Ding *et al. EURASIP Journal on Information Security* (2020) 2020:6
https://doi.org/10.1186/s13635-020-00109-8

EURASIP Journal on Information Security

RESEARCH Open Access

Swapped face detection using deep learning and subjective assessment

Xinyi Ding[1*], Zohreh Raziei[2], Eric C. Larson[1], Eli V. Olinick[2], Paul Krueger[3] and Michael Hahsler[2]

Abstract

The tremendous success of deep learning for imaging applications has resulted in numerous beneficial advances. Unfortunately, this success has also been a catalyst for malicious uses such as photo-realistic face swapping of parties without consent. In this study, we use deep transfer learning for face swapping detection, showing true positive rates greater than 96% with very few false alarms. Distinguished from existing methods that only provide detection accuracy, we also provide uncertainty for each prediction, which is critical for trust in the deployment of such detection systems. Moreover, we provide a comparison to human subjects. To capture human recognition performance, we build a website to collect pairwise comparisons of images from human subjects. Based on these comparisons, we infer a consensus ranking from the image perceived as most real to the image perceived as most fake. Overall, the results show the effectiveness of our method. As part of this study, we create a novel dataset that is, to the best of our knowledge, the largest swapped face dataset created using still images. This dataset will be available for academic research use per request. Our goal of this study is to inspire more research in the field of image forensics through the creation of a dataset and initial analysis.

Keywords: Face swapping, Deep learning, Image forensics, Privacy

1 Introduction

Face swapping refers to the process of transferring one person's face from a source image to another person in a target image, while maintaining photo-realism. It has a number of applications in cinematic entertainment and gaming. However, in the wrong hands, this method could also be used for fraudulent or malicious purposes. For example, "DeepFake" is such a project that uses generative adversarial networks (GANs) [1] to produce videos in which people are saying or performing actions that never occurred. While some uses without consent might seem benign such as placing Nicolas Cage in classic movie scenes, many sinister purposes have already occurred. For example, a malicious use of this technology involved a number of attackers creating pornographic or otherwise sexually compromising videos of celebrities using face swapping [2]. A detection system could have prevented this type of harassment before it caused any public harm.

Conventional ways of conducting face swapping usually involve several steps. A face detector is first applied to narrow down the facial region of interest (ROI). Then, the head position and facial landmarks are used to build a perspective model. To fit the source image into the target ROI, some adjustments need to be taken. Typically, these adjustments are specific to a given algorithm. Finally, a blending happens that fuses the source face into the target area. This process has historically involved a number of mature techniques and careful design, especially if the source and target faces have dramatically different position and angles (the resulting image may not have a natural look).

The impressive progress deep learning has made in recent years is changing how face swapping techniques are applied from at least two perspectives. Firstly, models like convolutional neural networks allow more accurate face landmarks detection, segmentation, and pose estimation.

*Correspondence: xding@mail.smu.edu
[1]Department of Computer Science, Southern Methodist University, Dallas, USA
Full list of author information is available at the end of the article

Secondly, generative models like GANs [1] combined with other techniques like auto-encoding [3] allow automation of facial expression transformation and blending, making large-scale, automated face swapping possible. Individuals that use these techniques require little training to achieve photo-realistic results. In this study, we use two methods to generate swapped faces [4, 5]. Both methods exploit the advantages of deep learning methods using contrasting approaches, discussed further in the next section. We use this dataset of swapped faces to evaluate and inform the design of a face swap detection classifier.

With enough data, deep learning-based classifiers can typically achieve low bias due to their ability to represent complex data transformations. However, in many cases, the confidence levels of these predictions are also important, especially when critical decisions need to be made based on these predictions. The uncertainty of a prediction could indicate when other methods could be more reliable. Bayesian deep learning, for example, assumes a prior distribution of its parameters $P(\mathbf{w})$ and integrates the posterior distribution $P(\mathbf{w}|\mathcal{D})$ when making a prediction, given the dataset $\mathcal{D}$. However, it is usually intractable for models like neural networks and must be employed using approximations to judge uncertainty. Test-time data augmentation is another recently proposed technique to measure data-dependent uncertainty [6]. We propose a much simpler approach by using the raw score difference of the neural network outputs. For a binary classification task, a neural network will usually output two numbers (scores) representing two classes. The input will be assigned the class the larger number represents. We assume, in a binary classification task, if the model has low confidence about a prediction, the difference of the two scores should be small compared with a high confidence prediction. We also show that the score difference of the neural network is correlated with the human perception of "fake" versus "real."

The end goal of malicious face swapping is to fool a human observer into believing that a person they recognize is depicted in the image. Therefore, it is important to understand how human subjects perform in recognizing swapped faces. To this end, we not only estimate the accuracy of human subjects in detecting fake faces, but we also elicit a consensus ranking of these images from the image perceived as most real to the image perceived as most fake using pairwise comparisons by a set of human subjects. We selected 400 images and designed a custom website to collect human pairwise comparisons of images. Approximate ranking is used [7] to help reduce the number of needed pairwise comparisons. With this ranking, we compare the score margin of our model outputs to the consensus ranking from human subjects, showing good, but not perfect correspondence. We believe future work can improve on this ranking comparison, providing a means to evaluate face swapping detection techniques that more realistically follow human intuition.

