard data 7 Fortunately, those are only the outliers that happen in the GAN-generated images only.

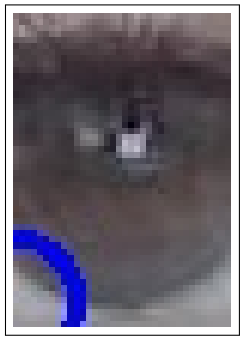

Fig. 6. Outlier

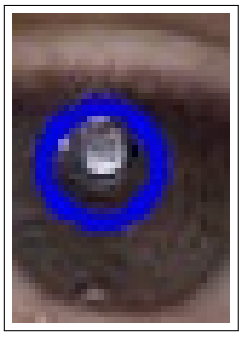

Fig. 7. Successful Extraction

3.5 Metrics

In order to evaluate the model the following metrics were used: accuracy, precision, recall, and F1 score.

Accuracy measures the overall correctness of the model and is calculated by dividing the correctly classified images be they real or not and dividing it by the total number of images.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

Precision measures the proportion of correctly classified GAN-generated images among the total images predicted as GAN-generated. This measure indicates whether the system is reliable in identifying GAN-generated images.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

Recall measures how many GAN-generated images were correctly classified among the total GAN-generated images in the dataset. This indicates how well the system captures GAN-generated images in the dataset.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The F1 score is a harmonic mean of both precision and recall. It indicates the overall performance of the model and takes into consideration both false positives and negatives.

By taking into consideration all the metrics described above, we believe that the system can be analyzed thoroughly and compared adequately to the other state-of-the-art systems.

4 EXPERIMENTS

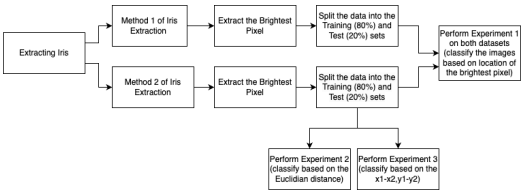

Fig. 8. The workflow expressed in a block diagram

In this research, three experiments were conducted to understand better the detection of GAN-generated images based on the specular corneal highlights. Experiment 1 focuses on training the model using only the coordinates of the brightest pixels, it is expected that those coordinates will match from the left and right eye. Experiment 2 involves calculating the Euclidean distance between the coordinates, to provide insights into the spatial relationships between the pairs. Experiment 3 involves calculating the distance between the respective x and y points of the left and right eye coordinates, capturing the alignment of the brightest points. In all the experiments a Logistic Regression model was used.

These experiments aim to enhance the interpretability of the model's predictions by considering different aspects of the coordinates of the highlights.

Which datasets are used for which experiments can be visualized in 8

4.1 Experiment 1

As previously stated the data obtained from the preprocessing step were the coordinates of the brightest spot of the image, the highlight. By taking the brightest pixel, it was expected that in the real images, the coordinates of those would align. Only the brightest pixel was taken to increase processing time and also since if the highlights align, due to the settings of pictures, so must their brightest points.

The first model was thus trained on specifically the coordinates, of the brightest spot from the left and right eye. An array of tuples was passed to the model, the first tuple represents the coordinates from the brightest pixel of the left eye, while the second tuple of the corresponding right eye. The training and test sets are flattened and passed into a Logistic Regression algorithm.

By using only the coordinates it was expected that the model would learn to differentiate real and generated images based on the alignment of the brightest pixel

4.2 Experiment 2

The second experiment was to take the Euclidean distance between the two sets of coordinates. The Euclidean distance between two points (x_1, y_1) and (x_2, y_2) in a two-dimensional space is given by the formula:

$$\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The Euclidean distance formula was utilized to measure the geometric distance between the two points in a two-dimensional space. Thus, giving insights into whether the system was looking at the spatial characteristics of the two points. The experiment was performed only on the set of data obtained from the second extraction method since for the understanding of the workings of the model it was enough to perform it only on one of the datasets.

