2023 24th International Conference on Digital Signal Processing (DSP) | 979-8-3503-3959-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/DSP58604.2023.10167876

COULD HUMAN GAZE AUGMENT DETECTORS OF SYNTHETIC IMAGES?

Nikolaos Fotopoulos[1], *Clara Riedmiller*[2], *Efe Bozkir*[2,4], *Panagiotis Tsinganos*[1], *Dimitris Ampeliotis*[3], *Gjergji Kasneci*[5], *Enkelejda Kasneci*[4] *and Athanassios Skodras*[1]

[1]*Department of Electrical and Computer Engineering, University of Patras*, Greece
[2]*Department of Computer Science, Eberhard Karls University of Tübingen*, Germany
[3]*Department of Digital Media and Communication, Ionian University*, Greece
[4]*Human-Centered Technologies for Learning, Technical University of Munich*, Germany
[5]*Responsible Data Science, Technical University of Munich*, Germany

ABSTRACT

Recent advances in generative adversarial networks (GANs) allow for the synthesis of extremely photo-realistic face images, deceiving even the most experienced observers, let alone the unsuspecting internet user. Due to this, there has been a considerable effort by the image forensics community to design appropriate tools for the detection of these images. This paper first implements one such detection technique based on spatial and cross-band co-occurrence matrices and convolutional neural networks (CNNs), and then attempts to improve it by introducing additional information obtained from the human gaze. We show that in cases where human observers correctly decide whether an image is real or fake, eye movement information in combination with spatial and cross-band co-occurrence matrices derived from observation regions can be informative towards the task of detecting fake images. However, only a limited increase in the detection accuracy is achieved.

***Index Terms*—** synthetic / fake images, eye tracking, gaze, image forensics, co-occurrence matrix, Generative Adversarial Networks

1. INTRODUCTION

With the advent of deep learning, and specifically Generative adversarial networks (GANs) [1], synthetic images have become more photo-realistic than ever. GANs not only enable the creation of entirely synthetic images from scratch, but they can also alter the characteristics (hair, race, gender, etc.) of an existing face [2]. In addition, the ease of access to such high-quality fake imagery[1] is a serious cause for concern, as the creation of fake online profiles and misinformation becomes incredibly easy.

The main purpose of this paper is to examine whether the human gaze can help synthetic image detectors increase their accuracy. The main idea is to quantify the human gaze with the help of gazemaps, and then incorporate the information from these gazemaps into an existing fake image classifier. This is an under-explored area in image forensics and can provide meaningful insights into how human and computer vision are connected.

First, an automatic detector is employed (base model), which relies on CNNs and pixel co-occurrences [3]. The human gaze is then incorporated into this detector by converting the available gazemaps to binary masks, and then calculating the co-occurrence matrices only in the areas that are not rejected by the masks. These matrices are then used to train the base model further. Our method relies on the assumption that the image regions that the observers focus the most on are the ones that carry the most important information (i.e., artifacts, asymmetrical irregularities, unnatural background, etc.) for the task of fake image detection.

2. RELATED WORK

Generative image modeling has advanced significantly lately, largely thanks to text-to-image generators like Stable Diffusion and DALL-E 2 [4]. In this work, however, we consider GAN-based architectures for the creation of synthetic images. There are primarily two lines of research for the discrimination of real from GAN-generated images: computer-driven (objective) methods and perceptual (subjective) studies.

2.1. Computer-driven methods

There are various approaches for the automatic detection of GAN images. Marra et al. [5] established that GANs embed a kind of fingerprint into each generated image, in a similar way that photos produced by real cameras bear a device-dependent signature due to manufacturing imperfections of the camera sensors. Zhang et al. [6] exploited the fact that the up-sampler of the GAN generator introduces image artifacts, which manifest as periodic peaks in the Fourier spectrum. They, therefore, proposed a classifier trained on the spectrum of the image rather than the image itself. Nataraj et al. [7] computed the co-occurrence matrices of the image's R, G, and B channels and fed them into a CNN. Barni et al. [3] built on this approach and made the detector more robust to image transformations, by additionally computing the co-occurrence matrices between the color channels (cross-band).

This research was supported by the DAAD exchange program OMEGA (Egocentric Perception, Interaction and Computing in the Deep Learning Era - project 57515461).

[1] `https://thispersondoesnotexist.com`, last access 03/17/2023.

