DO YOU KNOW IF I'M REAL?
AN EXPERIMENT TO BENCHMARK HUMAN RECOGNITION OF AI-GENERATED FACES

STIJN KAS[1], THOMAS HES[2], BRIAN JANSEN[2] &
RUBEN POST[1]

[1] HU University of Applied Sciences Utrecht, Institute for ICT, Utrecht, The Netherlands, e-mail: ruben.post@hu.nl, stijn.kas@hu.nl
[2] Utrecht University, Faculty of Science, Utrecht, The Netherlands, e-mail: b.n.janssen@students.uu.nl, t.l.a.hes@uu.nl

Abstract With the development of advanced machine learning techniques, it is now possible to generate fake images that may appear authentic to the naked eye. Realistic faces generated using Generative Adversarial Networks have been the focus of discussion in the media for exactly this reason. This study examined how well people can distinguish between real and generated images. 30 real and 60 generated were gathered and put into a survey. Subjects were shown a random 30 of these faces in random sequence and asked to specify whether or not they thought the faces were real. Based on a statistical analysis, the participants were not able to reliably distinguish between all real and generated images, but real images were correctly distinguished in 81% of cases, where generated images were correctly distinguished in 61% of cases. Some generated images did receive very high scores, with one generated image being classified as real in 100% of the cases.

Keywords: generative adversarial networks, face generation, authenticity, experiment, survey.

University of Maribor Press

DOI https://doi.org/10.18690/978-961-286-362-3.36
ISBN 978-961-286-362-3

1 Introduction

The generation of realistic faces could have many implications for society (Stehouwer, Dang, Liu, Liu, & Jain, 2019). Several examples of where this could have a profound impact can be envisioned. Because faces of people who do not exist have no rights reserved to them in terms of privacy and royalties (Icons8, 2019), the use of human faces in products becomes less bothersome. Additional benefits might be found for purposes such as art, education and even missing persons investigations (Westling, 2019). However, there is reason to be concerned: for instance, a social media-based political campaign used generated faces to create the illusion of legitimacy (Graphika, 2019). With the speed of modern social media, these impressions can have a long-lasting impact, even if the source is debunked afterward (Westling, 2019).

If people cannot distinguish real from generated faces, we as a society may also be faced with some significant problems. There are some proposed methods to detect Generative Adverserial Networks (GAN) generated images automatically (Nightingale, Wade, & Watson, 2017; Xuan, Peng, Wang, & Dong, 2019), but the question is if these can keep up with the pace of development. Nightingale et al. (2017) stress that the importance of this question becomes evident when considering that in today's society we still rely on people to make judgments about image authenticity" (Nightingale et al., 2017). This same sense of legitimacy can be applied for criminal usage, as a generated face will not show up on Google Image search and will, therefore, appear more authentic. There is currently no way to prevent this kind of fraud in the criminal area (Nightingale et al., 2017).

This research aims to provide evaluate the current state-of-the-art face generating To conduct this evaluation, an experiment is conducted whereby thirty faces are shown to participants, where they indicate whether each image is real or generated individually.

2 Background & related work

This section describes the background regarding GAN generated faces and work relevant to determining the authenticity of generated faces.

2.1 General Adversarial Networks

Complex machine learning techniques, deep learning in particular, have much potential to transfer knowledge previously only interpretable to humans over to machines (Bengio et al., 2009). One deep learning method is deep generative models: a method of unsupervised learning which aims to learn better how to predict data. A recent development in this technology was proposed by Goodfellow et al. (Goodfellow et al., 2014). They propose a framework for generative models which makes use of adversarial networks, where the generative model gets an adversary (competitor) to test their method against. This leads to better results as the models keep each other in check. In the framework proposed by Goodfellow et al. (Goodfellow et al., 2014) the model passes noise through a multilayer perceptron (a type of artificial neural network) (Pal & Mitra, 1992) to create randomness, which allows it to generate a new image based on the real world examples it has been taught. This method can be described as an adversarial network, using deep generative models. The synthesis of these methods created Generative Adversarial Networks, which is the technology used to generate the faces for this research.