4.3 Experiment 3

The third experiment was based on the distance between the (x,y) points from the left and the right eye. May (x1,y1) be the coordinates of the brightest spot from the left eye and (x2,y2) be the coordinates of the brightest spot from the right eye. Then we perform this operation for x $x_1 - x_2$ and for y $y_1 - y_2$ and create an array of tuples. To see whether the alignment of the highlights could be helpful in giving insights into whether the image was or was not GAN-generated. This would indicate the spatial relationship and alignment of the two bright spots. The experiment was performed only on the set of data obtained from the second extraction method since for the understanding of the workings of the model it was enough to perform it only on one of the datasets.

5 RESULTS

In order to best visualize the results, confusion matrices were generated. In the confusion matrix, the number of true positives, false positives, true negatives, and false negatives are visible, thus facilitating a better understanding of the further results and discussion.

For the first experiment, two confusion matrices were generated, 9 refers to the first method of extracting the iris, in which the iris is extracted using Hough Circles. 10 This confusion matrix refers to the second extracting technique, which utilizes masking.

Extraction Method	Accuracy	Precision	Recall	F1 score
One	0.66	0.75	0.5	0.6
Two	0.94	0.88	0.95	0.91

Table 1. The Metrics of the 2 experiments.

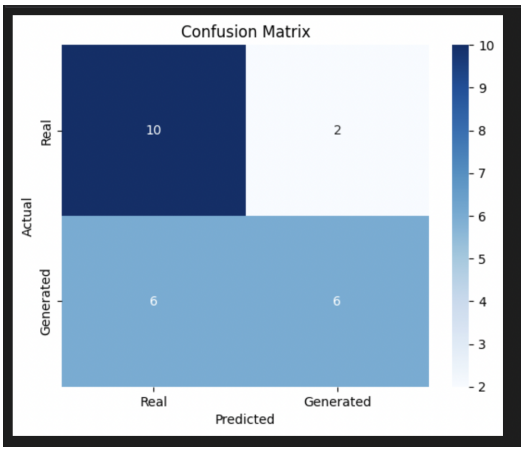

Fig. 9. Confusion Matrix of Experiment 1 when extracting the iris with Method 1

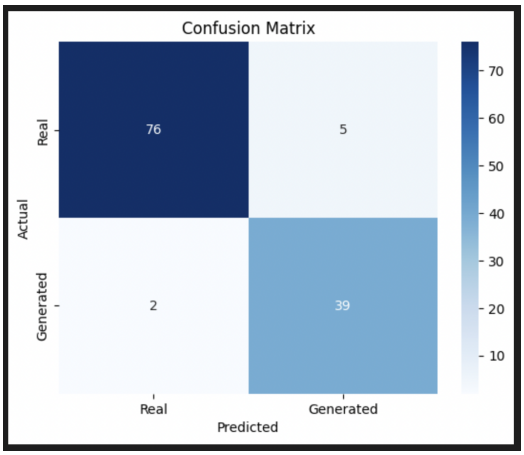

Fig. 10. Confusion Matrix of Experiment 2 when extracting the iris with Method 2

Table 2. Metrics of Experiment 2

Accuracy	Precision	Recall	F1 score
0.78	0.84	0.82	0.83

The above confusion matrices differentiate in the rates. Although the confusion matrix is a good guideline for what is happening, further metrics are needed to evaluate the results. Table 1 shows an overview of the metrics for experiment one and both extraction methods.

As for the second experiment the confusion matrix can be seen in 11 and the other metrics are visualized in table 2. Experiment 2 consisted of detecting based on the Euclidean distance.