We enumerate our contributions as follows:

- We provide a dataset comprising 420,053 images derived from pictures of 86 celebrities. This dataset is created using still images, different from other datasets created using video frames that may contain highly correlated images. In this dataset, each celebrity has approximately 1000 original images (more than any other celebrity dataset). The main use case is that the human observer recognizes the person in the image. Therefore, celebrities are a good fit. Also, we believe our dataset is not only useful for swapped face detection, it may also be beneficial for developing facial models.
- We investigate the performance of two representative face swapping techniques and discuss limitations of each approach. For each technique, we create thousands of swapped faces for a number of celebrity images.
- We build a deep learning model using transfer learning for detecting swapped faces. To our best knowledge, it is the first model that provides high accuracy predictions coupled with an analysis of uncertainties.
- We build a website that collects pairwise comparisons from human subjects in order to rank images from most real to most fake. Based on these comparisons, we approximately rank these images and compare to our model.

2 Related work

There are numerous existing works that target face manipulation and detection. Strictly speaking, face swapping is simply one particular kind of image tampering. Detection techniques designed for general image tampering may or may not work on swapped faces, but we expect specially designed techniques to perform superior to generic methods. Thus, we only discuss related works that directly target or involve face swapping and its detection.

2.1 Face swapping

Blanz et al. [8] use an algorithm that estimates a 3D textured model of a face from one image, applying a new facial "texture" to the estimated 3D model. The estimations also include relevant parameters of the scene, such as the orientation in 3D, the camera's focal length, position, and illumination intensity and direction. The algorithm resembles the Morphable Model such that it optimizes all parameters in the model in conversion from 3D to image. Bitouk et al. [9] bring the idea of face replacement without the use of 3D reconstruction techniques. The approach

involves the finding of a candidate replacement face which has similar appearance attributes to an input face. It is, therefore, necessary to create a large library of images. A ranking algorithm is then used in selecting the image to be replaced from the library. To make the swapped face more realistic, lighting and color properties of the candidate images might be adjusted. Their system is able to create subjectively realistic swapped faces. However, one of the biggest limitations is that it is unable to swap an arbitrary pair of faces. Mahajan et al. [10] present an algorithm that automatically chooses faces that are facing the front and then replaces them with stock faces in a similar fashion as Bitouk et al. [9].

Chen et al. [11] suggested an algorithm that can be used in the replacement of faces in referenced images that have common features and shape as the input face. A triangulation-based algorithm is used in warping the image by adjusting the reference face and its accompanying background to the input face. A parsing algorithm is used in accurate detection of face ROIs, and then, the Poisson image editing algorithm is finally used in the realization of boundaries and color correction. Poisson editing is explored from its basics by Perez et al. [12]. Once given methods to craft a Laplacian over some domain for an unknown function, a numerical solution of the Poisson equation for seamless domain filling is calculated. This technique can independently be replicated in color image channels.

The empirical success of deep learning in image processing has also resulted in many new face swapping techniques. Korshunova et al. [13] approached face swapping as a style transfer task. They consider pose and facial expression as the content and identity as the style. A convolutional neural network with multi-scale branches working on different resolutions of the image is used for transformation. Before and after the transformation, face alignment is conducted using the facial keypoints. Nirkin et al. [4] proposed a system that allows face swapping in more challenging conditions (two faces may have very different pose and angle). They applied a multitude of techniques to capture facial landmarks for both the source image and the target image, building 3D face models that allow swapping to occur via transformations. A fully convolutional neural network (FCN) is used for segmentation and for blending technique after transformation.

The popularity of auto-encoder[3] and generative adversarial networks (GANs) [1] makes face swapping more automated, requiring less expert supervision. A variant of the DeepFake project is based on these two techniques [5]. The input and output of an auto-encoder is fixed, and a joint latent space is discovered. During training, one uses this latent space to recover the original image of two (or more) individuals. Two different auto-encoders are trained on two different people, sharing the same encoder so that the latent space is learned jointly. This training incentivizes the encoder to capture some common properties of the faces (such as pose and relative expression). The decoders, on the other hand, are separate for each individual so that they can learn to generate realistic images of a given person from the latent space. Face swapping happens when one encodes person A's face, but then uses person B's decoder to construct a face from the latent space. A variant of this method in [5] uses an auto-encoder as a generator and a CNN as the discriminator that checks if the face is real or swapped. Empirical results show that adding this adversarial loss improves the quality of swapped faces.

Natsume et al. [14] suggest an approach that uses hair and faces in the swapping and replacement of faces in the latent space. The approach applies a generative neural network referred to as an RS-GAN (region-separative generative adversarial network) in the generation of a single face-swapped image. Dale et al. [15] bring in the concept of face replacement in a video setting rather than in an image. In their work, they use a simple acquisition process in the replacement of faces in a video using inexpensive hardware and less human intervention.

2.2 Fake face detection

Zhang et al. [16] created swapped faces using the labeled faces in the wild (LFW) dataset [17]. They used speeded up robust features (SURF) [18] and Bag of Words (BoW) to create image features instead of using raw pixels. After that, they tested on different machine learning models like random forests, SVMs, and simple neural networks. They were able to achieve accuracy over 92%, but did not investigate beyond their proprietary swapping techniques. The quality of their swapped faces is not compared to other datasets. Moreover, their dataset only has 10,000 images (half swapped) which is relatively small compared to other works.

Khodabakhsh et al. [19] examined the generalization ability of previously published methods. They collected a new dataset containing 53,000 images from 150 videos. The swapped faces in their dataset were generated using different techniques. Both texture-based and CNN-based fake face detection were evaluated. Smoothing and blending were used to make the swapped face more photo-realistic. However, the use of video frames increases the similarity of images, therefore decreasing the variety of images. Agarwal et al. [20] proposed a feature encoding method called Weighted Local Magnitude Patterns. They targeted videos instead of still images. They also created their own dataset. Korshunov et al. also targeted swapped faces detection in video [21]. They evaluated several detection methods of DeepFakes. What is more, they analyze the vulnerability of VGG- and FaceNet-based face recognition systems.

