EYES TELL ALL: IRREGULAR PUPIL SHAPES REVEAL GAN-GENERATED FACES

Hui Guo[1], *Shu Hu*[2], *Xin Wang*[3], *Ming-Ching Chang*[1], *Siwei Lyu*[2]

[1]University at Albany, State University of New York, USA. {hguo,mchang2}@albany.edu
[2]University at Buffalo, State University of New York, USA. {shuhu,siweilyu}@buffalo.edu
[3]Keya Medical, Seattle, Washington, USA. xinw@keyamedna.com

ABSTRACT

Generative adversary network (GAN) generated high-realistic human faces are visually challenging to discern from real ones. They have been used as profile images for fake social media accounts, which leads to high negative social impacts. In this work, we show that GAN-generated faces can be exposed via irregular pupil shapes. This phenomenon is caused by the lack of physiological constraints in the GAN models. We demonstrate that such artifacts exist widely in high-quality GAN-generated faces. We design an automatic method to segment the pupils from the eyes and analyze their shapes to distinguish GAN-generated faces from real ones. Qualitative and quantitative evaluations of our method on the Flickr-Faces-HQ dataset and a StyleGAN2 generated face dataset demonstrate the effectiveness and simplicity of our method.

Index Terms— image forensics, GAN faces detection, pupil segmentation, fake face detection

1. INTRODUCTION

The rapid development of the generative adversarial networks (GAN) models [1, 2, 3] has made it possible to synthesize highly realistic human face images that are difficult to discern from real ones [4]. These GAN-generated faces have been misused for malicious purposes. Recent years have seen an increasing number of reports that GAN-generated faces were used as profile images on fake social media accounts, which generates negative social impacts [5, 6, 7, 8].

Such pernicious impact of these fake faces have lead to the development of methods aiming to distinguish GAN-generated images from real ones. Many of those methods are based on deep neural network (DNN) models due to their high detection accuracy [9, 10, 11]. Albeit such success, these methods suffer from two significant limitations: (1) the lack of interpretability of the detection results and (2) the low capability to generalize across different synthesis methods [12, 13].

Another category of GAN-generated image detection methods aims to expose the inadequacy of the GAN models in handling the physical constraints of the face representation and synthesis process [14, 15, 16, 17]. Since these methods exploit knowledge of the physical world in distinguishing the fake images from real ones, these methods are more interpretable and can work robustly against various synthesis methods. The recent work [18] exploits the inconsistency of the corneal specular highlights between the two synthesized eyes to identify GAN-generated faces. However, this method is limited by its environmental assumptions regarding the light sources and the reflectors that must be visible in both eyes. This might cause high false negatives in fake face detection.

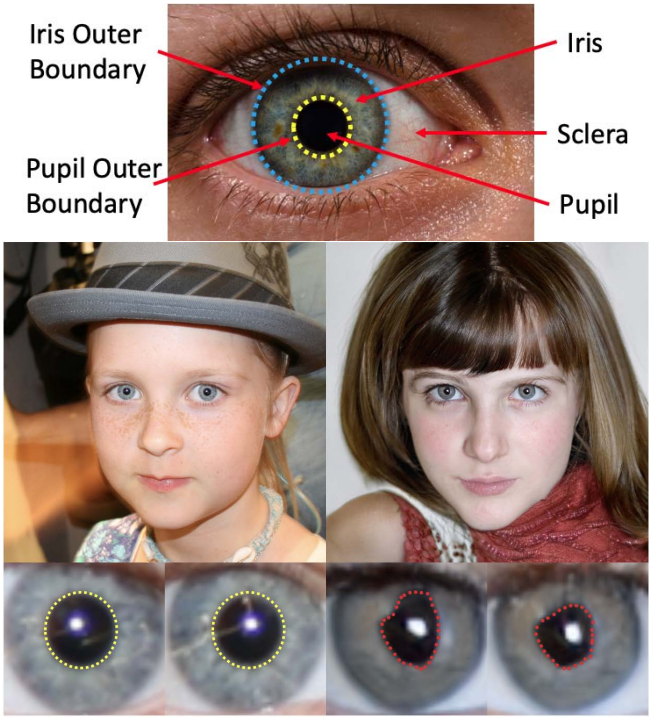

Fig. 1: ***Top:*** *Anatomy of a human eye: the iris and pupil are at the center surrounded by the sclera.* ***Bottom:*** *Examples of pupils of real human (left) and GAN-generated (right). Note that the pupils for the real eyes are in circular or elliptical shapes (**yellow**), while those for the GAN-generated pupils are much irregular (**red**). For GAN-generated faces, the shapes of the pupils are very different from each other when zoomed-in.*

In this work, we explore a universal physiological cue of the eye, namely the *pupil shape consistency*, to reliably identify GAN-synthesized faces. As shown in Figure 1, the eye is one of the few organs in the human body that is highly circular and

regular in geometry. We hypothesize that the human iris and pupil can provide rich physical and physiological cues that can improve GAN-synthesized face detection.