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

2.2. Perceptual studies

From the human point of view, recent studies showed that it has been getting increasingly challenging for humans to distinguish synthetic faces from real ones. Lago et al. [8] conducted a crowd-sourcing survey, in which they aimed to investigate how good humans are at recognizing GAN-generated images. The GAN images were taken from three different recent GAN architectures: ProGAN [9], StyleGAN [2], and StyleGAN2 [10], which they called AI-17, AI-18, and AI-19, respectively. They demonstrated that, on average, humans could recognize AI-17 images fairly well, but fail to accurately detect AI-18 and AI-19 images. More specifically, they classified AI-19 images as real more often than real images. Nightingale and Farid [11] performed a similar study, and they additionally concluded that synthetic faces are perceived as more trustworthy than real faces.

It is important to note that there are very few studies that measure the gaze of the observers when they are looking at an image. Caporusso et al. [12] carried out a study similar to the aforementioned studies, but they additionally measured each subject's gaze while they were observing each image. They found several statistically significant factors that characterize the high-accuracy group of subjects, such as the gaze spread and the gaze area.

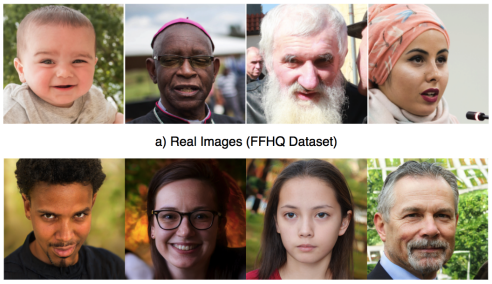

Fig. 1: a) Real faces from the FFHQ dataset (top row) and b) Generated faces from StyleGAN2 (bottom row).

Table 1: Results of the base model evaluated on the test set of 228 images ([11]).

	Detection Rate
REAL	111 out of 114 = 97.37%
AI-17	2 out of 38 = 5.26%
AI-18	14 out of 38 = 36.84%
AI-19	34 out of 38 = 89.47%
Total	161 out of 228 = 70.61%

3. AUTOMATIC METHOD

We used the approach of [3] as the base GAN image detector. In that work, the authors showed that while recent GAN architectures generate images that are extremely photo-realistic, they cannot accurately reproduce the spatial and spectral relationships between the image pixels. As a result, they used a CNN classifier which takes as input, not the image itself, but the spatial co-occurrence matrices of the image's R, G, and B color channels, as well as the cross-band co-occurrence matrices of pairs RG, RB, and GB.

3.1. Architecture

In terms of the architecture, we used the same model that was previously employed for GAN image classification in [3]. It consists of the following layers: Conv(32, 3x3) + ReLU + Conv(32, 5x5) + Pool + Conv(64, 3x3) + ReLU + Conv(64, 5x5) + Pool + Conv(128, 3x3) + ReLU + Conv(128, 5x5) + Pool + Dense(256) + ReLU + Dropout + Dense(256) + ReLU + Dropout + Sigmoid, where Conv(C, fxf) is a convolutional layer with C filters of size fxf, Pool is a max pooling layer and Dense(D) is a fully connected layer with D nodes.

Before calculating the co-occurrence matrices, the images were JPEG compressed with quality factors (QF) $\in \{75, 80, 85, 90, 95\}$. As a result, the model is able to recognize JPEG-compressed images, the most commonly found images on the web. After producing the $256 \times 256 \times 6$ tensor, we normalized it in the range [0,1] and then fed it into the network.

3.2. Dataset

The dataset used for training the base model consists of a total of 40000 images (20000 real and 20000 synthetic), of which for each class 12000 were used for training, 4000 were used for validation, and 4000 were used for testing. The real images were taken from the FFHQ Dataset [2], and the synthetic images were generated from the GAN model StyleGAN2 [10], which improves on the original StyleGAN model [2], and produces synthetic images of extremely high quality. It is visually very challenging to distinguish the face images generated by StyleGAN2 from the real faces, as they are mostly artifact-free, as some of them are depicted in Figure 1 (b).

3.3. Results

We employed the SGD optimizer with momentum, and used the following hyperparameters [3]: learning rate = 0.01, momentum = 0.9, decay = 2.5e-4, and gradient clip value = 0.5. The batch size was set to 40. The network was trained for a maximum of 40 epochs using early stopping which stopped the training process if the validation loss did not decrease after 8 epochs. After successful training, the model achieved a 94.85% accuracy on the test set.