Table 1 Dataset statistics

	Nirkin's method [4]	AE-GAN [5]	Total
Real face	72,502	84,428	156,930
Swapped face	178,695	84,428	263,123
Total	251,197	168,856	420,053

A recent work from Rössler et al. [22] provides an evaluation of various detectors in different scenarios. They also report human performance on these manipulated images as a baseline. Our work shares many similarities with these works. The main difference is that we provide a large-scale dataset created using still images instead of videos, avoiding image similarity issues. Moreover, we provide around 1000 different images in the wild for each celebrity. This is useful for models like auto-encoders that require numerous images for proper training. In this aspect, our dataset could be used beyond fake face detection. The second difference is that we are not only providing accuracy from human subjects, but also providing the rankings of images from most real to most fake. We compare this ranking to the score margin ranking of our classifier showing that human certainty and classifier certainty are relatively (but not identically) correlated.

3 Experiment

3.1 Dataset

Face swapping methods based on auto-encoding typically require numerous images from the same identity (usually several hundreds). There was no such dataset that met this requirement when we conducted this study; thus, we decided to create our own. Access to version 1.0 of this dataset is available for academic research use per request at the noted link[1]. The statistics of our dataset are shown in Table 1.

All our celebrity images are downloaded using the Google image API. After downloading these images, we run scripts to remove images without visible faces and remove duplicate images. Then, we perform cropping to remove extra backgrounds. Cropping was performed automatically and inspected visually for consistency. We created two types of cropped images as shown in Fig. 1 (Left). One method for face swapping we employed (Nirkin's method [4]) involves face detection and lighting detection, allowing the use of images with larger, more varied backgrounds. On the other hand, the Auto-Encoder-GAN (AE-GAN) [5] method is more sensitive to the background; thus, we eliminate as much background as possible. In a real deployment of such a method, a face detection would be run first to obtain a region of interest, then swapping would be performed within the region. In this study, for convenience, we crop the face first using a face detector as a pre-processing step.

The two face swapping techniques we use in this study are representatives of many algorithms in use today. Nirkin's method [4] is a pipeline of several individual techniques. On the other hand, the Auto-Encoder-GAN (AE-GAN) method is completely automatic, using a fully convolutional neural network architecture [5]. In selecting individuals to swap, we randomly pair celebrities within the same sex and skin tone. Each celebrity has around 1000 original images. For Nirkin's method, once a pair of celebrities is chosen, we randomly choose one image from these 1000 images as the source image and randomly choose one from the other celebrity as the target image. We noticed, for Nirkin's method, when the lighting conditions or head pose of two images differs too dramatically, the resulted swapped face is of low quality. On the other hand, the quality of swapped faces from the AE-GAN method is more consistent.

3.2 Classifier

Existing swapped face detection systems based on deep learning only provide an accuracy metric, which is insufficient for a classifier that is used continuously for detection. Providing an uncertainty level for each prediction is important for the deployment of such systems, especially when critical decisions need to be made based on these predictions.

In this study, we use the class probability produced by the softmax function for the final layer of the network for classification. We predict the class with the larger probability. For the binary classification problem, the winning class will have a probability of greater or equal to .5. Note that this is equivalent to choosing the class with the larger raw score (before applying softmax) and the raw score margin (i.e., the difference between the two classes) is proportional to the predicted class probability. For convenience, we will report score margins in this paper. We assume if the model is less certain about a prediction, the difference of these two scores should be smaller than that of a more certain prediction. We note that this method is extremely simple as compared to other models that explicitly try to model uncertainty of the neural network, such as Bayesian deep learning methods. The score margin, on the other hand, does not explicitly account for model uncertainty of the neural network—especially when images are fed into the network that are highly different from images from the training data. Even so, we find that the score margin is an effective measure of uncertainty for our dataset, though more explicit uncertainty models are warranted for future research.

Deep learning methods usually take days or weeks to train. Models like ResNet [23] can easily have tens or hundreds of layers. It is believed with more layers, more

[1] https://www.dropbox.com/sh/rq9kcsg3kope235/AABOJGxV6ZsI4-4bmwMGqtgia?dl=0

Fig. 1 Real and swapped faces in our dataset. Top row right: Auto-Encoder-GAN. Bottom row right: Nirkin's method

accurate hierarchical representation could be learned. Transfer learning allows us to reuse the learned parameters from one task to another similar task, thus avoiding training from scratch, which can save a tremendous amount of resources. In this study, we apply transfer learning using a ResNet model with 18 layers or ResNet-18, as described in the original paper [23], which is originally trained to perform object recognition on ImageNet [24]. We use the convolutional layers from ResNet-18 to process our images into a set of low-dimensional features (typically called the latent representation of the image). We note that no resizing of the images is needed since we use only the convolutional filters of the ResNet-18 architecture. These filters are used and can be applied to images of any size, which only changes the size of the outputs of the convolutions. Since we are performing binary classification in this study, we replace the final layers of ResNet-18 with custom dense layers and then train the model in stages. More specifically, we throw out the non-convolutional layers of ResNet-18, opting to train new layers to interpret the latent space feature representation of the images. During the first stage, we constrain the ResNet-18 architecture to be constant while the newly added layers are trained. After sufficient epochs, we then "fine tune" the ResNet-18 architecture, allowing the weights to be trained via back-propagation for a number of epochs. This method for implementing transfer learning on neural networks is also summarized by Shin et al. [25].

