ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414582

EXPOSING GAN-GENERATED FACES USING INCONSISTENT CORNEAL SPECULAR HIGHLIGHTS

Shu Hu[1], *Yuezun Li*[2], *and Siwei Lyu*[1]

[1]University at Buffalo, State University of New York, USA, {shuhu,siweilyu}@buffalo.edu
[2]Ocean University of China, China, liyuezun@ouc.edu.cn

ABSTRACT

Sophisticated generative adversary network (GAN) models are now able to synthesize highly realistic human faces that are difficult to discern from real ones visually. In this work, we show that GAN synthesized faces can be exposed with the inconsistent corneal specular highlights between two eyes. The inconsistency is caused by the lack of physical/physiological constraints in the GAN models. We show that such artifacts exist widely in high-quality GAN synthesized faces and further describe an automatic method to extract and compare corneal specular highlights from two eyes. Qualitative and quantitative evaluations of our method suggest its simplicity and effectiveness in distinguishing GAN synthesized faces.

Index Terms— Media Forensics, GAN Image Detection

1. INTRODUCTION

The rapid advancements of the AI technology, the easier access to a large volume of online personal media, and the increasing availability of high-throughput computing hardware have revolutionized the manipulation and synthesis of digital audios, images, and videos. A quintessential example of the AI synthesized media are the highly realistic human faces generated using the generative adversary network (GAN) models [1, 2, 3, 4], Figure 1. As the GAN-synthesized faces have passed the "uncanny valley" and are challenging to distinguish from images of real human faces, they quickly become a new form of online disinformation. In particular, GAN-synthesized faces have been used as profile images for fake social media accounts to lure or deceive unaware users [5, 6, 7, 8].

Correspondingly, there is a rapid development of detection methods targeting at GAN synthesized faces [9, 10]. The majority of GAN-synthesized image detection methods are based on extracting signal level cues then train classifiers such as SVMs or deep neural networks to distinguish them from real images. Although high performance has been reported using these methods, they also suffer from some common drawbacks, including the lack of interpretability of the detection results, low robustness to laundering operations and adversarial attacks [11], and poor generalization across different synthesis methods. A different type of detection methods takes advantage

Fig. 1: *Examples of GAN synthesized images of realistic human faces. These images are obtained from* http://thispersondoesnotexist.com *generated with the StyleGAN2 model [4].*

of the inadequacy of the GAN synthesis models in representing the more semantic aspects of the human faces and their interactions with the physical world [12, 9, 13, 14]. Such physiological/physical-based detection methods are more robust to adversarial attacks and afford intuitive interpretations.

In this work, we propose a new physiological/physical-based detection method of GAN-synthesized faces that uses the inconsistency of the corneal specular highlights between the two synthesized eyes. The corneal specular highlights are the images of light emitting or reflecting objects in the environment at the time of capture on the surface of the cornea. When the subject's eyes look straight at the camera and the light sources or reflections in the surrounding environment are relatively far away from the subject (*i.e*, the "portrait setting"), the two eyes see the same scene and their corresponding corneal specular highlights exhibit strong similarities (Figure 2, left

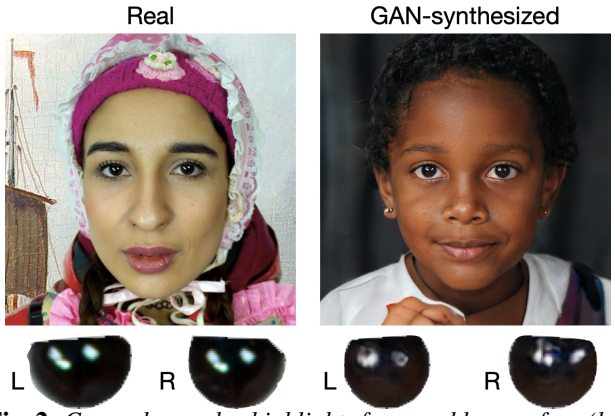

Fig. 2: *Corneal specular highlights for a real human face (left) and a GAN-synthesized face (right). The corneal regions are isolated and scaled for better visibility. Note that the corneal specular highlights for the real face have strong similarities while those for the GAN-synthesized face are different.*

image). We observe that GAN-synthesized faces also comply with the portrait setting (Figure 1), possibly inherited from the real face images that are used to train the GAN models. However, we also note the striking inconsistencies between the corneal specular highlights of the two eyes (Figure 2, right image). Our method automatically extracts and aligns the corneal specular highlights from two eyes and compare their similarity. Our experiments show that there is a clear separation between the distribution of the similarity scores of the real and GAN synthesized faces, which can be used as a quantitative feature to differentiate them.

