Detecting CNN-Generated Facial Images in Real-World Scenarios

SUPPLEMENTARY MATERIAL

Nils Hulzebosch[1,2] Sarah Ibrahimi[1,2] Marcel Worring[1]
[1]University of Amsterdam [2]Dutch National Police

This supplementary material discusses more implementation details of our algorithmic pipeline (Section A), additional information on how the survey is constructed (Section B), and insights about image cues that participants use to recognize fake images (Section C).

A. Implementation Details

A.1. Creation of StyleGAN$_{CAHQ}$ dataset

We generate images using a model pre-trained on CAHQ images, because there is no public dataset of such images. For generation we make use of the truncation trick [19], which refers to the *stylistic* sampling radius (denoted by ψ) in the latent style vector. In other words, it refers to how much the style of the image to be generated should be similar to or divergent from the average style in the training data, where style refers to the characteristics of the full image, with a large focus on the person (*i.e.* facial area) in the image, and a minor focus on the background. In our initial experiments, this latent sampling radius is uniformly sampled from $[0, 1]$. However, the set of images with $\psi \approx 0$ appears to be very homogeneous and predictable, without much geometrical variation. On the other hand, using a large value (*i.e.* $\psi \approx 1$) results in original but unrealistic images with many artefacts. This is demonstrated in Figure 1. Both types of images do not represent real-world scenarios, where images are realistic and varied. Based on visual inspection of many images within the range of $\psi \in [0, 1]$, it seems that a good trade-off between quality and variety seems to be somewhere around $\psi \approx 0.5$. Thus, the dataset is generated using $\psi = 0.5$, where each image is generated by passing a random noise vector (*i.e.* no style transfer).

(a) Generated with $\psi \approx 0$.

(b) Generated with $\psi \approx 0.5$.

(c) Generated with $\psi \approx 1$.

Figure 1: Manually selected images generated by StyleGAN$_{CAHQ}$ with different quantities of the truncation trick. Note that this results in a trade-off between visually realistic (*i.e.* with $\psi \approx 0$) and original/varied images (*i.e.* with $\psi \approx 1$).

A.2. Training procedure

We train all models using the settings of [12], unless otherwise specified. We use a batch size of 64 for ForensicTransfer and 32 for Xception due to its higher memory demands. We evaluate two optimizers (SGD and Adam) and find that on average, SGD slightly outperforms Adam. Thus, we use SGD using a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. We stop training after 3 epochs of no improvement, as we observe that overfitting tends to be slightly higher when we use 30 epochs as done by [12]. All models are trained on a single Nvidia Titan V GPU and take roughly 1-3 hours of training time per model.

We evaluate the influence of a pre-trained Xception model on Imagenet in combination with pre-processing methods, and find that it performs worse with pre-trained weights. This is likely due to the large difference between images using for pre-training and our pre-processed images. Thus, we choose Xception to be trained from scratch, using weights randomly initialized from a normal distribution.

Lastly, we evaluate the influence of a random seed. Based on initial experiments, we observed some models and pre-processing methods to be unstable. For example, training with one random seed leads to a high test set accuracy,

while another random seed leads to a much lower accuracy on the same dataset. This effect is even stronger for cross-model or cross-data test sets. To minimize the influence of a random seed on the results, 5 instances of each model–pre-processing pair are trained, each initialized with another random seed. Then, the performance (*i.e.* accuracies, not predictions) is averaged over the 5 instances. Every score reported in the results section is therefore an average of 5 model instances.

B. Survey design

This section describes six important elements of the survey design, including 1) the selection of images, 2) the gathering of participants, 3) the setup for testing the influence of feedback, 4) the setup for testing the influence of image resolution, 5) the image-questions, and 6) the meta-questions.

First, participants get to see an instruction screen with a motivation, the goal, and details of the survey, along with the guiding definitions of fake and real in this survey, as shown in Figure 2. These definitions are required since *fake* is a vague definition and could also mean digitally edited (*i.e.* photoshopped, or morphed together). Then, participants judge 18 images, and answer several meta-questions, as discussed later. Lastly, there is an overview page where participants see their total score (N out of 18 correct), and each of the 18 images, along with their own answer and the correct answers. Lastly, some information about the research is provided.

Real image: taken with a camera, from a scene that really happened. Possibly post-processed, for example by adjusting colors.
Fake image: a non-existing scene that is fully created by a computer. In other words, the person in the image does not exist.

Figure 2: Provided definitions of real and fake images in the survey.