2.1.1 Related work

There have been multiple studies that have researched methods for successfully recognizing generated fake images with artificial intelligence, machine learning, and other novel detection techniques (Hsu, Zhuang, & Lee, 2020; Kim, n.d.; Stehouwer et al., 2019; R. Wang et al., 2019; Yu, Davis, & Fritz, 2019). Xuan et al. (Xuan et al., 2019) for example propose training a forensics model that can detect GAN generated images on its own. Marra et al. (Marra, Gragnaniello, Cozzolino, & Verdoliva, 2018) managed to detect GAN generated images on social media with a 95% accuracy. A problem with these methods could be that they rely on artificial methods of recognition such as neural networks and other unsupervised learning methods, which means they do not entirely hold a solution that maintains human agency in recognizing what is fake and what is real. These methods could however

still be imperative in combating the negative implications of generated faces mentioned in the introduction.

Whether humans are as of yet capable of recognizing GAN generated faces is still the object of study. Nightingale et al. (Nightingale et al., 2017) proposed a similar research method as this paper, but with more generic images (no faces) that were doctored physically by humans. A website called whichfaceisreal.com has built an experimental design comparable to ours. It is part of an effort to make people more aware of deception: the 'calling bullshit project' (Bergstrom & West, 2019). Sadly it does not seem to use its potential for data collection to analyzing the degree of perceived authenticity of these faces. It is clear this is an angle that still has to be explored in detail. This gap in the state of the art is where this study finds its relevance.

Rossler et al. (Rossler et al., 2019) performed an experiment with a similar set-up as this study. 204 participants were shown either a real image or an image generated by one of five technologies. They were given only between 2 to 6 seconds to observe the image. Subsequently the participants had to indicate whether or not the image was real. Rossler et al. (Rossler et al., 2019) claim to have found a correlation between video quality and the ability to detect whether or not the image was fake. Important lessons can be learned from their experimental setup, namely that variables like image resolution and observer time constraints are important factors to consider. In a study researching manipulated image credibility across platform, Shen et al. (Shen et al., 2019) found photo-editing experience and social media use were significant predictors of image credibility evaluation. In other words, people who were more experienced with social media and photo-editing were better at spotting manipulation. The same may be true for people with experience in facial generation technology and must be taken into account for this study.

3 Methods

This research aims to evaluate the current state-of-the-art face generating algorithms. To formalize this research goal, a research question was formulated:

RQ: How well are humans able to distinguish between real and computer-generated faces?

3.1 Variables

As part of the research design, four independent variables and three dependent variables were formulated.

3.1.1 Independent variables

Timeout Whether a participant has a maximum time of 5 seconds to view the image.
Image Which image the participant is shown and whether this image is generated or not.

General participant information Information about each participant that gives an indication of representation of the taken sample for the population. This information is age, sex, highest received degree, and race.

Technology familiarity An indication of the familiarity of the participant with artificially generated faces and generative adversarial networks.

3.1.2 Dependent Variables

Correctness Whether the chosen answer for a given image is the correct answer.

Response-time Time needed for a participant to decide whether the image is generated or not.

Accuracy confidence An indication of the confidence the participant has that the selected answers are correct.

To complement the experiment, additional knowledge questions were formulated. Of particular interest is the influence of the various independent variables on the dependent variables. More specifically, factors that might influence the correctness DV such as age, race, and familiarity with the technology are to be elaborated, as well as the influence of the added timeout. Lastly, correlations between correctness, response time and accuracy confidence might provide additional insight.

3.2 Hypotheses

Following the research question defined in Section 3 a set of hypotheses is formulated.

Hypothesis 1: People are able to distinguish between real and generated images.

Hypothesis 2: Time pressure affects people's ability to distinguish between real and generated images.

Hypothesis 3: Response time has an effect on people's ability to distinguish between the images.

Hypothesis 4: People can accurately guess how well they can distinguish real and generated faces.

Hypothesis 5: Technology familiarity has an effect on people's ability to distinguish between the images.

Hypothesis 6: There are differences between the demographics

3.3 Method

To conduct the research, an independent measures experiment is conducted whereby each participant is asked to complete the experiment one time the experiment will be operationalized through online survey tool Qualtrics (Qualtrics, 2014). Through Qualtrics, a standardized experiment environment is created, while allowing the experiment to be conducted by the participants without the supervision of the researchers. In the Qualtrics environment, a survey is adapted from a different study, consisting of three sections: general information, facial images and then a set of reflection questions (Mathur & Reichling, 2019).