2. BACKGROUND

Anatomy of Human Eyes. The human eye provides the optics and photo-reception for the visual system. Figure 3 shows the main anatomic parts of a human eye. The center of an eye is the iris and pupil. The transparent cornea is the outer layer that covers the iris and dissolves into the white sclera at the circular band known as the corneal limbus. The cornea has a spherical shape and its surface exhibits mirror-like reflection characteristics, which generates the corneal specular highlights when illuminated by light emitted or reflected in the environment at the time of capture.

GAN Synthesis of Human Faces. A series of recent works known as StyleGANs [2, 3, 4] have demonstrated the superior capacity of GAN models [1] trained on large sets of real human faces in generating high-resolution realistic human faces. A GAN model consists of two neural networks trained in tandem. The generator takes random noises as input and synthesizes an image, and the discriminator aims to differentiate synthesized images from the real ones. In training the two networks compete with each other: the generator aims to create more realistic images to defeat the discriminator, while the discriminator aims to improve the accuracy in differentiating the two

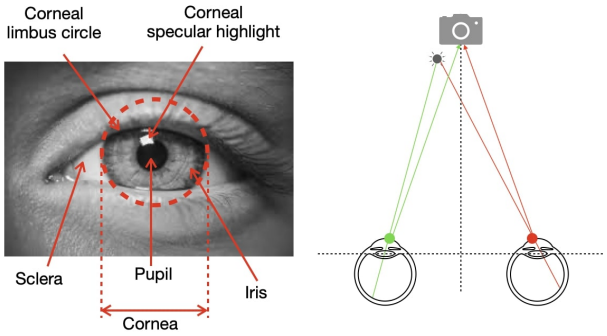

Fig. 3: *(left) Anatomy of a human eye. (right) The portrait setting with the corneal specular highlights.*

types of images. The training ends when the two networks reach an equilibrium.

Albeit the successes, GAN-synthesized faces are not perfect. Early StyleGAN model was shown to generate faces with asymmetric faces [12] and inconsistent eye colors [14]. However, the more recent StyleGAN2 model [4] further improves the synthesis quality and eliminate such artifacts. However, visible artifacts and inconsistencies can still be observed in the background, the hair, and the eye regions. One fundamental reason for the existence of such global and semantic artifacts in GAN synthesized faces is due to their lack of understanding of human face anatomy, especially the geometrical relations among the facial parts.

3. RELATED WORKS

Methods detecting GAN-synthesized faces fall into three categories. Those in the first category focus on signal traces or artifacts left by the GAN synthesis model. For example, earlier works, *e.g*, [15, 13], use color differences of first generation of GAN images. As color difference can be easily fixed, more sophisticated detection methods, *e.g*, [10, 16], seek more abstract signal-level traces or fingerprints in the noise residuals to differentiate GAN-synthesized faces. More recent works such as [17, 18, 19] extend the analysis to the frequency domain, where the upsampling step in the GAN generation leaves specific artifacts. The second category of GAN synthesized face detection methods are of data-driven nature [20, 21, 22, 23, 24], where a deep neural network model is trained and employed to classify real and GAN-synthesized faces. Methods of the third category look for physical/physiological inconsistencies by GAN models. The work in [12] distinguish GAN-synthesized faces by analyzing the distributions of facial landmarks, and [9] exposes the fake videos by detecting inconsistent head poses. The method in [14] further inspect more visual aspects to expose GAN synthesized faces. Such physiological/physical based detection methods are more robust to adversarial attacks and afford intuitive interpretations.