B.1. Selection of images

To achieve meaningful results, we use realistic and varied images. Therefore, we use real images from the FFHQ dataset, which is more varied and real-world than the CAHQ dataset. For fake images, we use the state-of-the-art StyleGAN$_{\boldsymbol{FFHQ}}$ images. Based on the findings of [51], along with our earlier experiments (Subsection A.1), we select images generated using the truncation with $\psi = 0.5$.

We *manually* select 1000 good StyleGAN$_{\boldsymbol{FFHQ}}$ images and exclude images with very obvious artefacts such as large blobs, because these images would disturb the results. As shown by [20], these blob-like artefacts are already vanished in newer versions of StyleGAN, and including them would not give an accurate representation of how these images would be used in real-world scenarios (where images with obvious artefacts would be excluded). Note that in the selected survey, there are still smaller artefacts and other cues present that could be detected if one knows where to pay attention to.

Next, 1000 real images are randomly selected from the FFHQ dataset, of which a handful of images of celebrities is manually removed to avoid bias towards real, and a handful of images that look really weird or obviously photo-shopped is manually removed to avoid bias towards fake. Furthermore, this helps preventing potential situations where participants who do not fully understand the definition of fake (*e.g.* thinking it means photoshopped) label a photoshopped image as fake. The resulting image pool consists of 1000 fake and 1000 real images, of which each participant sees 9 randomly selected images per class, in a random order.

B.2. Participants

In order to evaluate the detection capabilities of humans, a varied set of participants is tested. These participants vary in age, ethnicity, residence, education, AI-experience, *etc*. They are approached through several mediums such as Facebook, Instagram, email, Reddit, and WhatsApp. The survey is conducted during May 2019, and results in 591 participants. Of these participants, 496 completed the whole survey, while 95 terminated early, which could be at any point in the survey. The participants who terminated early are excluded from all results. Participants who have not answered meta-questions are only excluded from results where that specific meta-question is relevant (*e.g.* AI-experience). The amount of participants for different groups are shown in Table 1. As shown, the distribution of AI-experience (little or much) within the control group and feedback group is roughly equal.

B.3. Intermediate feedback

To evaluate whether participants are able to learn how to detect this type of fake images, two groups are constructed, to which respondents were randomly assigned. The first group is the control group and receives no intermediate feedback. Participants only get to see their results at the very end of the survey. The second group receives immediate feedback after labelling an image. This feedback is of the form *Correct, the image was indeed [real/fake]* or *Incorrect, the image was [real/fake]* and is shown above an image. Note that an image remains displayed in order to encourage people to see *why* an image is real or fake, without giving specific instructions on how to recognize fake images.

Participant group	Amount of participants	
Started survey	591	-
Completed survey	496	100%
Control-group *	263	53.0%
Feedback-group *	233	47.0%
Filled in 'AI-experience'	477	96.2%
Little AI-experience †	218	45.7%
Much AI-experience †	259	54.3%
Control-group - Little AI-exp. †	117	24.5%
Control-group - Much AI-exp. †	136	28.5%
Feedback-group - Little AI-exp. †	101	21.2%
Feedback-group - Much AI-exp. †	123	25.8%
Filled in 'image cues'	481	97.0%

Table 1: Overview of participant amounts per group. * randomly assigned, thus not precisely balanced. † calculated as part of people who filled in 'AI-experience' (477).

B.4. Image resolution

To evaluate whether image resolution influences the detection performance, three resolutions are evaluated: 256x256, 512x512, and 1024x1024 (the original size). They are resized using the standard interpolation method in web browsers. Each of these image sizes is tested with 3 real and 3 fake images, randomly chosen from the image pool, resulting in 18 images. Note that the random selection is without replacement, such that one participant cannot see the same image twice.

B.5. Labelling images

Each participant sees 18 images sequentially and answers on a 5-point scale how certain it is that an image is real or fake. The answers are the following: *certainly fake*, *probably fake*, *I don't know*, *probably real*, *certainly real*. Note that in the results, an answer is marked as correct if it is either the corresponding *probably [real/fake]* or *certainly [real/fake]* answer, and incorrect for the other three answers. A screenshot of our survey displayed in a web browser is shown in Figure 3.

There exists a website[1] where people can distinguish fake from real. On this website, a real and fake image are displayed next to each other, and users must select the one that is real. Such a setup is not appropriate for our survey, since we want to approximate real-world scenarios (*e.g.* a social media timeline or forensic applications) where one would make a choice (consciously or unconsciously) based on *one* image, and not a pair of images. Thus, we use an experimental setup with single images.