The achieved detection rate is very impressive, but how well does the base model generalize to images from different GAN architectures? To answer this, we tested it on a dataset [8] which consists of 228 images, separated as follows: 114 real images from the FFHQ Dataset (REAL), 38 GAN images from ProGAN (AI-17), 38 GAN from StyleGAN (AI-18), and 38 GAN from StyleGAN2 (AI-19).

The results obtained are shown in Table 1. We notice that while the model performs well on the real images and images from AI-19, its detection rate drops significantly when the synthetic images come from a GAN model that was not used in training.

4. GAZE INTEGRATION

As far as the human gaze is concerned, we used the eye-tracking data from [13]. In this study, 22 subjects viewed a total of 72 images that were separated as follows: 36 real images, 12 AI-17 images, 12 AI-18 images, and 12 AI-19 images. These images are a subset of Lago et al's. [8] dataset, which contains 300 images: 150 real images, 50 AI-17 images, 50 AI-18 images, and 50 AI-19 images. The remaining 228 images are the ones we used previously to evaluate the base model and they were not used in [13].

4.1. Gazemaps

As mentioned previously, we used 2D attention maps (gazemaps) to represent the human gaze. Given a sequence of 2D coordinates (i.e., (x_i, y_i)) of the viewer's gaze at time indices t_i when viewing image $I \in M \times N$, a gazemap J is created as summarized in Algorithm 1.

Algorithm 1 Gazemap generation process

1. Create J of size $M \times N$ initialized with zeroes.
2. For every coordinate (x_i, y_i) add 1 to the respective position of J, namely: $J(x_i, y_i) \mathrel{+}= 1$.
3. Perform Gaussian blurring to J, by appropriately selecting kernel size k and standard deviation σ.
4. Normalize J in the range [0,1].
5. Return J

4.2. Incorporating the gaze

In order to incorporate the human gaze in the base model, the strategy we adopted is based on the following assumption: as long as the observer correctly classifies an image as real or fake, the regions that the observer focused on contain the necessary information to determine if the image is real or fake.

As a result, we investigate whether co-occurrence matrices computed only at the regions that the observers focused on the most can help the already trained base model improve its performance. More specifically, the steps we take to integrate the gaze data for an image are described in Algorithm 2. Figure 2 illustrates this procedure.

Algorithm 2 Gaze integration process

1. Compute the gazemap from the observer's coordinates, as described in Algorithm 1.
2. Convert the gazemap to a binary mask by selecting a suitable threshold t $\in$ [0,1]. Values exceeding t remain unchanged. Otherwise, they are replaced by the value -1.
3. Keep only the image regions that are not rejected by the mask (critical regions), because they are the ones that humans rely on the most to determine whether the image is real or synthetic.
4. Compute the co-occurrence tensor only in the critical regions.

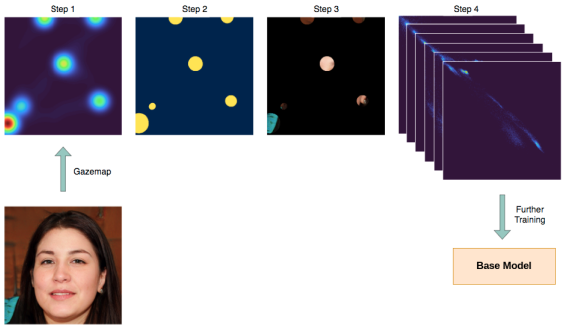

Fig. 2: Scheme of the gaze integration method. From left to right: gazemap with k = 199 and σ = 25, conversion to a binary mask with threshold t = 0.3, application of the mask on the image (critical regions), final co-occurrence tensor of size $256 \times 256 \times 6$.

Table 2: Number of available gazemaps for each class of images, separated by the observer's prediction, for each phase.

Class	Observer's Prediction	
	Correct	Incorrect
REAL	579	213
AI-17	241	23
AI-18	151	113
AI-19	78	186
Total Amount (1584)	1049	535

5. EXPERIMENTAL ANALYSIS

The study of Riedmiller [13] was split into two phases, namely, viewing and rating. The observers were split into two groups which we call Group 1 and Group 2. During the viewing phase, Group 1 was primed on the truthfulness of the image, while Group 2 had no such a-priori knowledge. The rating phase, however, was the same for both groups. In total, we obtain 22*72 = 1584 gazemaps for each phase. Table 2 shows the number of available gazemaps for each class of images. It is noted that there are more gazemaps corresponding to incorrect predictions for AI-19 images, as observers classified AI-19 incorrectly more often than correctly.