In the general information section, participants will need to answer general personal information about age, sex, race, and education. Additionally, the participants need to indicate their familiarity with technology to generate artificial faces and generative adversarial networks. Following the general information, the participants first get the instruction page to prepare them for the experiment section. In this section, each participant is shown a random image from a pool of 90 images. The page layout for this experiment section is derived from (Mathur & Reichling, 2019). For the images, three different datasets were used: thispersondoesnotexist.com (P. Wang, n.d.), Generated Photos (Generated Media, n.d.), and Flickr-Faces-HQ Dataset (NVlabs, n.d.). From each of these datasets, 30 random images were selected, meaning 60 images are generated faces and 30 images are real faces. For each image, the participants have to indicate if this image is real or fake. Half of the participants will have a 5 second time limit to view the image, after this time limit the image will disappear and the participant is encouraged to make a decision. At the very beginning of the survey, a random Boolean is generated and saved that indicates if this participant has a time limit. This value is used internally within Qualtrics to determine if a participant will receive a time limit and is used for data analysis.

Ending the questionnaire, the participants are asked what distinctive features made the participants decide whether a face was real or fake. Through this information, an attempt is made to distinguish new opportunities for either improving face generation or improving generated image detection.

3.3.1 Participants

By distributing the experiment through an online platform more participants can be reached. Sampling is mainly achieved through convenience sampling since the researchers are all students most of the participants are expected to be students as well. After data collection a total of 107 unique participants were obtained. However, a large proportion of these responses were not used as they did not pass data cleaning. This resulted in 59 unique participants with a correct response.

3.3.2 Context

Several contextual factors are accounted for through the use of advanced Qualtrics features. Firstly, participants are only able to perform the experiment on a non-mobile device, as asserted by a default Qualtrics feature to exclude all mobile users. Secondly, all participants perform the experiment on a screen larger than 600 by 600 pixels (HD), as asserted by a custom JavaScript setting. Lastly, participants were asked to set their screen brightness to the maximum and set their window to full screen and asked to perform the experiment individually without distraction by other people to minimize differences between participant settings.

3.3.2 Instrumentation

The survey first asks general demographics: age, sex, level of education, race and technology familiarity. Then, an instruction page is shown, explaining the general layout of the experiment. Lastly, the participants are informed their response time is measured and they will be asked about the experiment afterward to see if they found any distinctive features. Once the participant clicks the next button, they will be shown a layout with an image in the center of the screen, two buttons above it, and a title and progress above them. The second part asks how confident they were in discerning the faces. Lastly, two text entries are available, one to ask any questions and one to enter an email to receive the results of the study.

3.3.3 Data collection procedure

With the use of Qualtrics, the data collection procedure is fully automated, and a single URL was used to distribute the survey. To decide whether a participant receives a timeout, a Boolean variable is generated through JavaScript and added as an answer to an invisible question. To iterate through the dataset of 90 images but only show each participant 30 images, Qualtrics' loop and merge feature was used. All images were uploaded as loop and merge entries, and the order was set to random. A timing question was added below the image, automatically recording four values (in milliseconds after the page loaded): 1) first click, 2) last click, 3) page submit and 4) click count. Within the loop and merge JavaScript three settings were added: the title was enriched with a progress indicator, as shown in Appendix C, and if the participant received a timeout, the image would be automatically hidden five seconds after it fully loaded. To compensate for slower internet speeds, the third setting disabled the timing question until after the image was fully loaded. Once the survey is completed, the data is available through the Qualtrics platform, and exported to csv for analysis, as described in Section 4.

4 Analysis and execution

Basic descriptive statistics from the experiment are listed in Appendix E and Tables 2 to 4. Some Likert values had to be recoded from text to an integer. These values are listed in Appendix E. As Appendix E shows, the distribution between timeouts and no timeouts was a little skewed because of the random assignment to participants. Additionally, the participants were between 17 and 60 years old, with most of them between 21 and 26, as shown in Table 2.

