Supplemental Material for Deepfake Caricatures: Using Artifact Amplification to Expose Doctoring

Anonymous ICCV submission

Paper ID 9040

1. Human Experiments

1.1. Detection experiment

Task design The goal of this experiment was to determine which fake videos are relatively difficult to detect as fake, without scrutiny. The difficult videos would serve as good candidates to test the efficacy of our caricature method.

In our task, participants were instructed to maintain their fixation at a center cross while a pair of videos were presented, one on the left and one on the right (Figure 1). Participants were ask to select which video, left or right, was more likely to be fake by pressing a left or right keyboard key.

At the start of the experiment, people had the possibility to adjust the the distance between the videos (left and right videos were always at the same distance from the center). To help with the adjustment, the following instructions were given "The videos should not be too close to disturb you from fixating at the center, and not too far away to notably reduce your ability to distinguish them. Find the distance that feels more comfortable to you." After every 10 videos shown to users, users took a quick break and then had 3 seconds to refocus their attention on the fixation cross before examining the next set of videos. Whenever a gaze shifted towards either side of the videos, participants were asked to report and the trial was discarded.

Stimuli set We randomly selected 2000 videos (1000 real, 1000 fake) from the Deepfake Detection Challenge Dataset [1] and used FaceNet [2] to crop the face regions, resulting in videos of 360 $\times$ 360 pixels. 1000 video pairs were divided into five sessions, and were presented in a randomized order. One task consisted of 100 video pairs divided in 20 different levels.

Results We collected data from 7 participants from the university, resulting in 7 labels for each video on average. At the time the study was run, we were not able to access the eyetracker at our institution due to Covid-19 pandemic,

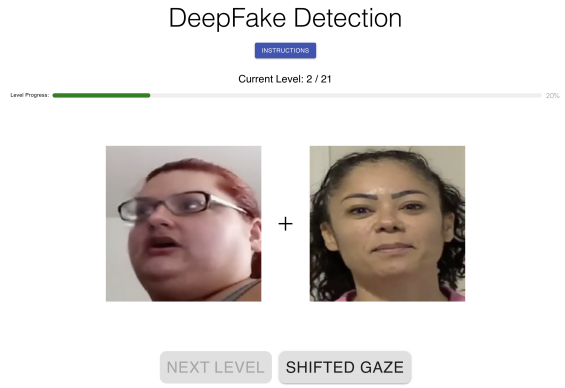

Figure 1. Interface to determine which of two videos is fake. Data collected from this experiment was used to identify deepfakes that were hard for humans to detect.

and asked the participants to self-report a trial if they could not maintain fixation. The average detection rate of fake videos is 0.79 ± 0.03. The average response time for each video pair was 2.41 seconds with a standard deviation of 0.81 seconds.

1.2. Caricatures Human Detection Results: Complementary analyses

Table **??** presents complementary analyses to Table 3 in the main paper. Here, we reorganize the presentation of the results according to the signal detection theory, focusing on the fake videos. As shown in Table **??**, the Grad-CAM caricatures have the highest HIT score (**71**% correct responses on the fake videos, which is significantly different from all the other conditions, with t(199)¿5, p¡.0001 for all comparison Grad-CAM with another condition). Grad-CAM caricatures elicited **26**% of false alarms (FA, responding *fake* on a real video) and a d prime (measure of sensitivity) of **1.2**. Importantly, the decision criterion *C* near zero indicates that in the Grad-CAM caricature condition, participants do not have a bias response towards fake or real, which is the best scenario for scaling up the method. As expected, this is in sharp contrast with the original video condition (no car-

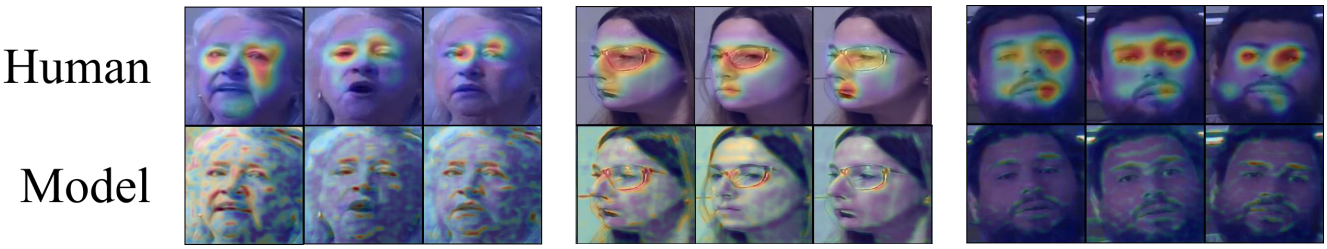

Figure 2. Examples of human artifact annotations and model artifact attention maps show that although the model maps bear some resemblance to human attention, the maps usually present more high frequency details. While humans focus on rather large areas around eyes and mouth, Our model identifies more fine grained areas of potential doctoring.

icature): people cannot discriminate well even when they can scrutinize the videos, which video is real or fake, and have a tendency to respond *real* more often. The results also emphasize the importance to have guided caricatures generation, like Grad-CAM, instead of a uniform artifact transformation, which creates caricatures that distress human perception (very low d prime of 0.5). Future work will concentrate on creating caricatures with a higher d-prime sensitivity, and no response bias.

2. Additional Models Results and Figures

We compare human annotations of fake indication areas with those generated by our model frame-by-frame (Figure 2). Humans tend to annotate areas of the face that are near the eyes or mouth, which are more salient facial features. Our model picks up on these features, and more fake indication areas. For example, there is attention to inflection regions of the face, such as the jawline or near the hairline.

2.1. Caricatures in Video Form

Our effect is best experienced in video form. We provide a compilation of examples of original and caricature videos in the supplementary mp4 files *caricature_applied_on_real_vids.mp4* and *caricature_applied_on_fake_vids.mp4*. The first one corresponds to the effect of applying our caricature model to real videos. Little to no impact on the underlying video can be seen. On the second movie, we show the caricature impact on deepfakes. The artifacts are amplified and the video is generally distorted. We show our best Grad-CAM caricatures in both cases.

We additionally provide a web gallery, *gallery.html*, that allows for easy visualization of several additional examples. Once the gallery.html file is extracted along with its accompanying videos, double click on the html file to view it on your browser.

References

[1] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 1

[2] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1