Because of the unique geometrical regularity, the corneal region of the eyes has been used in the forensic analysis of digital images. The work of [25] estimates the internal camera

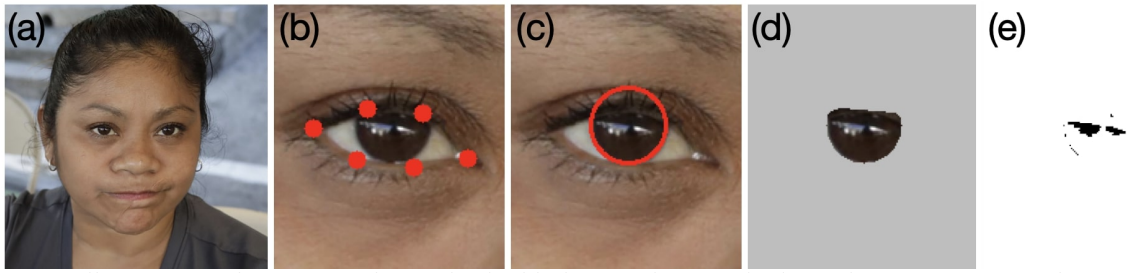

Fig. 4: *Overall process to obtain corneal specular highlight. (a) The input high-resolution face image. (b) Detection of facial landmarks around the eyes. (c) Hough circle detection of the corneal area. (d) Intersection of the eye region and circular corneal region. (e) Extracted corneal specular highlight area.*

parameters and light source directions from the perspective distortion of the corneal limbus and the locations of the corneal specular highlights of two eyes, which are used to reveal digital images composed from real human faces photographed under different illumination. The work of [14] identifies early generations of GAN synthesized faces [2] by noticing that they may have inconsistent iris colors, and the specular reflection from the eyes is either missing or appear simplified as a white blob. However, such inconsistencies have been largely improved in the current GAN synthesis models (*e.g*, [4]), see Figure 1.

4. METHOD

In this work, we explore the use of corneal specular highlight as a cue to expose GAN synthesized human faces. The rationale of our method can be understood as follows. In an image of a real human face captured by a camera, the corneal specular highlights of the two eyes are related as they are the results of the same light environment. Specifically, they are related by a transform that is determined by (1) the anatomic parameters of the two eyes including the distance between the centers of the pupils and the diameters of the corneal limbus; (2) the poses of the two eyeballs relative to the camera coordinate system, *i.e*, their relative location as a result of head orientation; and (3) the location and distance of the light sources to the two eyes, measured in the camera coordinates.

Under the following conditions, which we term as the *portrait setting* as it is often the case in practice when shooting closeup portrait photographs, the corneal specular highlights of the two eyes have approximately the same shape. To be more specific, what we mean by a portrait setting consists of the following conditions, which are also graphically illustrated in the right panel of Figure 3.

- The two eyes have a frontal pose, *i.e*, the line connecting the center of the eyeballs is parallel to the camera.
- The eyes are distant from the light or reflection source.
- All light sources or reflectors in the environment are visible to both eyes.

To highlight such artifacts and quantify them as a cue to expose GAN synthesized faces, we develop a method to automatically compare the corneal specular highlights of the two eyes and evaluate their similarity. Figure 4 illustrates the major steps of our analysis for an input image. We first run a face detector to locate the face, followed by a landmark extractor to obtain landmarks (Figure 4(b)), which are important locations such as the face contour, tips of the eyes, mouth, nose, and eyebrows, on faces that carry important shape information. The regions corresponding to the two eyes are properly cropped out using the landmarks. We then extract the corneal limbus, which affords a circular form under the portrait setting. To this end, we first apply a Canny edge detector followed by the Hough transform to find the corneal limbus (Figure 4(c)) and use its intersection with the eye region provided by the landmarks as the corneal region (Figure 4(d)).