Table 1: Age statistics

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
17.00	21.50	23.00	29.36	26.00	67.00	4.00

4.1 Analysis

From the experiment results, four additional metrics were developed. These metrics will be used to answer the research question and analyze the hypotheses. First, to gauge the accuracy of a participants distinction between the real and generated faces, a metric was developed which will be referred to as ”score” from this point on. The score metric was simply calculated by dividing the number of images where the participant got the right answer by the total number of images which was shown to the participant and can essentially be seen as the percentage the participant chose the right answer. Second, in order to analyze whether response time had any effect on the participants' scores, the 'meantime' metric was calculated, by taking the mean of all recorded times which were recorded for that participant. Next, for each participant individually, a metric was developed to measure the correlation between their scores and their time to answer, for each individual image. This was calculated using the built-in correlation feature from R on the participant's time to answer the question and a 1 for right answers and 0 for false. The last metric determined the percentage of real/fake choices per image, again by simply dividing the right options by the total amount the image was shown for each image.

4.2 Hypothesis testing

4.2.1 H1: People are able to distinguish between real and generated images

The scores for each participant gave insight into this hypothesis. As Table 3 shows, the mean of the scores is 0.69, with a standard deviation of 0.12. If people were to random guess each image then the score for each image would be 0,5. When comparing the scores of fake images with a set of data with the same size and all being 0,5 as score a t-test ($t(50)=3.37$, $p=0.00135$) indicates that people are able to distinguish a fake image and are not randomly guessing. This means that the null hypothesis cannot be rejected.

Table 2: General scoring statistics

Value	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.Dev.
Score	0.37	0.60	0.67	0.68	0.77	0.97	0.12
Meantime	1.62	2.70	3.51	3.73	4.27	9.34	1.55
Totaltime	48.55	81.08	105.42	111.89	128.04	280.32	46.42
Real percentage	0.10	0.27	0.33	0.34	0.40	0.53	0.08
True images	3.00	8.00	10.00	10.27	12.00	16.00	2.52
False images	14.00	18.00	20.00	19.73	22.00	27.00	2.52

Additionally, Table 4 shows the difference between the percentage each individual image was chosen as real or fake, for the images that were real and fake respectively. Results of the independent sample t-tests indicated that there was a significant difference in scores for real (M=0.83, SD=0.13) and fake (M=0.61, SD=0.25) images, ($t(87) = 6.07$, $p = 3.218e\text{-}08$).

However, six out of the 59 participants had a score of 0.5 or lower, and were not able to distinguish between the images as a score of 0.5 is equal to chance. The lowest score of 0.37, showed this one participant only choosing the right label in 37% of the images.

Table 3: Real and fake image accuracy statistics

Image	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.Dev.
Real	0.40	0.76	0.84	0.81	0.90	1.00	0.13
Fake	0.00	0.42	0.67	0.61	0.81	1.00	0.25

When the data is looked at individually in detail, some fake images scored remarkably high. One such image fooled all respondents, see Figure 1.

Figure 1: Generated face with an accuracy score of 0%

4.2.2 H2: Time pressure affects people's ability to distinguish between real and generated images

To analyze this hypothesis, a t-test was used based on the scores for participants which received a 5 second time limit and participants who did not. This independent sample t-test indicated that there were significant differences: the participants receiving a time limit (M=0.65, SD=0.098) scored significantly lower than participant not receiving a time limit (M=0.72, SD=0.13),($t(44) = -2.32$, $p = 0.025$). However, the differences in the same groups when comparing the average time each participant took to answer the questions was not significantly different, ($t(39) = -1.83$, $p = 0.075$).

This shows that participants, on average, scored significantly lower when receiving a five second penalty. Additionally, although the difference is not statistically significant, the participants with a timeout were around 18% quicker to decide between the images on average. This conclusion rejects the null hypothesis and affirms the alternative hypothesis: Time pressure negatively affects people's ability to distinguish between real and generated images.

4.2.3 H3: Response time has an effect on people's ability to distinguish between the images

For this hypothesis, the Pearson's product-moment correlation is used because it works well with the available data. The correlation analysis was performed on the average time the participants took to respond and their scores. Results of the test indicated that there was no significant correlation between the mean time and scores, ($r(57) = 0.14$, $p = 0.30$). However, when a correlation analysis is performed on a per-participant basis, and their time to answer is analyzed compared to their correct or incorrect choices, a slight negative mean is found, see Table 5. This is in contrast to what would be expected when looking at the total correlation.

