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, SUNY, USA. {hguo,mchang2}@albany.edu
[2]University at Buffalo, SUNY, USA. {shuhu,siweilyu}@buffalo.edu
[3]Keya Medical, Settle, USA. xinw@keyamedna.com

ABSTRACT

Generative adversary network (GAN) generated high-realistic human faces have been used as profile images for fake social media accounts and are visually challenging to discern from real ones. 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 and further describe an automatic method to extract the pupils from two eyes and analysis their shapes for exposing the GAN-generated faces. Qualitative and quantitative evaluations of our method suggest its simplicity and effectiveness in distinguishing GAN-generated faces.

Index Terms— Media Forensics, GAN Faces Detection, Pupil Detection

1. INTRODUCTION

The development of generative adversarial networks (GANs) [1] enables generating high-realistic human faces images, and the generated faces are difficult to discern from real ones visually [2, 3, 4]. Thus, fake social media accounts using the GAN-generated faces as profile images are more deceptive [5, 6, 7, 8]. Such behavior can be easily abused for malicious purposes, which can cause a significant social disturbance.

A recent development on the GAN-faces detection approaches using the Deep Learning model has shown the promising feasibility [9, 10]. However, these methods typically suffer from two significant challenges: the lack of interpretability of the detection results and low robustness of generalization across different synthesis methods due to the over-fitting problem. On the other hand, physical-based methods are proposed to overcome the above limitations by exposing the inadequacy of the GAN synthesis models in representing the human faces interaction with the physical world [11, 9, 12, 13]. The work [14] proposes to use the inconsistency of the corneal specular highlights between the two synthesized eyes to distinguish the real and the GAN-generated face images. However, the proposed method under strict portrait settings such as the light sources or reflectors in the environment are visible to both eyes, and the eyes are distant from the light or reflection source. However, when the portrait setting is not obeyed, the method will raise many false positives.

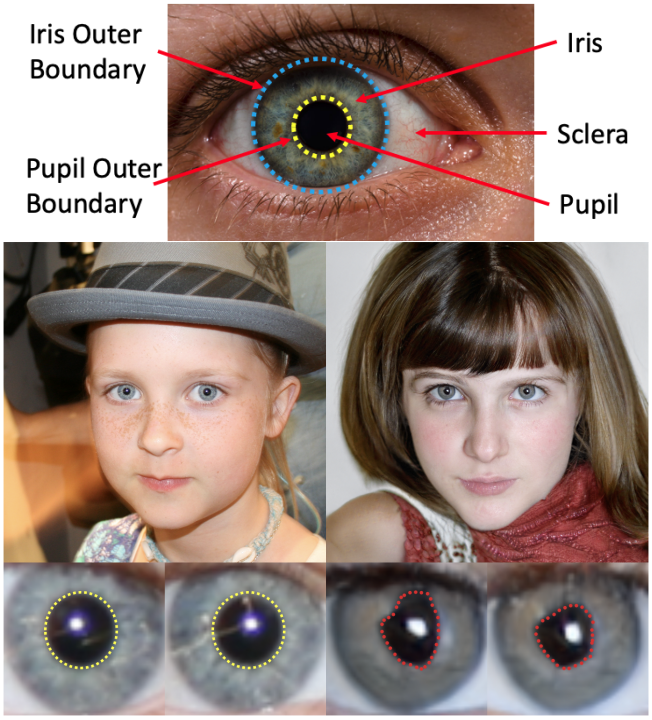

Fig. 1: ***Top:*** *Anatomy structures of a human eye.* ***Bottom:*** *Examples of pupils of real human (left) and GAN-generated (right). Note that the pupils for the real eyes have a strong circular or elliptical shapes (**yellow**) while those for the GAN-generated pupils are with irregular shapes (**red**). And also the shapes of both pupils are very different from each other in the GAN-generated face image.*

To eliminate these limitations and explore a more robust model, in this work, we propose a new physiological-based method based on pupil shapes. Concretely, we start with the main anatomic parts of a human eye (see Figure 1(top)). The center of an eye is the iris and pupil, and the white area is the sclera. The pupils have near-circular shapes for healthy adults. Comparing with the real faces, we observe that visible artifacts and inconsistencies can be observed in the eye regions of the