The confusion matrix of the third experiment can be seen in 12 and the other metrics are visualized in table 3. Experiment three

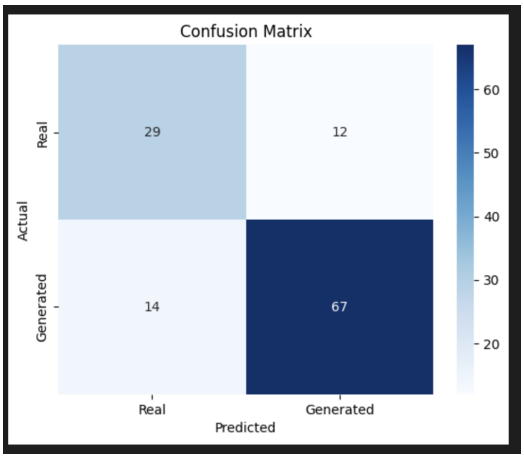

Fig. 11. Confusion Matrix of Experiment 2

Table 3. Metrics of Experiment 3

Accuracy	Precision	Recall	F1 score
0.95	0.97	0.95	0.96

was based on the detection based on the (x,y) coordinates difference between the 2 eyes.

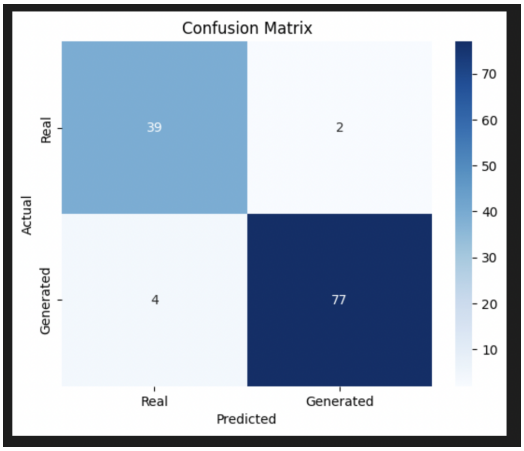

Fig. 12. Confusion Matrix of Experiment 3

6 DISCUSSION

6.1 Method 1 of extraction of iris VS Method 2 in experiment 1

From the above results, it is easy to notice the differences in the accuracy level of the two extraction methods. It is easy to assume that the differences lie in the extraction methods themselves and assume that the second one is simply more precise. However, that is not entirely true. By looking at the data it is easy to notice the

fact that the second method of extraction masks everything outside the iris, so everything around it is black, while the first one leaves some parts of the cornea, especially the white ones. Due to the nature of generated images, the minMaxLoc function detects the brightest pixel as being located in the white part of the cornea for some images. This is an exciting result.

Moreover, for the same image, the left brightest pixel could be in the white part of the cornea, while the right one is in the correct spot. This is an interesting discovery since it shows that there are some colour attributes of GAN-generated images that represent distinctive features for the GAN-generated images.

On the other hand, the fact that in the second method, the accuracy score lies at 94% indicates that detecting GAN-generated images using the specular corneal highlights is worth pursuing and might represent a reliable method of doing so.

Those results are consistent with the existing literature, in [7], the obtained accuracy was 94%, which is precisely the same accuracy as obtained in this research paper. In [7] the shape of the specular highlight was taken and compared between the two eyes. In the aforementioned research, the methodology is tested on GAN-synthesized faces, which means that the faces have been generated by a GAN algorithm from scratch.

6.2 Where are the results coming from?

As observed in the tables 1, 2, and 3. The results are mostly coming from the spatial position and alignment of the highlights rather than their distance from each other. Even though distance represents an important metric, the spatial position still represents 95% accuracy. This aligns with the assumption that the highlights need to be in the same relative spots in the bonafide images.

7 CONCLUSION

In this research, a way to detect morphed portrait images was proposed by analyzing the positions of the brightest pixel of the specular corneal highlights. Through a series of experiments, we aimed to enhance the interpretability and accuracy of the method. The face morphing technique involves blending two or more images together to generate a new one.

In Experiment 1 we looked solely at the location of the specular highlights. The expectations were that the bonafide images would have the same location of the highlights. By making this assumption we have achieved an accuracy of 94%, which are some promising results.

The aim of Experiment 2 was to gain further insights into the spatial characteristics of the highlights. In Experiment 2 the Euclidean distance was used to dwell deeper into the importance of the spatial characteristics of the highlights. As a result, the system achieved 76% accuracy.