In the following experiments, we only use gazemaps that correspond to correct observer predictions, of which 80% of the co-occurrence gaze data is used to further train the base model and the 20% is used for validation. After obtaining the validation results and seeing how well the model learns from the gaze data, the base model is retrained with all the available gaze data and tested on the test set of 228 images.

Due to the large number of experiments, it was not possible to conduct an exhaustive grid search in order to find the best hyperparameters for each experimental setup. Instead, we compared the results for the same set of hyperparameters, i.e., under the same conditions, which are reported as follows: optimizer = SGD with 0.5 momentum, batch size = 32, learning rate = 0.005, epochs = 5, decay = 0.001, and gradient clip value = 0.5. It is noted that the training is done with a small number of epochs and a small learning rate to avoid overfitting the gaze data and losing the knowledge that the base model already has.

Table 3: Experiments performed for different values of threshold t, kernel size k, and standard deviation σ. The chosen parameters and the validation results are displayed for each setup.

Setup Number	Conditions					
	t	Groups	Phases	k, σ	Val. Loss	Val. Acc.
1.1.1	0.3	1, 2	V, R	15, 3	0.6184	68.70%
1.1.2	0.6	1, 2	V, R	15, 3	0.6905	56.91%
1.2.1	0.3	1, 2	V, R	65, 15	0.4564	79.67%
1.2.2	0.6	1, 2	V, R	65, 15	0.6020	69.31%
1.2.3	0.9	1, 2	V, R	65, 15	0.6877	58.74%
1.3.1	0.3	1, 2	V, R	199, 25	0.3309	86.59%
1.3.2	0.6	1, 2	V, R	199, 25	0.5917	73.98%
1.3.3	0.9	1, 2	V, R	199, 25	0.6371	65.04%
1.4.1	0.3	1, 2	V, R	299, 40	0.2283	92.68%
1.4.2	0.6	1, 2	V, R	299, 40	0.3665	87.60%
1.4.3	0.9	1, 2	V, R	299, 40	0.6064	67.89%

Table 4: Results on the test set for t = 0.3 and different values of parameters (k,σ). Details of each setup are found in Table 3.

Setup Number	Accuracy in the test set of 228 images				
	Real images	Synthetic images			Total
		AI-17	AI-18	AI-19	
1.1.1	62/114	33/38	34/38	38/38	167/228
1.2.1	66/114	31/38	31/38	38/38	166/228
1.3.1	57/114	33/38	35/38	38/38	163/228
1.4.1	67/114	33/38	31/38	38/38	169/228

5.1. Experiments with parameters t, k, σ

Table 3 summarizes the experiments performed with threshold t, kernel size k, and standard deviation σ. The corresponding parameters and results in the validation set are listed. For each configuration, we obtain 1158 gazemaps from real images and 940 gazemaps from synthetic images (579 real and 470 synthetic for each of the two phases). Since they are not equal in number, we balance the dataset to have the same number of gazemaps for real and synthetic images, by randomly keeping only 940 gazemaps from real images. We also take into account gazemaps from both groups and phases. We see that the best performance on the validation set is observed for (t, k, σ) = (0.3, 299, 40). This is because larger values of these three parameters give larger critical regions, which allows the network to learn more information. However, the critical regions should not be too large, otherwise, they cover most of the image and the local information is lost.

In general, we observe that a lower threshold value leads to better results for all pairs (k, σ). Therefore, we examine the performance of the following experimental setups: 1.1.1, 1.2.1, 1.3.1, and 1.4.1 (all of which correspond to t = 0.3) on the test set of 228 images. We retrain the model using all the available gaze data for training (940 + 940 gazemaps) and obtain the results shown in Table 4. We observe that the overall recognition accuracy is slightly higher than that of the original base model. Specifically, for setup 1.4.1 we obtain an accuracy of 74.12%, which corresponds to a 4% increase in the recognition rate. Furthermore, the model has learned to recognize the synthetic images of all three GAN models (AI-17, AI-18, and AI-19) very well, but the recognition accuracy drops significantly on the set of real images; from 97.37% to only 58.77% for the best experimental setup 1.4.1.