GAN-generated faces (e.g., StyleGAN2 [4]). In particular, the boundary of the GAN-generated pupils is not elliptical shapes. Note that the real pupil in an image should be elliptical due to the different face orientations. The bottom row of Figure 1 gives an illustrative example that compares pupils from the real faces and the GAN-generated faces. One fundamental reason for such artifacts in the GAN-generated faces is that the current GAN models lack understanding of human eye anatomy, especially the geometrical shapes of the pupils. Motivate by this observation, we propose a new detection method of GAN-generated faces that can automatically segment pupils from two eyes and extract their boundaries. Then calculate the Boundary intersection-over-union (BIoU) scores [15] between the predicted pupil mask and the ellipse-fitted pupil mask to evaluate and detect if they are with elliptical shapes. Our experiments show a clear separation between the distribution of the BIoU scores of the real and GAN-generated faces, which can be used as a quantitative measurement to differentiate them. In summary, the main contributions of this work are three-fold:

- We found irregular pupil shapes widely exist in the high-quality StyleGAN-generated faces, which are different from the real human pupils.
- We propose a new physiological-based method that can use the irregular pupil shapes as a cue to detect the GAN-generated faces, which is simple yet effective.
- Our findings can not only be used for designing the automatic detection methods but also is a good cue for the human to distinct the GAN-generated face visually (See Figure 7).

2. RELATED WORKS

We briefly introduce the works that inspired our method and other GAN-faces detection methods in the literature.

GAN-generated Faces Detection. A series of recent GAN models have demonstrated superior capacity in generating high-resolution realistic human faces. However, the works [11, 13] indicate that the faces generated by using the early StyleGAN model [2] have obvious artifacts such as asymmetric faces, inconsistent iris colors, etc [10, 16, 17, 12]. More recently, the StyleGAN2 model further improves the generated face quality [2, 3, 4].

With the development of the GAN models, move advanced GAN-generated faces detection methods using Deep Learning models have been developed correspondingly [18, 19, 20, 21, 22]. Furthermore, many explainable models that utilize physical/physiological inconsistencies of GAN models for GAN-generated faces detection methods have been proposed recently [13]. The work in [11] distinguish GAN-generated faces by analyzing the distributions of facial landmarks. More works analyze the internal camera parameters and light source directions from the perspective distortion of the locations of the specular highlights of two eyes and use them to reveal digital images composed from real human faces photographed under different illumination [23, 24]. Such physiological/physical-based detection methods are more robust to adversarial attacks and afford intuitive interpretations [14].

Iris and Pupil Segmentation. Iris and pupil segmentation are essential tasks in biometric identification and have been widely studied. Recently, IrisParseNet [25] was proposed for iris segmentation in the non-cooperative environments, where the captured iris images suffer from various noises due to limited user cooperation such as using a moving camera, poor illumination or long distance, etc. It is a complete iris segmentation solution including iris mask and inner and outer iris boundaries estimation, jointly modeled in a unified multi-task network. Experimental results demonstrate the proposed model is robust to various types of noises. Furthermore, a lightweight stacked hourglass network is proposed for iris segmentation of noisy images acquired by mobile devices [26]. As the method is end-to-end trainable, it can be used in any regular iris recognition system. More recent methods can be found in the survey paper from the public challenge [27].

3. METHOD

To detect such artifacts of the pupil shapes and use them as the basis of a method to expose GAN-generated faces, we first automatically extract the pupil masks of the two eyes and then evaluate if they have elliptical shapes.

3.1. Pupil Segmentation and Boundary Detection

Figure 2 illustrates the significant steps of our pupil segmentation and boundary detection from an input image. We first run a face detector to locate the face, followed by a landmark extractor to obtain the face landmarks. The regions corresponding to the two eyes are properly cropped out using the landmarks (Figure 2(b)). We then extract the pupil mask and boundary using EyeCool [27].

The EyeCool[1] segments the mask, the inner and outer boundary mask of both pupil and iris simultaneously, Figure 2(c) is the predicted pupil mask, which is the central part used in our method. Specifically, EyeCool is an improved U-Net-based model, where the EfficientNet-B5 [28] is used as an encoder, and a boundary attention block is added in the decoder to improve the ability of the model to focus on the object boundaries. Moreover, both Dice loss and MSE loss are used to train the model, where the Dice loss is used to evaluate the segmentation part, and the MSE is used to calculate the regression loss of the pupils' boundary heatmaps.