Our method is based on a simple physiological assumption that human pupils should be nearly circular in their shapes in a face image. Due to different facial orientations and camera angles, the actual pupil shapes can be elliptical. Our observation is that *this simple property is not well preserved in the existing GAN models*, including StyleGAN2 [3], the state-of-the-art face synthesis method. The pupils for the StyleGAN-generated faces tend to have non-elliptical shapes with irregular boundaries. Figure 1 shows an example with zoom-in views of the pupils. Such artifacts in the GAN-generated faces are due to the difficulty or negligence of physiological constraints on human anatomy when training the GAN models via standard data-driven machine learning.

The proposed GAN-generated face detector consists of several automatic steps. We first segment the pupil regions of the eyes and extract their boundaries automatically. We next fit an ellipse parametric model to each pupil, and calculate the Boundary Intersection-over-Union (BIoU) scores [19] between the predicted pupil mask and the ellipse-fitted model. The BIoU score provides a quantitative measure of the regularity of the pupil shape, that determines if the eyes (and the face) are real or not. Experiments are conducted on a dataset containing both real and machine-synthesized faces. Results in § 4 show that there is a clear separation between the distributions of the BIoU scores of the real and GAN-generated faces.

The main contributions of this work are two-fold:

- We are the first to propose the idea of exploiting pupil shape consistency as an effective way to distinguish fake faces from real ones. This new cue is effective for humans as well to visually identify GAN-generated faces.
- The proposed method for fake face detection is based on explainable physiological cues. It is simple, effective, and explainable. Evaluations on the Flickr-Faces-HQ dataset and an in-house collected StyleGAN2 face dataset show its effectiveness and computational efficiency.

2. RELATED WORKS

GAN-generated faces. A series of recent GAN models have demonstrated superior capacity in generating or synthesizing realistic human faces. However, the works [14, 17] indicate that faces generated by the early StyleGAN model [2] have considerable artifacts such as fingerprints [9, 20], inconsistent iris colors [16, 21], *etc.* More recently, StyleGAN2 [3] has greatly improved the visual quality and pixel resolution, with largely-reduced or undetectable artifacts in the generated faces.

GAN-generated face detection. With the development of the GAN models for face generation/synthesis, methods for distinguishing GAN-generated faces have progressed accordingly as well. Most of these methods are Deep Learning based [22, 23, 10, 24, 25]. Notably, several methods exploit the physiological cues (which suggest inconsistency in the physical world) to distinguish GAN-generated faces from the real ones [17]. In [14], GAN-generated faces are identified by analyzing the distributions of the facial landmarks. The work of [18] analyzes the light source directions from the perspective distortion of the locations of the specular highlights of the two eyes. Such physiological/physical-based methods come with intuitive interpretations and are more robust to adversarial attacks [26, 27].

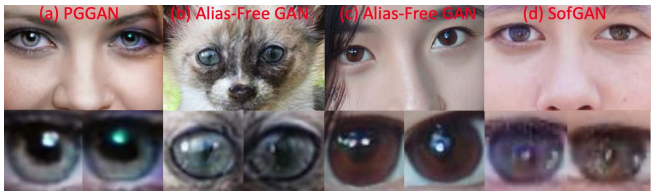

Fig. 2: *Examples of GAN-synthesized faces additional to StyleGAN and StyleGAN2. The images are from their original papers* ***(a)*** *PGGAN [1],* ***(b,c)*** *Alias-Free GAN (StyleGAN3) [31],* ***(d)*** *SofGAN [32]. Observe in the zoomed-in view that the pupils appear in irregular, inconsistent shapes, which tell them apart from real faces.*

Iris and pupil segmentation is an important task in biometric identification that has been studied well. The IrisParseNet [28] provides complete iris segmentation solutions including iris mask and inner and outer iris boundaries extraction, which are jointly modeled in a unified multi-task neural network. Iris segmentation in non-cooperative environments is supported, while the iris pixel quality might be low due to the limited user cooperation (moving camera, poor illumination, or long-distance views). An end-to-end trainable lightweight stacked hourglass network is presented in [29] for iris segmentation from noisy images acquired by mobile devices. More recent methods can be found in the NIR Iris Challenge survey paper [30].