[1] http://www.whichfaceisreal.com/

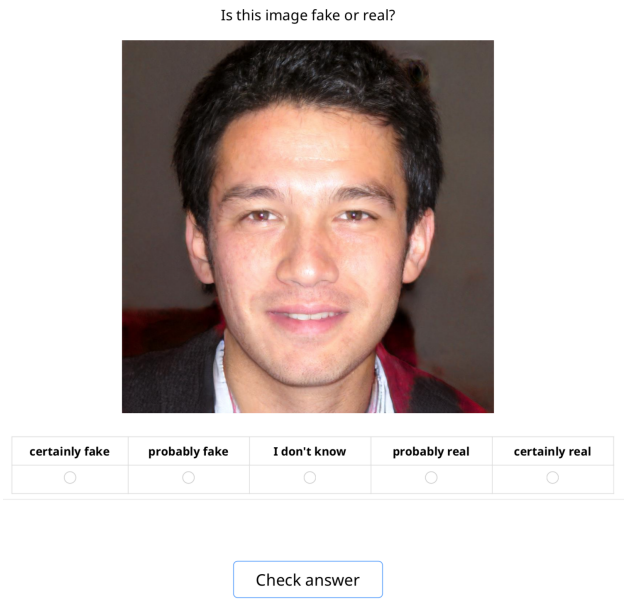

Figure 3: Screenshot taken from the online survey for a random fake image. Note that "Check answer" is only visible for participants from the feedback group, for the control group participants see a "Next" button. This image is generated by StyleGAN$_{\boldsymbol{FFHQ}}$.

B.6. Meta-questions

After labelling all images, participants have the choice to answer several meta-questions. Note that these questions are posed after the experiment itself to prevent any biases, and are not mandatory such that people can still finish the survey when they do not want to answer these questions. Most important are the questions about their AI-experience and cues they use to label images.

In order to evaluate the impact of domain knowledge, the amount of AI-experience is questioned using a 5-point scale, with the following answers: *0 - none*, *1 - heard of it*, *2 - indirect experience*, *3 - AI study*, and *4 - AI-professional (PhD or work)*. We expect that this gives more meaningful results than having the answers *little* and *much*, since these answers might be too subjective for the participants. Based on their answers, we choose to group the first three into *little* and the last two into *much* AI-experience, where *much* refers to AI-experts and *little* refers to everyday people.

In order to find out how humans can distinguish fake from real, respondents are posed the question: *You have labeled 18 images on a scale from fake to real. What aspects in the images contributed to your decisions?* The choice for an open instead of closed question is simple: it is not desirable to bias the respondents towards certain answers. For example, if a list of *eyes, nose, hair,* etc. would be presented, they would easily reason further with that list in

Object cue	Percentage
Background	26.6
Hair	12.3
Teeth	8.7
(A)symmetry	8.5
Eyes	7.7
Composition	7.3
Accessories / Context	7.3
Ears	6.1
'Other'	6.1
Expression	5.4
(Im)perfections	5.0
Skin	4.4
Originality	2.4
Mouth	2.4

Table 2: Object view image cues, ordered from most to least occurring. Percentage refers to how often the cue is mentioned as part of total amount of participants.

mind, resulting in, for example, a user input of *mouth, teeth*. However, if the list would be too broad, such as *eyes, nose, background, lighting conditions,* etc., the respondent might select multiple aspects without having actually thought of them during the experiment, resulting in biased backwards reasoning. The choice for an open question leads to a varied set of answers, as discussed in Section C.

C. Image cues

This final section discusses the image cues participants use to label an image as real or fake, based on their own answers after labelling all images. It becomes clear that the answers are very varied, ranging from specific answers such as *blurry eyes* to more abstract answers such as *something with the teeth* or *unoriginal*.

Based on all answers, we decide to group them into two categories. First, there are *object* cues, referring to *physical* properties of the objects and background in the images. A few examples of such cues are *weird shape of nose*, *something with eye*, *originality of background*, and *expression*. The second category is referred to as *display* cues, referring to *how* these objects are displayed in an image as if they were captured by a camera. Several examples include *artefacts*, *blurry nose*, and *lighting/shadows*. Clustering each of these cues is extremely difficult due to differences in jargon and specificity. Thus, our results should be taken with caution, since they approximate the distribution of image cues used by humans. Furthermore, some participants only answer with one example, while some answer with six examples, making this categorization even more difficult.

Display cue	Percentage
Blur	40.1
Artefacts	27.4
Transitions	10.5
Lighting / Shadow	9.3
Reflections	4.8
Details	4.0
Color	2.4
Focus / Depth of field	2.2
'Other'	1.6

Table 3: Display view image cues, ordered from most to least occurring. Percentage refers to how often the cue is mentioned as part of total amount of participants.