Table 5: Experiments performed when training with gazemaps only from one phase or only from one group of subjects. The chosen parameters and the validation results are displayed for each setup.

Setup Number	Conditions				Validation	
	Phase	Group	Number of Gazemaps		Loss	Accuracy
			Train	Validation		
1.5.1	V	1, 2	347+347	123+123	0.3416	85.77%
1.5.2	R	1, 2	347+347	123+123	0.4514	83.33%
1.5.3	V, R	1	390+390	118+118	0.3230	88.14%
1.5.4	V, R	2	304+304	112+112	0.3423	86.61%

Table 6: Results on the test set for (t, k, σ) = (0.3, 299, 40), when training only with gazemaps of one group or one phase. Details of each setup are found in Table 5.

Setup Number	Accuracy in test set of 228 images				
	Real images	Synthetic images			Total
		AI-17	AI-18	AI-19	
1.5.1	72/114	31/38	34/38	38/38	175/228
1.5.2	61/114	33/38	36/38	38/38	168/228
1.5.3	79/114	29/38	29/38	38/38	175/228
1.5.4	65/114	32/38	35/38	38/38	170/228

5.2. Experiments with groups and phases

Due to the nature of the survey, members of Groups 1 and 2 had different recognition accuracies [13]. As a result, we trained the model with the gaze data of each group separately to see if any of them resulted in a greater improvement.

In addition, the regions on which observers focused in the first and second phases differed. The reason is that during the first phase, observers were simply looking at the image on their screen, whereas during the second phase, they were actively attempting to determine whether the image was real or synthetic. As such, we compare the results of the two phases by training with the gazemaps of each phase separately.

Table 5 reports the results for (t, k, σ) = (0.3, 299, 40). We observe that there are no large differences between the validation accuracies, although the gazemaps of Group 1 lead to slightly better results (setup 1.5.3). For each experimental setup of Table 5, we retrain the network using all available gazemaps and evaluate it on the test set of 228 images. The results reported in Table 6 show that the accuracies are comparable if we train with only one group or phase.

6. CONCLUSIONS

In this paper, we proposed a method to incorporate human gaze information into an automatic fake image detector. The CNN-based detector receives as input the spatial and cross-band co-occurrence matrices of the image. To integrate the gaze data in the model, we calculated the co-occurrence matrices only in the areas where the observers focused on (i.e., critical regions), and used these to further train the detector. The results did not show any significant improvement in the total accuracy when training with the gaze data. Specifically, the gaze-augmented model yields better performance on the synthetic images, which demonstrates that the critical regions hold some discriminative power. Yet, it performs worse at classifying real images. Thus, more investigation is needed.

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 *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[3] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi, "Cnn detection of gan-generated face images based on cross-band co-occurrences analysis," in *2020 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2020, pp. 1–6.

[4] Ali Borji, "Generated faces in the wild: Quantitative comparison of stable diffusion, MidJourney and DALL-E 2," 2022.

[5] 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)*, Los Alamitos, CA, USA, mar 2019, pp. 506–511, IEEE Computer Society.

[6] 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.

[7] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, B. Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Md Jawadul Bappy, and Amit Roy-Chowdhury, "Detecting GAN generated fake images using co-occurrence matrices," *Electronic Imaging*, vol. 2019, pp. 532–1, 01 2019.

[8] Federica Lago, Cecilia Pasquini, Rainer Bohme, Helene Dumont, Valerie Goffaux, and Giulia Boato, "More Real Than Real: A Study on Human Visual Perception of Synthetic Faces [Applications Corner]," *IEEE Signal Processing Magazine*, vol. 39, no. 1, pp. 109–116, Jan. 2022.

[9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," Feb. 2018, arXiv:1710.10196.

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and Improving the Image Quality of StyleGAN," Mar. 2020, arXiv:1912.04958.

[11] Sophie J Nightingale and Hany Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, pp. e2120481119, 2022.

[12] Nicholas Caporusso, Kelei Zhang, and Gordon Carlson, "Using eye-tracking to study the authenticity of images produced by generative adversarial networks," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020, pp. 1–6.

[13] Clara Riedmiller, "Gaze-driven discrimination of computer graphics from photo images," 2022, Eberhard Karls Universitaet Tuebingen, https://www.hci.uni-tuebingen.de/publications/riedmiller-thesis/.