Table 4: Time vs score correlation

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.Dev.
-0.53	-0.22	-0.08	-0.06	0.09	0.37	0.20

A negative correlation means that for images where participants took longer to react, they scored slightly lower on average. This can be interpreted as harder questions taking longer to answer or people's first instinct being better than a nuanced answer, but the data cannot provide the true answer.

Given the results of the Pearson correlation test the null hypothesis cannot be rejected, as there is no significant correlation.

4.2.4 H4: People can accurately guess how well they can distinguish between the images

For this hypothesis, the participants gave an answer to the following question on a 5-point likert scale: "I am confident I was always able to discern which faces were generated and which ones were real".

From this 5 point Likert scale, a correlation analysis was ran compared to their actual score. The results of the Pearson's product-moment correlation test indicated that there was a weak positive association between their answer to the question and their accuracy score, ($r(57) = 0.21$, $p = 0.104$). While there is a positive relation, it is not statistically significant, and therefore the null hypothesis cannot be rejected.

4.2.5 H5: Technology familiarity has an effect on people's ability to distinguish between the images

For this hypothesis, the participants gave an answer to a different question on a 5 point Likert scale: "To what degree are you familiar with technology to generate artificial faces, including generative adversarial networks?".

From this 5 point Likert scale, a correlation analysis was ran compared to their actual score with their familiarity. The results of the Pearson correlation indicated that there was a slightly positive association between their answer to the question and their accuracy score, ($r(57) = 0.16$, $p = 0.23$). While there is a positive relation, it is not statistically significant, and therefore the null hypothesis cannot be rejected.

4.2.6 H6: There are differences between the demographics

For the demographics, three different aspects are investigated: 1) age, 2) gender and 3) education level.

For differences in age results, two correlation analyses were performed: the first Pearson correlation indicated that there was a significant positive association between the participants' age and their time to decide whether the images were real or fake, ($r(52) = 0.31$, $p = 0.020$). The second Pearson correlation indicated a statistically insignificant negative association between the participants' age and their accuracy scores, ($r(53) = -0.25$, $p = 0.06$). Thus, we can conclude that, with statistical significance, older participants scored took longer to make their choices, but did not necessarily score lower.

For gender and education level, two t-tests were performed. The first results for the independent sample t-test indicated that there were no significant differences in scores between males (M=0.67, SD=0.14) and females (M=0.69, SD=0.09), ($t(54) = -0.55$, $p = 0.582$).

For the education level, two groups were formed and another t-test was performed. One group was formed of participants with a bachelor degree or higher and the other of participants without a bachelor degree or higher. This independent t-test indicated that there was no significant difference between students with a bachelor degree or higher (M=0.68, SD=0.11) and students without a bachelor degree (M=0.67, SD=0.13), ($t(39) = 0.44$, $p = 0.664$). Therefore, because the age has a significant correlation with time to decide, the null hypothesis can be rejected.

5 Discussion

In this research, internal validity is fairly well covered. However, some selection bias may still occur during sampling as convenience sampling was the main type of sampling used. Additionally, as the convenience sample originated from students within the IT domain, the sample might be skewed towards IT. Since people in the IT domain are, on average, more familiar with face-generating technology and how to recognize generated faces, this may provide an unbalanced sample. Many external validity threats were mitigated. However, the experimenter effect might still occur if the researcher was present and gave further instructions during the experiment. For this reason, the researchers were vigilant to not instruct the participants more than the instructions given in the experiment. Lastly, since the experiment was performed through an online survey, not all situational and context factors could be accounted for, potentially negatively impacting the reliability of the research.

The findings might be generalizable in a broader sample, as the difference between industries has not explicitly been measured within the experiment. Additionally, one related technology might see similar results in a similar setup: doctored videos. Through use of deep fakes, these videos have risen in popularity and notoriety, and the results might be comparable. Furthermore, the experimental design used in this research is easily adaptable to measuring the same variables for doctored videos.