3. METHOD

Our fake face detection method is motivated by the observation that GAN-generated faces exhibit a common artifact that the pupils appear with irregular shapes or boundaries, other than a smooth circle or ellipse. This artifact is universal for all known GAN models (at least for now, *e.g.* PGGAN [1], Alias-Free GAN [31], and SofGAN [32]), as shown in Figure 2. This artifact occurs in both the synthesized human and animal eyes.

Our automatic fake face detection pipeline starts with a face detector to identify any face in the input image. We then extract the facial landmark points to localize the eyes and then perform pupil segmentation. The segmented pupil boundary curves are next analyzed to determine if the pupil shape is irregular. We perform parametric fitting of the pupil to an ellipse following the mean squared error (MSE) optimization. This provides a way to define a distance metric to quantify the irregularity for decision-making. The following subsections describe each step of our pipeline in details.

3.1. Pupil Segmentation and Boundary Detection

We adopt the Dlib [33] face detection to locate the face and extract the 68 facial landmark points provided in Dlib, as shown in Figure 3. We next focus on the eye regions to perform pupil segmentation. We use EyeCool [30] [1] to extract the pupil segmentation masks with corresponding boundary contours. EyeCool provides an improved U-Net-based model with EfficientNet-B5 [34] as the encoder. A boundary attention block is added in the decoder to improve the ability of the model to focus on the object boundaries. Specifically, considering the subpixel accuracy, we focus on the outer boundary of the pupil for the irregularity analysis.

3.2. Ellipse Fitting to the Pupil Boundary

We next fit an ellipse to the pupil mask via least-square fitting. As there might be multiple components in the predicted masks, we keep the largest component for ellipse fitting. Specifically, the method of [35] is used to fit an ellipse to the outer boundary of the extracted pupil mask. Figure 3(d) shows an example. Denotes u as the coordinates of the outer boundary points from the pupil mask. The least-square fitting determines the ellipse parameters θ minimizing the distance between the pupil boundary points and a parametric ellipse represented by:

$$F(\mathbf{u};\theta) = \theta \cdot \mathbf{u} = ax^2 + bxy + cy^2 + dx + ey + f = 0,$$

where $\theta = [a, b, c, d, e, f]^T$ and $\mathbf{u} = [x^2, xy, y^2, x, y, 1]^T$; T denotes transpose. $F(\mathbf{u};\theta)$ represents the algebraic distance of a 2D point (x, y) to the ellipse, and a perfect fit is indicated by $F(\mathbf{u};\theta) = 0$. The fitting solution is obtained by minimizing the sum of squared distances (SSD) over the N data points from the pupil boundary:

$$\min_{\theta} \mathcal{D}(\theta) := \sum_{i=1}^{N} F(u_i;\theta_i)^2, \quad \text{s.t. } ||\theta||^2 = 1, \;\; b^2 \geq ac,$$

where the constraints are imposed to avoid the trivial solution of $\theta = \mathbf{0}$ and ensure the *positive definiteness* of the quadratic form. The solution is calculated using the gradient-based optimization described in [35].

3.3. Estimating the Pupil Shape Irregularity

To accurately estimate the irregularity of the segmented pupil boundary and the fitted ellipse, we adopt the Boundary IoU (BIoU) [19] as a distance metric. BIoU is widely used in image segmentation where the sensitivity of the object boundary is important. Instead of considering all pixels, BIoU calculates the IoU for mask pixels within a certain distance from the boundary contours between the predicted mask and the corresponding ground truth mask. Thus, BIoU can better focus on the matching of the boundaries of the two shapes. We use

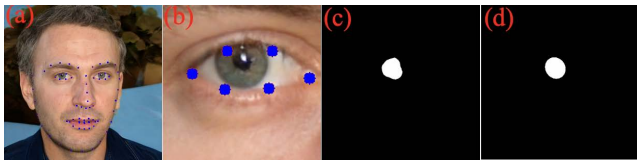

Fig. 3: *The proposed pipeline for face detection, facial landmark localization, pupil segmentation, and pupil ellipse fitting.* ***(a)*** *The input high-resolution face image,* ***(b)*** *The cropped eye image using landmarks,* ***(c)*** *The predicted pupil mask of the eye from (b),* ***(d)*** *The fitted ellipse mask. This example shows a GAN-generated face.*

