Double Face: Leveraging User Intelligence to Characterize and Recognize AI-synthesized Faces

Matthew Joslin, Xian Wang, and Shuang Hao, *University of Texas at Dallas*

https://www.usenix.org/conference/usenixsecurity24/presentation/joslin

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

Double Face: Leveraging User Intelligence to Characterize and Recognize AI-synthesized Faces

Matthew Joslin Xian Wang Shuang Hao
University of Texas at Dallas
{matthew.joslin, xxw161030, shao}@utdallas.edu

Abstract

Artificial Intelligence (AI) techniques have advanced to generate face images of nonexistent yet photorealistic persons. Despite positive applications, AI-synthesized faces have been increasingly abused to deceive users and manipulate opinions, such as AI-generated profile photos for fake accounts. Deception using generated realistic-appearing images raises severe trust and security concerns. So far, techniques to analyze and recognize AI-synthesized face images are limited, mainly relying on off-the-shelf classification methods or heuristics of researchers' individual perceptions.

As a complement to existing analysis techniques, we develop a novel approach that leverages crowdsourcing annotations to analyze and defend against AI-synthesized face images. We aggregate and characterize AI-synthesis artifacts annotated by multiple users (instead of by individual researchers or automated systems). Our quantitative findings systematically identify where the synthesis artifacts are likely to be located and what characteristics the synthesis patterns have. We further incorporate user annotated regions into an attention learning approach to detect AI-synthesized faces. Our work sheds light on involving human factors to enhance defense against AI-synthesized face images.

1 Introduction

Recent advances in deep neural networks have significantly improved AI synthesis techniques such as Generative Adversarial Networks (GANs) that automatically produce realistic-appearing images [1, 2, 3]. Despite positive applications [4, 5, 6], increasingly realistic models are being abused in online fraud and mischief, which causes severe trust and security threats. Especially, the abuses of AI-synthesis techniques generate images of *nonexistent yet photorealistic* persons to deceive users or spread propaganda.

Deceptive content historically used low-quality content or plagiarized others' images, whose origins are easier to identify (such as by reverse image search [7]). But AI-synthesis techniques allow miscreants to circumvent existing defenses by generating realistic-appearing and unique face images. Recently, we have witnessed a surge in real-world incidents. For example, an AI-synthesized photo was used in a suspected spy's LinkedIn profile as part of espionage efforts [8], a fake political candidate on Twitter (set up by a high school student) was created with the portrait image of a nonexistent person that AI algorithms produced [9], and groups of AI-generated profile photos of fake accounts were found on Facebook distributing polarization [10].

Challenges. Unfortunately, so far, techniques to counter AI-synthesized facial images are limited. Existing detection approaches [11, 12, 13, 14, 15] mostly train off-the-shelf classifiers, susceptible to the risk of picking superficial features [16]. While some research [17, 18] suggested specific facial parts as detection features, the features were based on the heuristics of individual researchers, limited to the individual perception or assumptions. We lack a systematic understanding of what detectable AI artifacts are more prevalent quantitatively in such images, specifically what regions are suspicious and what specific patterns resemble synthesis.

Our approaches and findings. In this paper, we develop a complementary novel approach that uses crowdsourcing intelligence to characterize and defend against AI-synthesized face images. We derive and analyze AI-synthesis artifacts annotated by multiple users (instead of by individual researchers or automated systems), which provides new insights to investigate and identify AI-synthesized images. While an individual user has limited capability, the annotations of multiple users will aggregate in finding constantly marked artifact regions (example user annotations in Figure 1, where heat color indicates the number of annotations aggregated for a position). We focus on human face synthesis, as this poses high practical threats in social engineering attacks [10, 19, 20]. Research in neuropsychology has shown that the human visual system is sensitive to face perception [21, 22], which has natural advantages for the analysis of face images.