3.2. Ellipse Fitted Pupil Mask

To obtain the ellipse fitted pupil mask (Figure 2(d)), the Least Square-based ellipse fitting method [29] can be used on the outer boundary of the predicted pupil mask to estimate the ellipse fitted pupil boundaries.

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

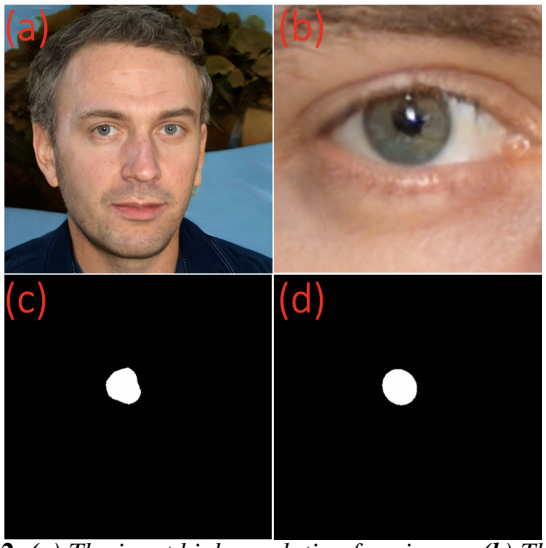

Fig. 2: ***(a)*** *The input high-resolution face image,* ***(b)*** *The the cropped eye image using landmarks,* ***(c)*** *Predicted pupil mask of image (b),* ***(d)*** *Ellipse fitted pupil mask of (c). Note that this example is a GAN-generated face.*

Formally, denotes u as the coordinates of the points on the outer boundary of the predicted pupil mask. The Least-squares based method aims to find the set of parameters θ that minimize a distance measure between the data points and the ellipse. The ellipses can be represented by a function,

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

where $\theta = [a, b, c, d, e, f]^T$ and $u = [x^2, xy, y^2, x, y, 1]^T$. $F(u;\theta)$ is the algebraic distance of a point (x, y) to the ellipse $F(u;\theta) = 0$. The fitting of an ellipse by minimizing the sum of squared algebraic distances over the N data points is formulated as follow,

$$\mathcal{D}(\theta) = \sum_{i=1}^{N} F(u_i;\theta_i)^2, \quad \text{subject to } ||\theta||^2 = 1 \qquad (1)$$

where the constraint is to avoid the trivial solution $\theta = 0$. The problem can be optimized by the gradient-based method inventively.

3.3. Measure the Irregular Pupil Shapes

To evaluate pupil shapes, a naive approach would be to measure the mask between the predicted pupil mask and the ellipse-fitted pupil mask by using Mask intersection-over-union (IoU). However, the Mask IoU divides the intersection area of two masks by the area of their union. This measure values all pixels equally. Thus, it is less sensitive to boundary quality.

Recently, Boundary IoU (BIoU) was proposed in [15] that aims to identify a measure for image segmentation that is sensitive to boundary quality. Instead of considering all pixels, it calculates the IoU for mask pixels within a certain distance

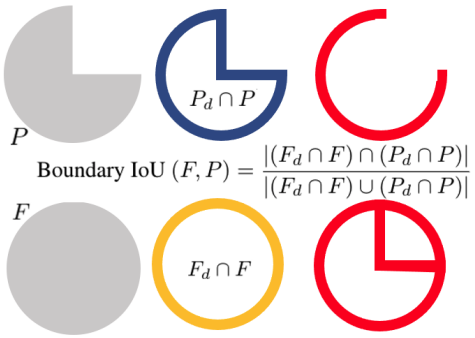

Fig. 3: *A toy example to explain the Boundary IoU.* ***Left:*** *The predicted pupil mask P and the ellipse fitted pupil mask F.* ***Middle:*** *P_d and F_d are the mask pixels within distance d from the boundaries (blue and yellow).* ***Right:*** *Boundary IoU calculation between predicted pupil mask and the ellipse fitted pupil mask with distance parameter d.*

from boundary contours between the predicted mask and the corresponding ground truth mask.

Thus, to focus more on the shape of the boundary of the pupil, we use the BIoU to evaluate only mask pixels of pupil that are within pixel distance d from the pupil outer boundary (See Figure 3), where P indicates the predicted pupil mask and F indicates the ellipse fitted pupil mask, the parameter d is the distance to the boundary that controls the measure's sensitivity to the boundary.