3.3 Human subjects

Because face swapping attacks are typically aimed at misleading observers, it is important to understand how human beings perform at detecting swapped faces. In this research, it is not only our aim to provide the accuracy of human subjects at detecting swapped faces, but also to establish a ranking of images from perceived as most real to most fake. For example, if a rater thinks that an image is fake, is it obvious or is that rater not quite sure about their decision? We argue that this uncertainty is important to model. Moreover, we argue that if fake images are visually unidentifiable, then the human ranking should be similar to the ranking produced by using the predicted score of the machine learning model.

For n images, we have $\frac{n(n-1)}{2}$ pairs, and to get reliable rating data, we would have to elicit for each pair multiple ratings by different raters. This is impractical for even a modest number if images. Therefore, we apply two techniques to mitigate the ranking burden. First, we only rank a subset of the total images, and second, we perform approximate ranking of image pairs. As a subset of images, we manually select 100 high-quality swapped faces from each method together with 200 real faces (400 images in total). The manual selection of high-quality images is justified because badly swapped faces would be easily recognized. Thus, an attacker would likely perform the same manner of "re-selecting" only high-quality images before using them for a malicious purpose. It is of note that even with only 400 images and only considering a single rating per pair, the number of pairwise ratings required for ranking (over 79,000) poses a monumental task. The second technique is to use an active learning scheme that dynamically selects the most informative next image pair to compare for converging to an approximate ranking.

3.3.1 *Active scheme for approximate ranking*

To elicit the ranking, we designed and deployed a website that implements an adaptation of the approximate ranking algorithm Hamming-LUCB [7] to the specific application of swapped face detection. The input and output of the algorithm are based on "crowd sourced" comparisons made on our website by a large number of users. Users on the website are asked to compare two images and select which image they think is more likely to be fake. The idea is that different users are more likely to select a clearly fake looking image, while they will disagree more when the two images both look real. This is exactly the type of information that is needed to create the consensus ranking. We use approximate ranking with active comparison selection because of the impractical number of pairwise

comparisons that would be required for an exact ranking using passive comparison selection.

For a set of $[n] = \{1, 2, \ldots, n\}$ of n images, Hamming-LUCB seeks to estimate an approximate ranking by assigning each image $i \in [n]$ a ranking score τ_i that is sufficiently accurate to identify two ordered sets of images S_1 and S_2, representing the highest and lowest ranked images, respectively. Subsets S_1 and S_2 consist of a range of items of size $k - h$, where h is a predefined number of allowable "mistakes" in each set, and the $n - k - h$ items. Images in S_1 are perceived as most fake, while images in S_2 are perceived as least fake. Between the two sets, there is a high confidence that the items contained in the first set score higher as compared to those items that are contained in the second set.

The main goal of the algorithm is collect enough information to estimate the image ranking score for each image to be able to identify the two sets, S_1 and S_2, given a required confidence. In each iteration, the algorithm determines which pair of items to present for comparison based on the outcomes of previous comparisons. In this strategy, the current score estimations and the intervals of confidence associated with the scores are the parameters underlying the decision about which images to compare next. The result is an approximate ranking that is most accurate for images that are on the borderline between S_1 and S_2, i.e., images that do not look too fake, but also not too real. For more details about the algorithm and its implementation, we refer the reader to [7].

3.3.2 Website ratings collection

The inspiration of our website comes from that of the GIFGIF project for ranking emotions of GIFs[2]. Figure 2 shows a screenshot of the website. The text "Which of the following two faces looks MORE FAKE to you" is displayed above two images. When the evaluator moves the mouse above either image, it is highlighted with a box. The evaluator could choose to login using a registered account or stay as an anonymous evaluator. In this website, there are two instances of Hamming-LUCB running independently for two investigated face swapping algorithms, AE-GAN and Nirkin's method. The probability of selecting either swapped type is 50%. Over a 3-month period, we recruited volunteers to rate the images. When a new rater is introduced to the website, they first undergo a tutorial and example rating to ensure they understand the selection process. We collected 36,112 comparisons in total from more than 90 evaluators who created login accounts on the system. We note that anyone using the system anonymously (without logging in) was not tracked, so we do not know exactly how many different evaluators used the website.

[2] http://gifgif.media.mit.edu

4 Results

To evaluate the performance of our classifier, we have separated train and test by person (i.e., celebrity). That is, if a person has images in the training set, they do not have images in the testing set. In this way, the algorithm cannot learn specifics of a person's face but, instead, must rely on learning generalizing artifacts within swapped face images. We used the default hyperparameters and a fixed number of epochs for models, and therefore, there was no need for a separate validation dataset. Also, we do not distinguish between the two types of swapped faces during training. In other words, we mix the swapped faces generated using both methods during training, but we report prediction performance on each method separately.

Table 2 gives the overall detection performance of our classifier for the entire dataset (5-fold cross-validation) and for the 400 images that were ranked by human subjects (using the entire dataset for training, but not including these 400 images). We also report the accuracy with which humans were able to select images as real or fake based on the pairwise ranking. That is, any fake images ranked in the top 50% or any real images ranked in the bottom 50% were considered as errors. From the table, we can see that both human subjects and the classifier achieve good accuracy when detecting swapped faces. Our classifier is able to achieve comparable results to human subjects in 200 manually selected representative images (100 fake, 100 real) for each method.