We then separate the corneal specular highlights using an adaptive image thresholding method [26]. Because the specular highlights tend to have brighter intensities than the background iris, we keep only pixel locations above the adaptive threshold (Figure 4(e)). We align the extracted corneal specular highlights of the two eyes (denoted as R_L and R_R) with a translation, and use their IoU scores, $\frac{|R_L \cap R_R|}{|R_L \cup R_R|}$, as a similarity metric. The IoU score takes range in $[0, 1]$ with a smaller value suggesting lower similarity of R_L and R_R, and hence more likely the face is created with a GAN model.

5. EXPERIMENTS

The images of real human eyes are obtained from the Flickr-Faces-HQ (FFHQ) dataset [3], and the GAN synthesized human faces are from `http://thispersondoesnotexist.com`, which are created by the StyleGAN2 method [4]. There are 500 images for each of the types. The images have resolution of $1,024 \times 1,024$ pixels. We use the face detector and landmark extractor provided in DLib [27], and the Canny edge detector and Hough transform are from scikit-image [28].

Figure 5 shows examples of the analysis results for images of both real and GAN-synthesized human eyes. As described in the previous section, real human eyes captured by a camera under the portrait setting exhibit a strong resemblance between the corneal specular highlights of the two eyes, which are reflected by the higher IoU scores. On the other hand, the corneal specular highlights of the two GAN synthesized eyes may exhibit various types of inconsistencies, such as different numbers, different geometric shapes, or different relative lo-