The results of our clustering are shown in Table Table 2 ('object cues') and Table Table 3 ('display cues'). Lastly, we provide one visual example (Figure 4) to refer to several of the cues shown in these tables.

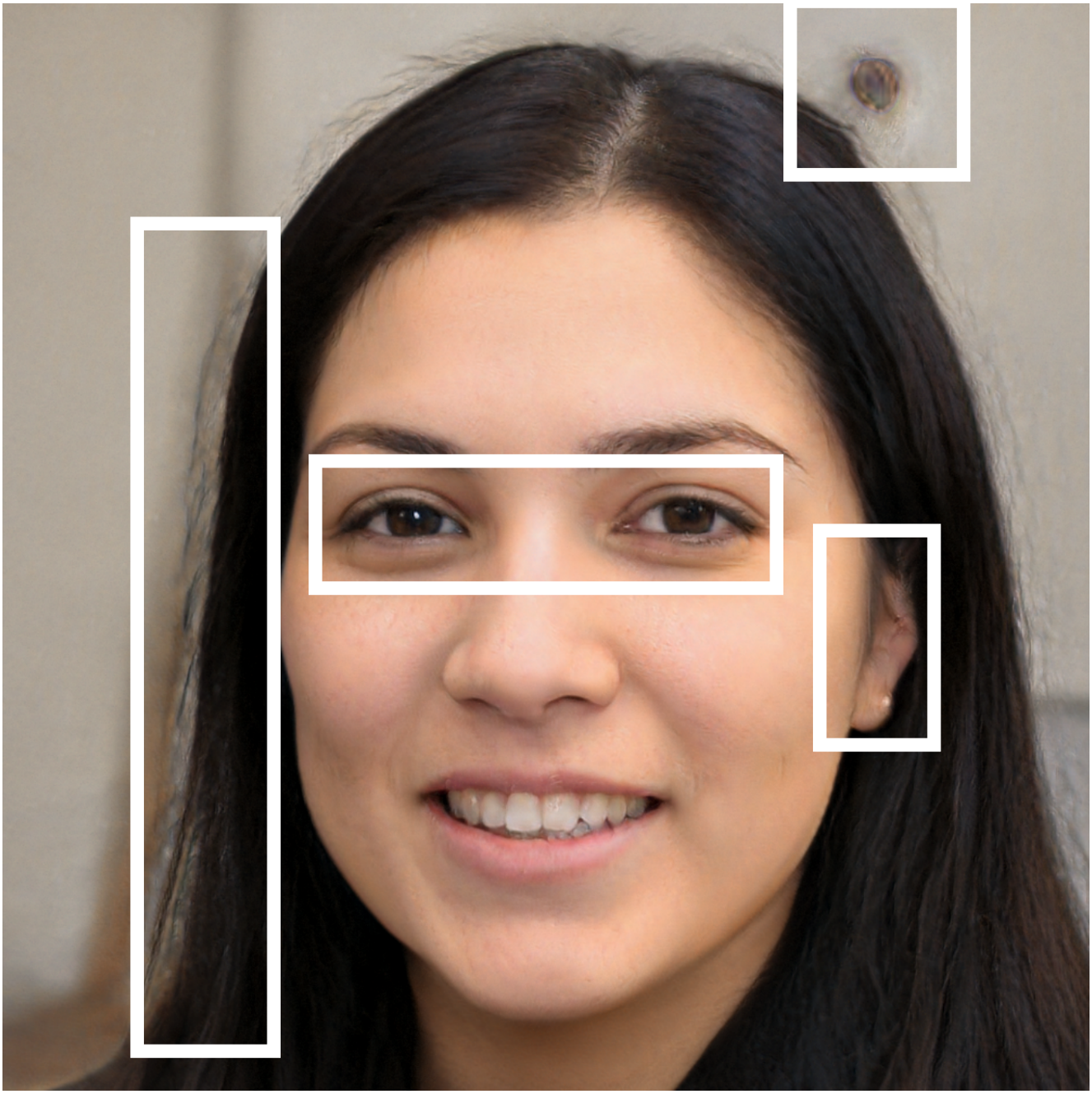

Figure 4: StyleGAN$_{\boldsymbol{FFHQ}}$ image with several unrealistic cues: 1) unnatural artefact (top-right), 2) blurry ear (right), 3) unrealistic/blurry hair (left), 4) asymmetric eyes (center). To elaborate on the last aspect: the iris colors, sizes, and shapes are slightly different between left and right eye. Furthermore, the pupil reflection only occurs at the left eye. When zooming in, artefacts (or lack of details) are better visible.