To test the generalizability of our model, we also use the Chicago Face Dataset[3] (CFD). It provides high-resolution standardized photographs of male and female faces of different ages and varying ethnicity. The CFD also provides photographs of faces of different expressions for a subset of targets (neutral, happy, fearful, etc). For these targets that have more than one expression, we only use the photograph of the neutral expression. We used 600 real faces from CFD and generated 662 swapped faces using Nirkin's method. Thus, we have 1262 faces in total as a separate testing set. It was not possible to generate swapped faces using AE-GAN method, as mentioned above, because AE-GAN requires hundreds of images for the same person (and is therefore only suitable for databases with numerous images of one person, like celebrities).

4.1 Classification accuracy

As we mentioned above, we created two types of cropped images for each method. The AE-GAN method contains minimal background, and Nirkin's method contains more background. We can see from Table 2 that our classifier is able to detect face swapping better for the AE-GAN generated images—this holds true regardless of testing upon the entire dataset or using the manually selected 200. As

[3] https://chicagofaces.org/default/

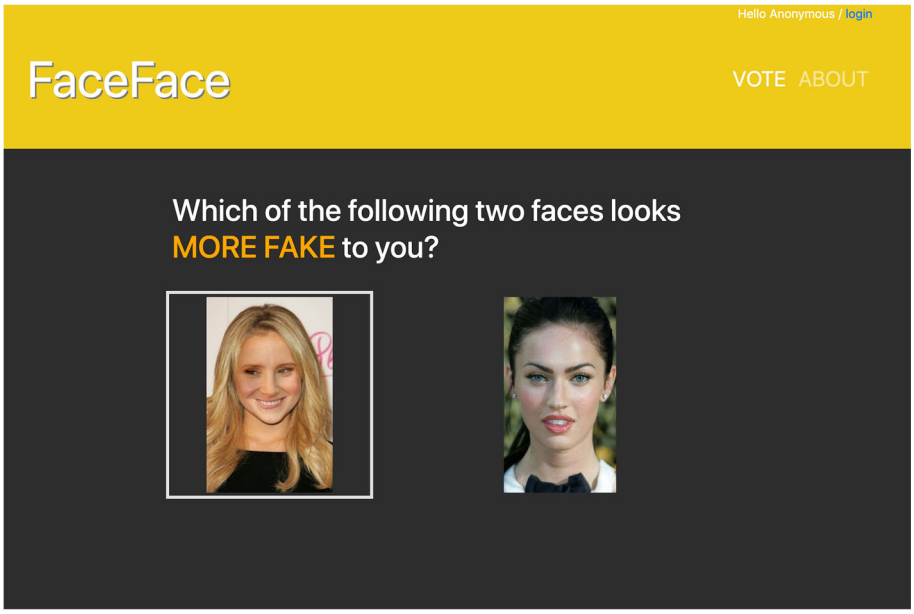

Fig. 2 A screenshot of the website collecting pairwise comparisons. As the mouse hovers over the left image, it is highlighted

we can see from Fig. 1 (Right), the AE-GAN generates swapped faces that are slightly blurry, which we believe our model exploits for detection. On the other hand, Nirkin's method could generate swapped faces without a decrease in sharpness. Thus, it may require the model to learn more subtle features, such as looking for changes in lighting condition near the cropped face or stretching of facial landmarks to align the perspectives.

When testing the Chicago Face Dataset, we use all the images in our celebrity dataset for training. Table 2 shows the results after a single epoch of training. As we can see, our model is generalizable to the non-celebrity faces in the Chicago Face Dataset, but performs slightly worse in terms of true positive rate and false alarm rate. However, we do not attribute this to celebrity status. Instead, we hypothesize that the reduced variability in lighting and relatively consistent head pose for the Chicago Face Dataset eases the process of swapping faces. That is, there is less inconsistency in the face for the model to find, resulting in reduced ability to detect swapped faces. Even so, the performance on the dataset is greater than 90.0%, which supports the conclusion that the model is robust to images from different sources, such as the Chicago Face Dataset.

For version 1.0 of the dataset, we have collected more than 36,112 pairwise comparisons from more than 90 evaluators (approximately evenly split between each method). Human subjects may have different opinions about a pair of images; thus, it requires many pairwise comparisons, especially for these images in the middle area. However, we can see human subjects still give a reasonable accuracy, especially for the AE-GAN method. It is interesting to see that both our classifier and human subjects perform better on the AE-GAN generated images.

4.2 Classifier visualization

To elucidate what spatial area our classifiers are concentrating upon to detect an image as real or fake, we employ the Gradient-weighted Class Activation Mapping (Grad-CAM) visualization technique [26]. This analysis helps mitigate the opaqueness of a neural network model and enhance explainability for applications in the domain of

Table 2 Overall results

		Nirkin's method [4]			AE-GAN [5]		
		True positive (%)	False positive (%)	Accuracy (%)	True positive (%)	False positive (%)	Accuracy (%)
Entire dataset	ResNet-18	96.52	0.60	97.19	99.86	0.08	99.88
Manually selected 200	ResNet-18	96.00	0.00	98.00	100.00	0.00	100.00
	Human subjects	92.00	8.00	92.00	98.00	2.00	98.00
Chicago Face Dataset	ResNet-18	90.18	6.03	91.97			