Moreover, the BIoU equals the Mask IoU when enlarging d large enough to include all pixels inside the masks. To make the BIoU more sensitive to the boundary quality, one can reduce the parameter d to ignore interior mask pixels.

The BIoU score between the predicted pupil mask and the ellipse fitted pupil mask takes range in $[0, 1]$ with a larger value suggesting the boundary the pupil is more similar with an elliptical shape. Hence, more likely the face is real; otherwise, it is generated with a GAN model.

4. EXPERIMENTS

Dataset. The images of real human faces are from the Flickr-Faces-HQ (FFHQ) dataset [3], and the GAN-generated human faces are created by the StyleGAN2[2] [4]. There are 1000 images for each class with resolution of 1024×1024.

Implementation details. We use the face detector and landmark extractor provided in DLib [30] to crop the eye region. The EyeCool predicted mask might include multiple components. The largest consent is selected as the final predicted mask for ellipse fitting and BIoU calculation. We set d=4 in our experiments because it is close to an optimal hyper-parameter.

Results. We first show examples of the analysis results for the pupil of both real and GAN-generated human faces in Figure 4. As described in the previous section, real human pupils have strong elliptical shapes, which are reflected by the higher

[2] http://thispersondoesnotexist.com

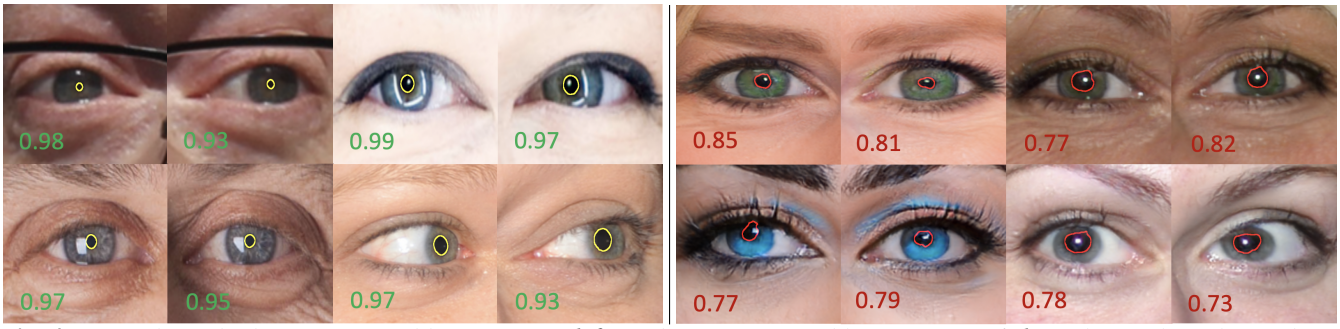

Fig. 4: *Examples of both eyes from real human faces (**left**) and GAN generated human faces (**right**). The pixels of the predicted pupil mask within a distance $d = 4$ from the prediction boundary contours are highlighted. The Boundary IoU score ($d = 4$) between the predicted pupil mask and the ellipse-fitted pupil mask for each pupil is shown on the images.*

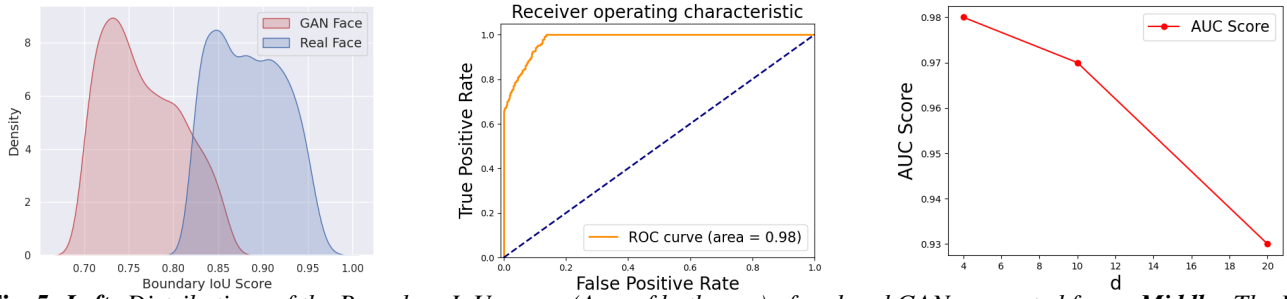

Fig. 5: ***Left:*** *Distributions of the Boundary IoU scores (Ave. of both eyes) of real and GAN-generated faces.* ***Middle:*** *The ROC curve is based on the Boundary IoU scores. The $d = 4$ for both figures.* ***Right:*** *BIoU hyper-parameter analysis, where x axis indicates the variation of hyper-parameter d and y axis is the AUC score.*

BIoU scores between the predicted pupil mask and ellipse fitted pupil mask. However, the artifacts of irregular pupil shapes lead to significantly lower BIoU scores.