Our work systematically identifies *where the synthesis artifacts are likely to be located* and *what characteristics the*

synthesis patterns have. We design a user study that prompts users to visually draw suspicious regions on AI-synthesized face images, and input text to describe annotated regions. While prior work [23, 24, 25] has explored crowdsourcing studies to analyze AI-synthesized images, these studies used designs limited to only binary questions or rating questions. Instead, our design (drawing regions and inputting text to describe suspicious regions) allows deep quantitative analysis of what specific artifacts often occur in AI-synthesized images and how these results can be leveraged for detection.

Our analysis compares and quantifies 15 different regions, including facial and non-facial regions. For comparative analysis, we include a control group of real images in experiments. For AI-synthesized images, we use the images generated by StyleGAN2 [3] and StyleGAN3 [26], which are widely used in practice [27, 28] and have been reported in real-world incidents [29, 9]. We discover what synthesis artifacts are prevalent and distinct from real images as opposed to prior work which only suggested the existence of artifacts [18, 17]. We summarize our main findings on AI-synthesized face images. (1) Facial regions show various degrees of artifacts despite exhibiting AI synthesis defects. Ear and hair regions in synthesized images are more likely to exhibit defects that users discern, ranging from 2.3 to 4.3 times higher likelihood compared to the control group of real images. On the other hand, other regions, such as nose and eye, present less effectual artifacts to distinguish synthesized images without showing statistical significance. In general, we observe that the facial regions distant from the central face area lead to increasing levels of distinguishable artifacts. (2) Non-facial regions show consistently high likelihoods of AI synthesis artifacts. Non-facial objects of hat, earring, eyeglass, or clothes are frequently annotated as suspicious regions in synthesized images, with an overall 6.6 times higher likelihood compared to the control group of real images. We find that non-facial regions have higher probabilities to exhibit synthesis artifacts compared to facial regions. (3) The prevalent pattern of artifacts in synthesized images is blur, e.g., correlated with ear, eyeglass, and clothes regions. Additionally, the facial and non-facial regions show other unique patterns, such as skin anomalies in facial regions and unknown objects in non-facial regions. The patterns in suspicious regions in synthesized images are generally different from the control group of real images. Our findings provide new insights into synthesis artifacts to characterize and recognize AI-synthesized faces. Based on the results, we compile actionable suggestions for content moderators to screen AI-synthesized images empirically.

Furthermore, with the intelligence of user annotations, we investigate the potential of enhancing the detection of AI-synthesized faces. We adopt an attention learning approach, which guides inference to the regions of interest. The artifact regions from crowdsourcing annotations provide attention guidance during the training of the detection model. We include eight synthesis models in detection experiments, spanning GAN, autoencoder, and diffusion models. The evaluation on AI-synthesized faces shows that the detection guided by user annotations outperforms the state-of-the-art detection approaches [30, 15, 17], and achieves comparatively high generalization performance. We further evaluate common evasions on the synthesized images, and observe that the detection with user annotations retains high robustness against evasions.

Figure 1: Examples of user annotations on AI-synthesized face images. Multiple users draw bounding boxes to locate suspicious regions. We use heat color to indicate the number of annotations aggregated for a position (red color parts have high annotation numbers). The aggregation of multiple user annotations focuses on key artifact regions.

To summarize, we make the following contributions in this paper.

- We take a novel analysis perspective and perform an empirical study to characterize artifacts that users commonly perceive in AI-synthesized face images. We develop a crowdsourcing annotation approach to systematically aggregate multiple user annotations that locate suspicious regions and extract artifact patterns.
- We characterize what synthesis artifacts and patterns are prevalent in AI-synthesized face images. We find that facial regions distant from the center (such as ear and hair) are more likely to exhibit synthesis artifacts, and non-facial regions consistently show high likelihoods of defects (compared to facial regions). The prevalent pattern of artifacts in synthesized images is blur. Our findings provide empirical insights to recognize AI-synthesized faces.
- We incorporate an attention learning method with user annotations to detect AI-synthesized faces. The experiments show our approach outperforms the state-of-the-art detection approaches in accuracy and remains robust against

evasions. The results demonstrate the potential benefits of human intelligence to defend against synthesized face images.

2 Background

We introduce the background of AI-synthesis techniques that automatically generate face images, and describe general human visual system which is evolved to recognize and perceive faces.