Fig. 3 Grad-CAM visualization of our proposed model on real and swapped faces. Top row: original images before fed into the network. Bottom row: original images with heatmap

privacy and security. Grad-CAM starts by calculating the gradients of the score for class c (before the softmax) with respect to the feature maps of the last convolutional layer. The gradients are then global average-pooled as weights. By inspecting these weighted activation maps, we can see which portions of the image have significant influence in classification. For both types of generated swapped faces, our classifier focuses on the central facial area (i.e., the nose and eyes) rather than the background. This is also the case for real faces as we can see from Fig. 3. We hypothesize that the classifier focuses on the nose and eyes because these areas contain more intricate details of faces. Figure 4 gives one such example. As we zoom in, we can see the blurring on the left side of the nose is

Fig. 4 An example of blurring difference contained on a nose from a swapped image

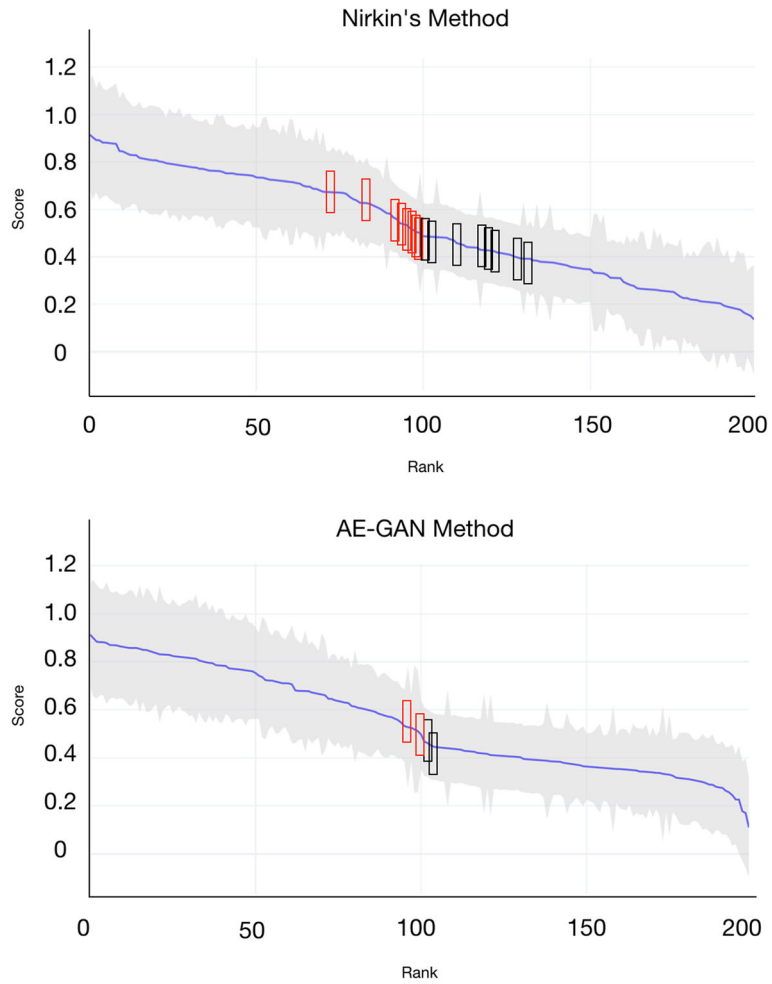

Fig. 5 Human subjects rank of the manually selected 200 images. Left to right, from most real to most fake

slightly different from that on the right side. It is interesting that the eyes and nose are focused upon by the classifier because human gaze also tends to focus on the eyes and nose when viewing faces [27].

4.3 Comparing image rankings

Rather than reporting only accuracy of detecting swapped faces from human subjects, we also compare the rankings. Ranking gives us more information to compare the models with humans, such as does the ResNet model similarly rate images that are difficult to rate for humans? Or, on the contrary, is the ranking from the model very different from human ranking?

Figure 5 gives the overall ranking for faces generated using two methods using the Hamming-LUCB consensus ranking from human evaluators. Red boxed points are false negatives, and black boxed points are false positives. The confidence interval based on the Hamming-LUCB is shown as the gray-shaded area. As we can see, human subjects have more difficulty classifying the faces generated using Nirkin's method. As mentioned, the AE-GAN generated faces are blurrier compared with Nirkin's method. Human subjects seemingly are able to learn such a pattern from previous experience. While some mistakes are present for the AE-GAN, these mistakes are very near the middle of the ranking. Swapped faces generated using Nirkin's method keep the original resolution and are more photo-realistic—thus, they are also more difficult to discern as fake.

To compare the human ranking to our model, we need to process the outputs of the neural network. During training, the model learns a representation of the input data using convolutions. Instances belonging to different classes usually are pushed away in a high-dimensional space. But this distance between two instances is not necessarily meaningful to interpret. Despite this, the output

of the activation function can be interpreted as a relative probability that the instance belongs to each class.

We assume for the score margin of the last fully connected layer (before the softmax activation) that the wider the margin, the more confident the classifier is that the instance is real or fake (i.e., a measure of certainty). Figure 6 gives the comparison of score margin of our model and human rating (score) for the 200 faces used with each method. For Nirkin's method, the Pearson correlation coefficient between the scores is 0.7896 and Spearman's rank order correlation coefficient between the resulting rankings is 0.7579. For the AE-GAN Method, the Pearson correlation is 0.8332 and Spearman's rank order correlation is 0.7576. However, the overall correlation may only indicate that these two classes are well separated by humans and the classifier. To eliminate this effect, we also give the correlation within each class to understand if the correlation is related to human perception or "realism." For the AE-GAN method, the Pearson correlation coefficient is 0.2865 for the real faces and 0.0415 for the fake faces. Spearman's rank order correlation is 0.1106 for the real faces and −0.0027 for the fake faces. For Nirkin's method, the Pearson correlation is 0.3701 for the real faces and 0.4229 for the fake faces. Spearman's rank order correlation is 0.1175 for the real faces and 0.3741 for the fake faces. This indicates that the certainty