BIoU to evaluate the pupil mask pixels that are within a distance of d pixels from the pupil boundary. For each extracted pupil mask, we use P to indicate the predicted pupil mask and F for the fitted ellipse mask. Denote P_d and F_d the mask pixels within distance d from the predicted and fitted boundaries, respectively. BIoU is calculated as:

$$\text{BIoU}\ (F, P) = \frac{|(F_d \cap F) \cap (P_d \cap P)|}{|(F_d \cap F) \cup (P_d \cap P)|}. \qquad (1)$$

The distance parameter d controls the sensitivity of the BIoU measure to the object boundary. Reducing the value of d causes the fitting to be more sensitive to the boundary pixels while ignoring the interior pixels of the pupil mask. We set $d = 4$ for the BIoU calculation, which leads to the best empirical segmentation performance in our experiments.

Given the predicted pupil mask and the ellipse fitted pupil mask, the BIoU score takes range in $[0, 1]$. A larger BIoU value suggests the pupil boundary better fits the parametrized ellipse. In our case, higher BIoU values suggest more regular pupil shapes and thus the face is more likely real. In comparison, GAN-generated faces should produce lower pupil BIoU scores.

4. EXPERIMENTS

Datasets. We use the real human faces from the Flickr-Faces-HQ (FFHQ) dataset [2]. Since StyleGAN2 [3] [2] is currently the state-of-the-art GAN face generation model with the best synthesis quality, we collect GAN-generated faces from it. We only use images where the eyes and pupils can be successfully extracted. In total, we collected $1,600$ images for each class (of real *vs.* fake faces) with a resolution of $1,024 \times 1,024$.

Results. Figure 4 shows examples of the segmented pupils for both the real and GAN-generated faces. These results clearly show that pupils in the real faces are in strongly regular, elliptical shapes. Such high pupil shape regularity is also reflected in the high BIoU scores computed for the pupil mask and the fitted ellipse. On the other hand, irregular pupil shapes lead to significantly lower BIoU scores, which represents the artifacts of GAN-generated faces.

[1] Code at https://github.com/neu-eyecool/NIR-ISL2021.

[2] http://thispersondoesnotexist.com

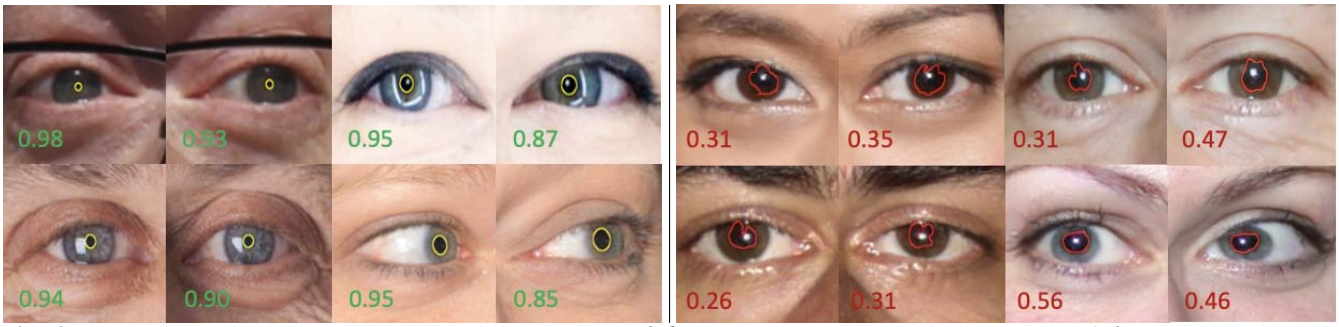

Fig. 4: *Examples of both eyes from the real human faces (**left**) and GAN generated human faces (**right**). The pixels of the predicted pupil mask within distance $d = 4$ from the prediction boundary contours are highlighted. The BIoU scores with $d = 4$ between the predicted pupil mask and the ellipse-fitted one are shown on each image.*