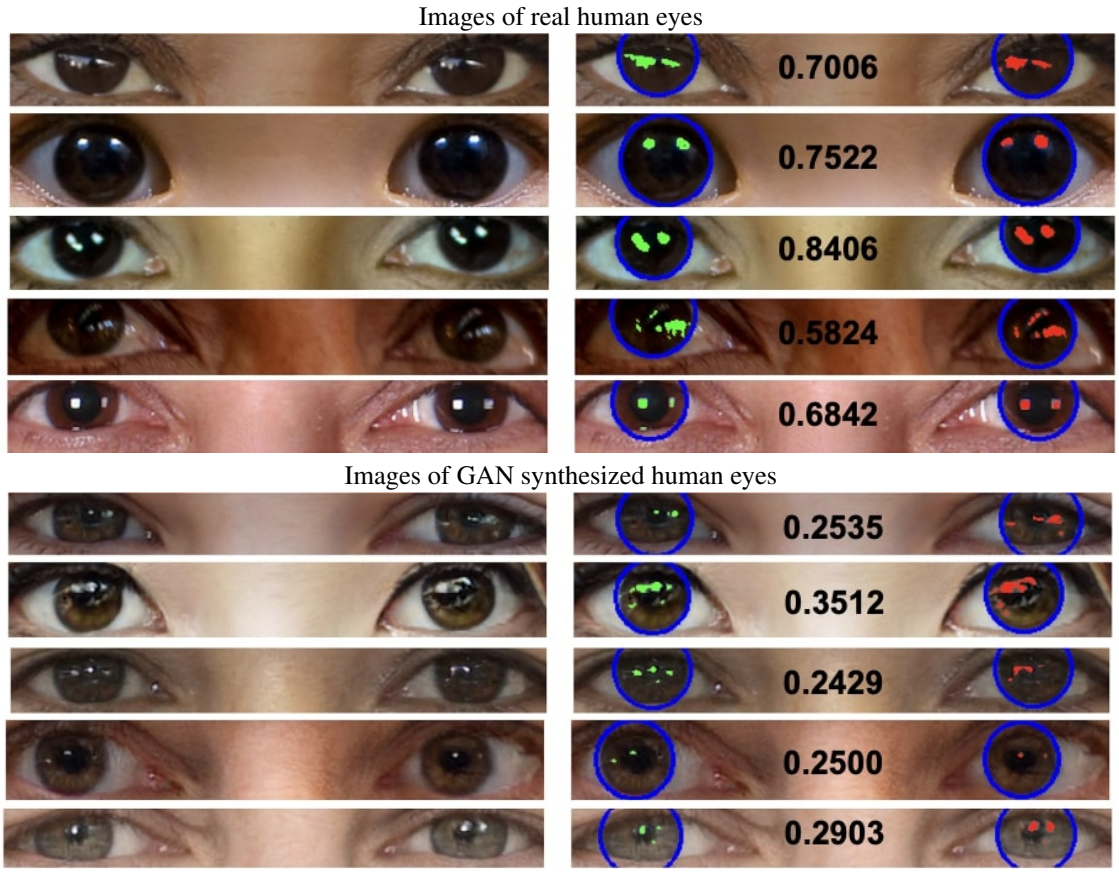

Fig. 5: *Corneal specular highlights from real human eyes (top) and GAN generated human faces (bottom). The right column corresponds to the detected corneal region (blue) and the specular highlights of two eyes (green and red). The IoU scores of the two corneal specular highlights are shown alongside the detections.*

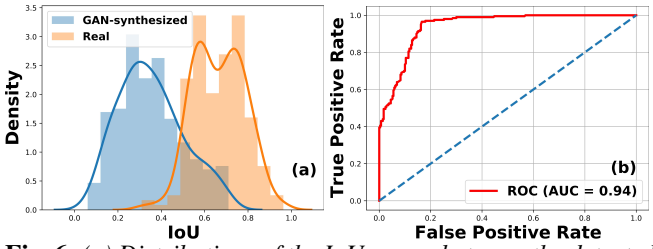

Fig. 6: *(a) Distributions of the IoU scores between the detected corneal specular highlights of two eyes for real and GAN synthesized faces. (b) The ROC curve based on the IoU scores.*

cations of specular highlight regions of the two eyes. These artifacts lead to significantly lower IoU scores. Figure 6(a) shows the distributions of the IoU scores of two eyes' corneal specular highlights for the real images and GAN generated images we collected. Consistent with the visual examples, there is a clear separation between the distributions, indicating that consistency of corneal specular highlights is an effective measure differentiating real and GAN generated faces. We also show the *receiver operating characteristic* (ROC) curve in Figure 6(b), which corresponds to an AUC (Area under the ROC curve) score of 0.94, indicating that corneal specular highlights are effective to identify GAN synthesized faces.

6. DISCUSSION

In this work, we show that GAN synthesized faces can be exposed with the inconsistent corneal specular highlights between two eyes. Although inconsistencies of specular patterns can be fixed with manual post-processing, it is expected to be non-trivial. Our method has several limitations. We only compare pixel differences without considering inconsistencies in geometry and scene. Also, we may have false positives when the portrait setting is not obeyed, *e.g*, when a light source is very close to the subject or a peripheral light source that is not visible in both eyes. It does not apply to images where specular patterns are not present. In the future, we will investigate these aspects and further improve the effectiveness of our method.

Acknowledgments. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123 and the Center for Identification Technology Research and the National Science Foundation under Grant No. 1822190. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense, the National Science Foundation, or the U.S. Government.

7. 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[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)*. IEEE, 2019, pp. 506–511.

[11] Nicholas Carlini and Hany Farid, "Evading deepfake-image detectors with white- and black-box attacks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. 2020, pp. 2804–2813, IEEE.

[12] 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.

[13] 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.

[14] 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.

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

[16] Ning Yu, Larry S Davis, and Mario Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7556–7566.

[17] Xu Zhang, Svebor Karaman, and Shih-Fu Chang, "Detecting and simulating artifacts in gan fake images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2019, pp. 1–6.

[18] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz, "Leveraging frequency analysis for deep fake image recognition," *arXiv preprint arXiv:2003.08685*, 2020.

[19] Ricard Durall, Margret Keuper, and Janis Keuper, "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7890–7899.

[20] 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.

[21] Michael Goebel, Lakshmanan Nataraj, Tejaswi Nanjundaswamy, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, and BS Manjunath, "Detection, attribution and localization of gan generated images," *arXiv preprint arXiv:2007.10466*, 2020.

[22] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, vol. 7.

[23] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr, "Global texture enhancement for fake face detection in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8060–8069.

[24] 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.

[25] 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.

[26] Jui-Cheng Yen, Fu-Juay Chang, and Shyang Chang, "A new criterion for automatic multilevel thresholding," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370–378, 1995.

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

[28] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, pp. e453, 2014.