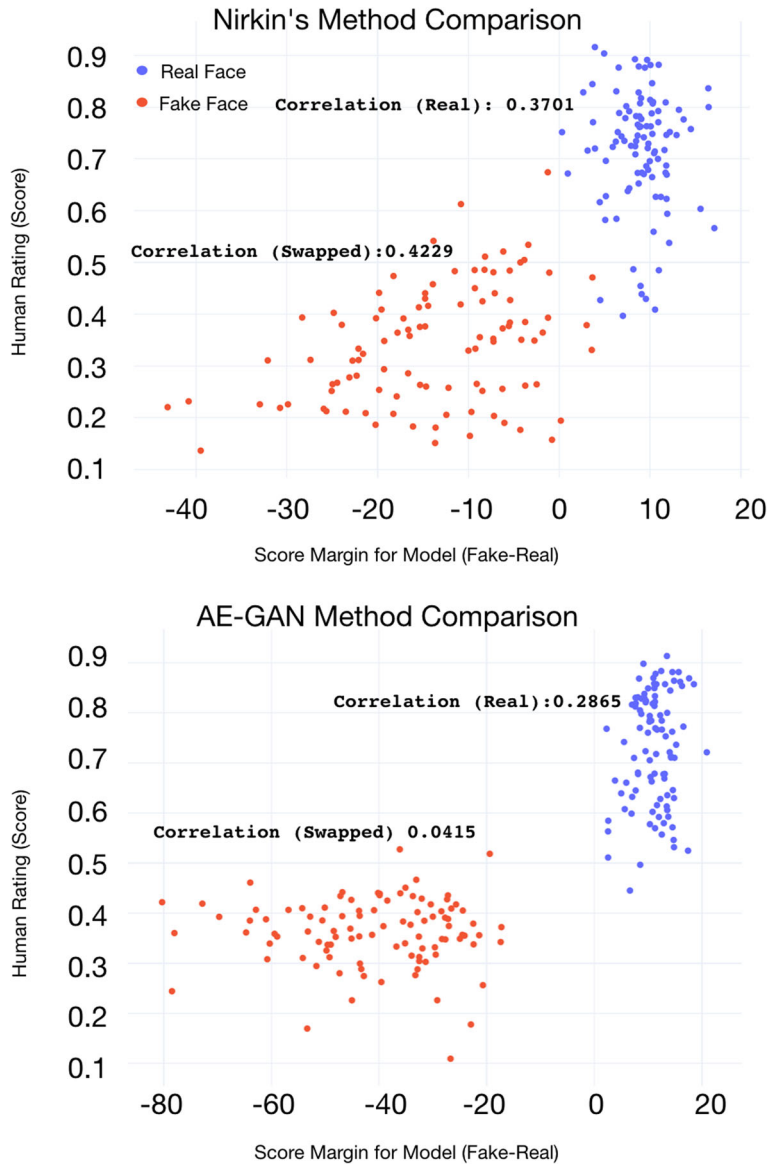

Fig. 6 Top: Nirkin's method Pearson's correlation for real face 0.3701 and swapped face 0.4229. Bottom: AE-GAN method Pearson's correlation for real face 0.2865 and swapped face 0.0415

level of our model and human subjects rating is related, but not perfect—especially for fake images using the AE-GAN method where almost no correlation to humans is observed. We also notice the correlations for the real faces for Nirkin's method and for the AEGAN methods are very close (around 0.09 difference). However, with about 50 data points in each calculation, the linear correlation is sensitive to relatively few outliers. Therefore, one cannot interpret these correlations in a conclusive way. Even so, the correlation in Nirkin's method is encouraging because it shows the model learns not only a binary threshold, but captures some similarity in ranking of the images from most fake to most real. This analysis supports a conclusion that human ranking of realism may be easier to mimic than human ranking of "how fake" an image might be. We anticipate that future work can further improve upon this ranking similarity.

5 Conclusion

In this study, we investigated using deep transfer learning for swapped face detection. For this purpose, we created the largest, to date, face swapping detection dataset using still images. Moreover, the dataset has around 1000 real images for each individual (known largest), which is beneficial for models like the AE-GAN face swapping method. We use this dataset to inform the design and evaluation of a classifier, and the results show the effectiveness of the model for detecting swapped faces. More importantly, we compare the performance of our model with human subjects. We designed and deployed a website to collect pairwise comparisons for 400 carefully picked images from our dataset. Approximate ranking is calculated based on these comparisons. We compared the ranking of our deep learning model and find that it shows good correspondence to human ranking. The code used in this study is available at the noted link[4]. Because deepfake algorithms are continually improving, we hope making our code available will allow the research community to use our model as a baseline for improved methods. We hope this work will assist in the creation and evaluation of future image forensics algorithms.

Finally, we point out that the models created are evaluated against "fakeness" instead of "identity theft" or "identity masquerading." While these two problems are related, the latter are dependent upon the former. That is, our model is able to detect swapped faces, which could be used to detect identity theft because such a swapping is the first step in masquerading someone's identity. We would like to acknowledge that people may have been able to detect a person's identity was fake even if the image looked authentic. In that case, our model would not be needed because the context around the individual was such that a reasonable person could detect that this person was not in the image scenario. In this way, the proposed model is well suited to work with a human to detect identity in an image.

[4] https://github.com/dxywill/swapped_face_detector

Acknowledgements
We thank all the participants rating the images in this study.