We also show the distributions of the BIoU scores of pupils from the real faces and GAN-generated faces in Figure 5(left). We can find a clear separation between the distributions, indicating that irregular pupil shapes are an effective measure differentiating real and GAN-generated faces. The *receiver operating characteristic* (ROC) curve is shown in Figure 5(middle), which corresponds to an AUC (Area under the ROC curve) score of 0.94, indicating that irregular pupil shapes are effective to identify GAN-generated faces.

Hyper-parameter analysis. The BIoU measure has an essential parameter d, which indicates the distance to the boundary. Figure 5(right) shows how the performance of our method varies with the parameter d. As we have described, the BIoU will reduce to the Mask IoU when d is large enough and thus less sensitive to the boundaries, this is why the AUC score decreases as d increases.

Limitations. Our method has several limitations. We use irregular geometry shapes of the pupil to detect GAN-generated faces. And we may have false positives when the shapes are non-elliptical in the real faces; it happens when diseases and infection at the pupil and iris regions (See Figure 6(left)). Thus, our approach does not apply in this situation. However, this phenomenon is infrequent. We did not find such abnormal pupils in the real images of our dataset.

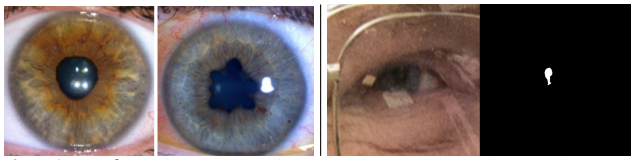

Fig. 6: ***Left:*** *There are abnormal pupils in the real images with non-elliptical shapes due to diseases and infection at pupil and iris regions, these real images examples are from [31].* ***Right:*** *Occlusions, noises around the pupil and fail pupil segmentation.*

Furthermore, occlusions on the pupil or fail pupil segmentation may also lead to wrong predictions (Figure 6(right))). Still, as we mentioned, with this cue, a human can visually find whether the face is real or not easily (See Figure 7).

5. DISCUSSION

In this paper, we presented that GAN-generated faces can be exposed with the irregular shapes of the pupils. The proposed approach is extremely effective and has good interpretations by measuring the elliptical shapes with the Boundary IoU scores. In the future, we will investigate other types of inconsistencies between two pupils of the GAN-generated face, such as different geometric shapes or different relative locations of pupils in the two eyes. These aspects may further improve the effectiveness of our method.

6. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

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

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] Ning Yu, Larry S Davis, and Mario Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *CVPR*, 2019.

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

[18] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva, "Incremental learning for the detection and classification of gan-generated images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2019, pp. 1–6.

[19] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring, "Detecting cnn-generated facial images in real-world scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 642–643.

[20] 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.

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

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

[23] Micah K. Johnson and Hany Farid, "Exposing digital forgeries through specular highlights on the eye," in *Information Hiding*, Teddy Furon, François Cayre, Gwenaël J. Doërr, and Patrick Bas, Eds., 2008, vol. 4567 of *Lecture Notes in Computer Science*, pp. 311–325.

[24] Luisa Verdoliva, "Media forensics and deepfakes: an overview," *arXiv preprint arXiv:2001.06564*, 2020.

[25] 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*, vol. 15, pp. 2944–2959, 2020.

[26] Caiyong Wang, Yunlong Wang, Boqiang Xu, Yong He, Zhiwei Dong, and Zhenan Sun, "A lightweight multi-label segmentation network for mobile iris biometrics," in *ICASSP*. IEEE, 2020, pp. 1006–1010.

[27] Caiyong Wang, Yunlong Wang, Kunbo Zhang, and etc., "Nir iris challenge evaluation in non-cooperative environments: Segmentation and localization," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2021.

[28] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[29] Andrew Fitzgibbon, Maurizio Pilu, and Robert B Fisher, "Direct least square fitting of ellipses," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 5, pp. 476–480, 1999.

[30] Davis E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[31] 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.