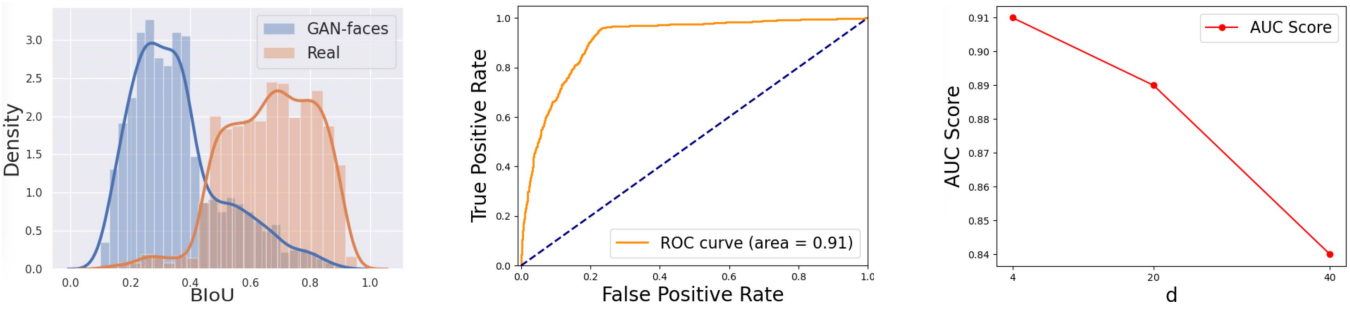

Fig. 5: ***Left:** Distributions of the BIoU scores (of the averages of both pupils) for the real and GAN-generated faces. **Middle:** The ROC curve based on the BIoU with $d = 4$. **Right:** BIoU hyper-parameter analysis, where the x-axis indicates distance parameter d, and y-axis indicates the AUC score.*

Figure 5(left) shows the distributions of the BIoU scores of pupils from the real faces and GAN-generated ones. Observe that there is a clear separation between the two classes of distributions, indicating that the proposed *pupil shape regularity* can indeed serve as an effective feature to distinguish the GAN-generated faces from the real ones.

Figure 5(middle) shows the *receiver operating characteristic* (ROC) curve of our GAN-generated face detection evaluation. The Area under the ROC curve (AUC) score is 0.91, which indicates the effectiveness of the proposed method.

Sensitivity Analysis of d. The BIoU boundary distance d is an essential parameter that controls the matching sensitivity of the pixels near shape boundary. Figure 5(right) shows how the fake face detection ROC varies *w.r.t.* parameter d. As the value of d grows too large, sensitivity telling the differences of pupil boundary decreases, which also reduces fake face detection performance.

Limitations. The proposed method still contains several limitations. Since our method is based on the simple assumption of pupil shape regularity, false positives may occur when the pupil shapes are non-elliptical in the real faces. This may happen for infected eyes of certain diseases as shown in Figure 6(left). Also imperfect imaging conditions including lighting variations, largely skew views, and occlusions can also cause errors in pupil segmentation or thresholding errors, as shown in Figure 6(right).

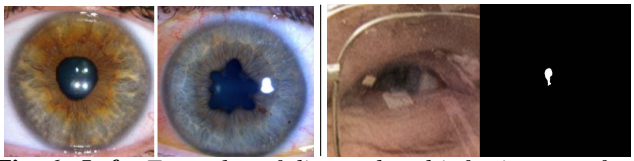

Fig. 6: ***Left:** Examples of diseased and infection eyes from [36]. These pupils from images of real faces contain abnormal non-elliptical pupil shapes, which only occurs rarely in real life. **Right:** Occlusions and environmental variations can cause pupil segmentation failure.*

5. CONCLUSION

In this paper, we show that GAN-generated faces can be identified by exploiting the regularity of the pupil shapes. We propose an automatic method for pupil localization and segmentation, and perform ellipse fitting to the segmented pupils to estimate a Boundary IoU score for forensic classification. The proposed approach is simple yet effective. The detection results are interpretable based on the BIoU score.

Future Work. We will investigate other types of inconsistencies between two pupils of the GAN-generated face, such as the different geometric shapes and relative locations of pupils in the two eyes. These cues in combination may further improve forensic detection effectiveness. Future work also includes the deployment to an online platform that can further expand the impact in addressing issues in social media forensics.

6. REFERENCES

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *ICLR*, 2018.

[2] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of StyleGAN," in *CVPR*, 2020.

[4] Sophie Nightingale, Shruti Agarwal, Erik Härkönen, Jaakko Lehtinen, and Hany Farid, "Synthetic faces: how perceptually convincing are they?," *Journal of Vision*, 2021.