Authors' contributions
All authors contributed to the design of this research. XD contributed to the implementation of the research, creation of the dataset, creation of the website for human subject data collection, analysis of the results, and writing of the main part of the manuscript. ZR contributed to the design and implementation of the approximate ranking algorithm, and writing of the manuscript. EVO and MH contributed to the design of the approximate ranking algorithm and writing and editing parts of the manuscript. ECL contributed to the design of the neural network architecture and analysis of the human subjective assessment. PK provided general consultation and contributed to the fine-tuning of the hyperparameters used in the approximate ranking algorithm. All authors read and approved the final manuscript.

Funding
None.

Availability of data and materials
The dataset we used in this study and human subject assessment data are available for academic research use per request at https://www.dropbox.com/sh/rq9kcsg3kope235/AABOJGxV6Zsl4-4bmwMGqtgia?dl=0.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Computer Science, Southern Methodist University, Dallas, USA. [2]Department of Engineering Management, Information and Systems, Southern Methodist University, Dallas, USA. [3]Department of Mechanical Engineering, Southern Methodist University, Dallas, USA.

Received: 23 December 2019 Accepted: 30 April 2020
Published online: 19 May 2020

References

1. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, in *Advances in Neural Information Processing Systems*. Generative adversarial nets (Neural Information Processing Systems, Montreal, 2014), pp. 2672–2680
2. Deepfakes porn has serious consequences - BBC News. https://www.bbc.com/news/technology-42912529. Accessed 28 May 2019
3. J. Schmidhuber, Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
4. Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, G. Medioni, in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. On face segmentation, face swapping, and face perception (IEEE, 2018). https://doi.org/10.1109/fg.2018.00024
5. GitHub - shaoanlu/faceswap-GAN: a denoising autoencoder + adversarial losses and attention mechanisms for face swapping. https://github.com/shaoanlu/faceswap-GAN. Accessed 11 May 2019
6. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing. **338**, 34–45 (2019)
7. R. Heckel, M. Simchowitz, K. Ramchandran, M. J. Wainwright, Approximate ranking from pairwise comparisons. arXiv preprint arXiv:1801.01253 (2018)
8. V. Blanz, K. Scherbaum, T. Vetter, H.-P. Seidel, Exchanging faces in images. Comput. Graph. Forum. **23**, 669–676 (2004). https://doi.org/10.1111/j.1467-8659.2004.00799.x
9. D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, S. K. Nayar, in *ACM Transactions on Graphics (TOG), vol. 27*. Face swapping: automatically replacing faces in photographs (ACM, 2008), p. 39

10. S. Mahajan, L.-J. Chen, T.-C. Tsai, in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*. Swapitup: a face swap application for privacy protection (IEEE, 2017). https://doi.org/10.1109/aina.2017.53
11. D. Chen, Q. Chen, J. Wu, X. Yu, T. Jia, Face swapping: realistic image synthesis based on facial landmarks alignment. Math. Probl. Eng. **2019**, 1–1 (2019). https://doi.org/10.1155/2019/8902701
12. P. Pérez, M. Gangnet, A. Blake, Poisson image editing. ACM Trans. Graph. (TOG). **22**(3), 313–318 (2003)
13. I. Korshunova, W. Shi, J. Dambre, L. Theis, in *2017 IEEE International Conference on Computer Vision (ICCV)*. Fast face-swap using convolutional neural networks (IEEE, 2017). https://doi.org/10.1109/iccv.2017.397
14. R. Natsume, T. Yatagawa, S. Morishima, Rsgan: face swapping and editing using face and hair representation in latent spaces. arXiv preprint arXiv:1804.03447 (2018)
15. K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlasic, W. Matusik, H. Pfister, in *ACM Transactions on Graphics (TOG), vol. 30*. Video face replacement (ACM, 2011), p. 130
16. Y. Zhang, L. Zheng, V. L. Thing, in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*. Automated face swapping and its detection (IEEE, 2017). https://doi.org/10.1109/siprocess.2017.8124497
17. G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, *Labeled Faces in the Wild: a Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49.* (University of Massachusetts, Amherst, 2007)
18. H. Bay, T. Tuytelaars, L. Van Gool, in *European Conference on Computer Vision*. Surf: speeded up robust features (Springer, Graz, 2006), pp. 404–417
19. A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, C. Busch, in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. Fake face detection methods: can they be generalized? (IEEE, 2018). https://doi.org/10.23919/biosig.2018.8553251
20. A. Agarwal, R. Singh, M. Vatsa, A. Noore, in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. Swapped! Digital face presentation attack detection via weighted local magnitude pattern (IEEE, 2017). https://doi.org/10.1109/btas.2017.8272754
21. P. Korshunov, S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
22. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: learning to detect manipulated facial images. arXiv preprint arXiv:1901.08971 (2019)
23. K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Deep residual learning for image recognition (IEEE, 2016). https://doi.org/10.1109/cvpr.2016.90
24. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Imagenet: a large-scale hierarchical image database (IEEE, 2009). https://doi.org/10.1109/cvpr.2009.5206848
25. H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging. **35**(5), 1285–1298 (2016)
26. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, in *2017 IEEE International Conference on Computer Vision (ICCV)*. Grad-cam: visual explanations from deep networks via gradient-based localization (IEEE, 2017). https://doi.org/10.1109/iccv.2017.74
27. K. Guo, D. Tunnicliffe, H. Roebuck, Human spontaneous gaze patterns in viewing of faces of different species. Perception. **39**(4), 533–542 (2010)

Publisher's Note