6 Conclusions and future work

In this experiment, 59 participants iterated through 30 random selected images, where the participant can choose if they think it is real or generated. Based on a statistical analysis, the participants were able to distinguish between real and generated images, but not reliably. Half of the participants were given only five seconds to decide between the images, after which the image disappeared. These participants performed significantly worse than the participants who did not receive such a time limit. How familiar a participant was with face generating technology had no effect on their ability to recognize generated faces, and people who thought they scored well scored a little higher than people who did not. Lastly, correlation analysis showed older participants took significantly longer to decide than younger participants.

While participants were able to distinguish between real and generated images, some individual generated images were thought to be real very often, with one image fooling every participant into thinking it was real. While this does not reject the null hypothesis, it does provide valuable insight: if someone were to use these generated images for malicious purposes, they might filter through them first and pick ones they consider normal looking. If they pick "good" images, people will not be able to distinguish between real and generated images. Because the image selection for this study was performed completely randomly from an image generator this was outside of the scope of this research.

For future work, researchers may manually select images they think are good and compare them to real images, in order to see if they can find different results. Additionally, the experiment could be repeated with a larger, more diverse sample, possibly in an offline, controlled setting. Lastly, the same experimental design could be applied to computer-generated videos such as deepfakes, for which the societal impact is also very high.

References

Bengio, Y., et al. (2009). Learning deep architectures for ai. Foundations and trends R in Machine Learning, 2 (1), 1{127.

Bergstrom, C., & West, J. (2019). Calling bullshit: Data reasoning in a digital world. Retrieved from https://callingbullshit.org/

Generated Media, I. (n.d.). Generated photos. Retrieved from https://generated.photos/

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial networks.

Graphika. (2019). Operation #FFS:Fake Face Swarm. https://graphika.com/uploads/Graphika%20Report%20-%20OperationFFS Fake Face Storm.pdf

Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences, 10 (1), 370.

Icons8. (2019, Sep). AI-Generated Faces: Free Resource of 100K Faces Without Copyright. prototypr.io. Retrieved from https:// blog.prototypr.io/generated-photos-free-resource-of-100k -diverse-faces-generated-by-ai-2144a8615d1f

Kim, G. L. (n.d.). Quality evaluation of synthesized human face images from generative adversarial network.

Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018). Detection of gan-generated fake images over social networks. In 2018 ieee conference on multimedia information processing and retrieval (mipr) (pp. 384{389).

Mathur, M. B., & Reichling, D. B. (2019). Open-source software for mouse-tracking in qualtrics to measure category competition. Behavior research methods, 51 (5), 1987{1997.

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes? Cognitive research: principles and implications, 2 (1), 30.

NVlabs. (n.d.). Flickr-Faces-HQ Dataset (FFHQ). Retrieved from https:// github.com/NVlabs/ffhq-dataset of California, S. (n.d.). Bill ab-730. Retrieved from https:// leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill id=201920200AB730

Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classi ca-tion. IEEE Transactions on neural networks, 3 (5), 683{697.

Qualtrics, L. (2014). Qualtrics [software].

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nie ner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the ieee international conference on computer vision (pp. 1{11).

Shen, C., Kasra, M., Pan, W., Bassett, G. A., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The e ects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. new media & society, 21 (2), 438{463.

Stehouwer, J., Dang, H., Liu, F., Liu, X., & Jain, A. (2019). On the detection of digital face manipulation. arXiv preprint arXiv:1910.01717 .

Wang, P. (n.d.). This person does not exist. Retrieved from https:// thispersondoesnotexist.com/

Wang, R., Ma, L., Juefei-Xu, F., Xie, X., Wang, J., & Liu, Y. (2019). Fakespotter: A simple baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122 .

Westling, J. (2019). Are deep fakes a shallow concern? a critical analysis of the likely societal reaction to deep fakes. A Critical Analysis of the Likely Societal Reaction to Deep Fakes (July 24, 2019).

Xuan, X., Peng, B., Wang, W., & Dong, J. (2019). On the generalization of gan image forensics. In Chinese conference on biometric recognition (pp. 134{141).

Yu, N., Davis, L. S., & Fritz, M. (2019). Attributing fake images to gans: Learn-ing and analyzing gan ngerprints. In Proceedings of the ieee international conference on computer vision (pp. 7556{7566).