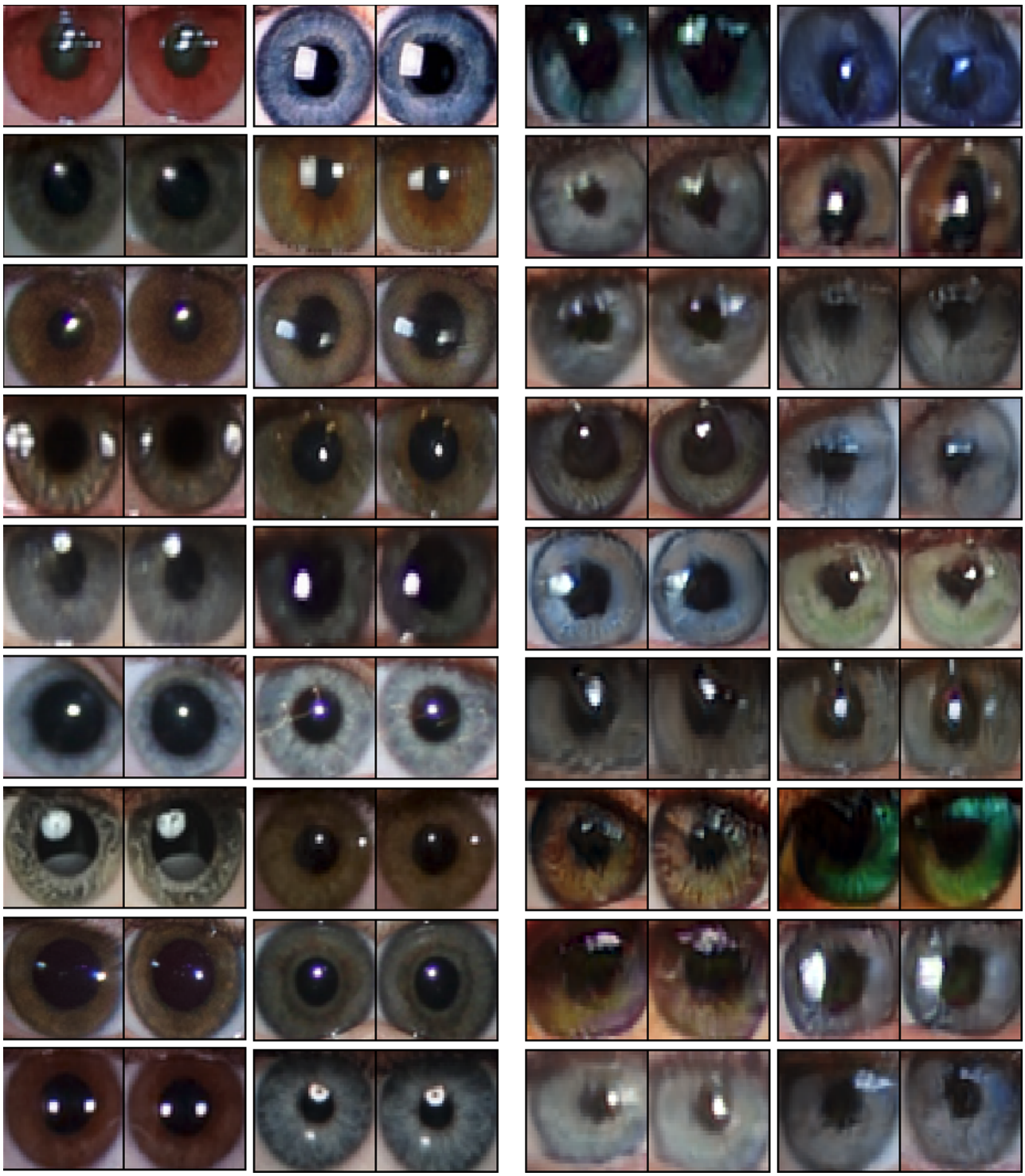

Fig. 7: *More pair of pupil examples from real (**left**) human faces and GAN-generated (**right**) faces. As we mentioned before, the irregular pupil shape is a good sign for the human to expose the GAN-generated face visually, even there are no boundary labels around the pupils, we can easily see that the shapes of GAN-generated pupils are very irregular, and the shapes of both pupils are very different in the same GAN-generated face image. In practice, people can zoom a face image large enough and then check the pupil shapes to find whether the face is real or not easily.*