[5] "A spy reportedly used an AI-generated profile picture to connect with sources on LinkedIn," https://bit.ly/35BU215.

[6] "A high school student created a fake 2020 US candidate. Twitter verified it," https://www.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html.

[7] "How fake faces are being weaponized online," https://www.cnn.com/2020/02/20/tech/fake-faces-deepfake/index.html.

[8] "These faces are not real," https://graphics.reuters.com/CYBER-DEEPFAKE/ACTIVIST/nmovajgnxpa/index.html.

[9] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi, "Do GANs leave artificial fingerprints?," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.

[10] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, "CNN-generated images are surprisingly easy to spot... for now," in *CVPR*, 2020, pp. 8695–8704.

[11] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, "Robust attentive deep neural network for exposing gan-generated faces," *arXiv preprint arXiv:2109.02167*, 2021.

[12] Shu Hu, Yiming Ying, Xin Wang, and Siwei Lyu, "Learning by minimizing the sum of ranked range," *NeurIPS*, vol. 33, 2020.

[13] Shu Hu, Yiming Ying, Xin Wang, and Siwei Lyu, "Sum of ranked range loss for supervised learning," *arXiv preprint arXiv:2106.03300*, 2021.

[14] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu, "Exposing GAN-synthesized faces using landmark locations," in *ACM Workshop on Information Hiding and Multimedia Security (IHMMSec)*, 2019.

[15] Xin Yang, Yuezun Li, and Siwei Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP*, 2019.

[16] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, "Detection of deep network generated images using disparities in color components," *arXiv preprint arXiv:1808.07276*, 2018.

[17] Falko Matern, Christian Riess, and Marc Stamminger, "Exploiting visual artifacts to expose Deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 2019, pp. 83–92.

[18] Shu Hu, Yuezun Li, and Siwei Lyu, "Exposing GAN-generated faces using inconsistent corneal specular highlights," in *ICASSP*, 2021.

[19] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *CVPR*, 2021.

[20] Ning Yu, Larry S Davis, and Mario Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *ICCV*, 2019.

[21] Scott McCloskey and Michael Albright, "Detecting GAN-generated imagery using color cues," *arXiv preprint arXiv:1812.08247*, 2018.

[22] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva, "Incremental learning for the detection and classification of GAN-generated images," in *IEEE WIFS*, 2019.

[23] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring, "Detecting CNN-generated facial images in real-world scenarios," in *CVPR Workshops*, 2020, pp. 642–643.

[24] Michael Goebel, Lakshmanan Nataraj, and etc, "Detection, attribution and localization of GAN generated images," *arXiv:2007.10466*, 2020.

[25] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr, "Global texture enhancement for fake face detection in the wild," in *CVPR*, 2020, pp. 8060–8069.

[26] Luisa Verdoliva, "Media forensics and DeepFakes: an overview," *arXiv:2001.06564*, 2020.

[27] Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *ICCV*, 2021, pp. 7649–7657.

[28] Caiyong Wang, Jawad Muhammad, Yunlong Wang, Zhaofeng He, and Zhenan Sun, "Towards complete and accurate iris segmentation using deep multi-task attention network for non-cooperative iris recognition," *IEEE TIFS*, 2020.

[29] Caiyong Wang, Yunlong Wang, Boqiang Xu, and etc, "A lightweight multi-label segmentation network for mobile iris biometrics," in *ICASSP*, 2020.

[30] Caiyong Wang, Yunlong Wang, Kunbo Zhang, and etc., "NIR iris challenge evaluation in non-cooperative environments: Segmentation and localization," in *IEEE IJCB*, 2021.

[31] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Alias-free generative adversarial networks," *arXiv preprint arXiv:2106.12423*, 2021.

[32] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu, "SofGAN: A portrait image generator with dynamic styling," *ACM transactions on graphics*, 2021.

[33] Davis E King, "Dlib-ml: A machine learning toolkit," *JMLR*, vol. 10, pp. 1755–1758, 2009.

[34] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[35] Andrew Fitzgibbon, Maurizio Pilu, and Robert B Fisher, "Direct least square fitting of ellipses," *IEEE TPAMI*, 1999.

[36] A. R. Ramli R. A. Ramlee and Z. M. Noh, "Pupil segmentation of abnormal eye using image enhancement in spatial domain," in *International Technical Postgraduate Conference*, 2017.