The main focus of Experiment 3 was the specific distance between the x and y points of the left and right eye coordinates. Creating an array of tuples with the differences of x and y points of the left and right eye, allowed us to capture the spatial relationship between the brightest pixels from both eyes.

To answer the research question the method of detecting GAN-generated images using the specular corneal highlights, specifically their relative position to each other is quite accurate. The obtained 94% accuracy is a clear indicator of that. Moreover, according to the results, the spatial relationship represents a more important metric than Euclidean distance. This can be seen in the 95% achieved accuracy for this experiment.

This method of distinguishing between GAN-generated morphs and bonafide images has significant implications in the real world, especially in cyber security, where morphed image attacks happen.

In conclusion, our research demonstrates the potential of analysing the location of the brightest pixel in specular corneal highlights to determine the veridicality of a portrait image. The conducted experiments demonstrate the importance of considering both the location and spatial relationship of the highlights. These findings can be used to develop the detection methodologies further.

8 LIMITATIONS AND FUTURE WORK

While the method shows promising results, there are some limitations to consider.

The experiments were conducted on a limited dataset. The dataset was only made of specifically GAN-generated morphs and thus further experiments need to be conducted to establish the generalizability of the method. Thus, the usability is unclear for other types of morphs. In this way, future research could be centralized around generalizing the method and testing it against other types of morphs.

Moreover, in the detection process, only the coordinates of the brightest pixel were taken. Future research could focus on the context of the highlights, as well as their shape. During the extraction process, we noticed that in morphed images the brightest pixel is sometimes in the sclera instead of the highlight, which is never the case in the real images. So future research could focus on the colour attributes of the highlights.

ACKNOWLEDGMENTS

To Luuk Spreeuwers, for being the most amazing supervisor and mentor. I am very thankful I got to work with you on this research. You inspired me by showing what it means to be truly passionate about your work.

REFERENCES

[1] 2023. dlib C++ Library. http://dlib.net/
[2] 2023. OpenCV: Hough Circle Transform. https://docs.opencv.org/4.x/da/d53/tutorial_py_houghcircles.html
[3] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2019. ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection. arXiv:1812.02510 [cs.CV]

[4] Luca Debiasi, Naser Damer, Alexandra Moseguí Saladié, Christian Rathgeb, Ulrich Scherhag, Christoph Busch, Florian Kirchbuchner, and Andreas Uhl. 2019. On the Detection of GAN-Based Face Morphs Using Established Morph Detectors. In *Image Analysis and Processing – ICIAP 2019*, Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe (Eds.). Springer International Publishing, Cham, 345–356.

[5] Lisa DeBruine and Benedict Jones. 2017. Face Research Lab London Set. https://doi.org/10.6084/m9.figshare.5047666.v3

[6] Diego Gragnaniello, Francesco Marra, and Luisa Verdoliva. 2022. *Detection of AI-Generated Synthetic Faces.* Springer International Publishing, Cham, 191–212. https://doi.org/10.1007/978-3-030-87664-7_9

[7] Shu Hu, Yuezun Li, and Siwei Lyu. 2021. Exposing GAN-Generated Faces Using Inconsistent Corneal Specular Highlights. 2500–2504. https://doi.org/10.1109/ICASSP39728.2021.9414582

[8] Micah Johnson and Hany Farid. 2007. Exposing Digital Forgeries Through Specular Highlights on the Eye. *Proceeding of the 9th International Workshop on Information Hiding*, 311–325. https://doi.org/10.1007/978-3-540-77370-2_21

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813

[10] Ali Khodabakhsh, Raghavendra Ramachandra, Kiran Raja, Pankaj Wasnik, and Christoph Busch. 2018. Fake Face Detection Methods: Can They Be Generalized? https://doi.org/10.23919/BIOSIG.2018.8553251

[11] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 83–92. https://doi.org/10.1109/WACVW.2019.00020

[12] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. 2020. Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs? - Vulnerability and Detection. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. 1–6. https://doi.org/10.1109/IWBF49977.2020.9107970