2.1 AI-synthesis Techniques and Models

AI techniques with neural networks are used to generate photorealistic-looking photos of people who do not exist in reality. Among various open-source synthesis techniques, Generative Adversarial Networks [31] (GANs) achieve the state-of-art performance and efficient computation, in which two neural networks compete to improve quality iteratively. A series of GAN models are developed to achieve high synthesis effects. ProGAN [1] added neural network layers and learned progressively to generate large-resolution images. StyleGAN [32] (based on ProGAN) introduced finer latent representation and style information, and provided a dataset of real human faces FFHQ which became a common training set for later models. StyleGAN2 [3] (based on StyleGAN) simplified dataflow and overcame previously noticed weaknesses. MSG-GAN [33] (based on StyleGAN2) was developed to improve the stability of synthesizing high-resolution images. StyleGAN3 [26] (based on StyleGAN2) used continuous representation for detail transformation. Anyres-GAN [34] (based on StyleGAN3) sampled patches to synthesize images at arbitrary scales. In addition, other AI-synthesis approaches have been developed for producing high-quality images. For example, StarGAN v2 [35] performed an image-to-image translation using multiple modules to generate images with diverse styles. Nouveau Variational Autoencoder (NVAE) [36] learned a latent encoding through a variational autoencoder (VAE) and generated images using the decoder. Latent Diffusion Model (LDM) [37] employed a denoising autoencoder to model images as a diffusion process from a latent vector.

In our user study (Section 4), we use face images generated from StyleGAN2 and StyleGAN3, the milestone GAN models which have attained high popularity in practice [27, 28]. Anecdotes have shown that face images generated from StyleGAN2 were used to deceive users [9]. The image quality of StyleGAN3 is similar with StyleGAN2 (in terms of FID, Frechet inception distance) as presented by the original work [26]. The difference is that StyleGAN3 has changes in internal representations to address texture sticking and improve the equivariance metrics. In detection experiments (Section 6), we include more models to examine detection generalization (including various GAN models, and autoencoder and diffusion models).

2.2 Human Perception Sensitivity on Faces

Humans use vision perception as a primary sensory means and have dedicated neurobiological capability to recognize faces. A significant amount of neural signals are for visual information processing. Prior research has shown that human visual system is specialized for the recognition of faces [38, 22]. For socialization, humans evolve to be capable of quickly extracting traits of faces and identifying face differences. Humans are sensitive to processing facial information [38]. For vision tasks, humans have advantages to achieve high-quality results.

On the other hand, human attention is more attracted by face images. Photographic face images promote affection and social attraction of users [39]. Social networks and service websites typically have user accounts to set with profile photos. Displaying profile face photos is an important factor to influence decision making online, such as friend request acceptance [40, 41] or online purchasing [42, 43]. Therefore, face images have become a target of attackers to manipulate and falsely gain users' trust to perform malicious activities.

3 Design Overview and Research Questions

AI-synthesized face images have been increasingly abused by miscreants to falsely gain users' trust in social engineering attacks, such as creating fake social media profiles (real-world incidents [8, 9, 19, 10]). Existing detection approaches mostly rely on black-box classifiers or heuristics of researchers' individual perception [11, 12, 13, 17, 30, 15]. In contrast, we develop a novel method to leverage user perceptions to systematically characterize and defend against AI-synthesized face images. We derive and analyze AI-synthesis artifacts annotated by multiple users (instead of by individual researchers or automated systems). Our quantitative analysis answers the following research questions.

RQ1. Do AI-synthesized face images contain artifacts that users commonly perceive?

RQ2. Where in the AI-synthesized face images are the synthesis artifacts located?

RQ3. What patterns do the perceived artifact regions exhibit in AI-synthesized images?

RQ4. How can user perceptions be used to facilitate detecting AI-generated face images?

Figure 2 shows the design overview of our study, and the corresponding components that investigate the research questions. We conduct a user study that collects crowdsourcing annotations to find and characterize commonly perceived synthesis artifacts (Section 4 and Section 5). Specifically, for each face image, users are requested to rate the face fidelity

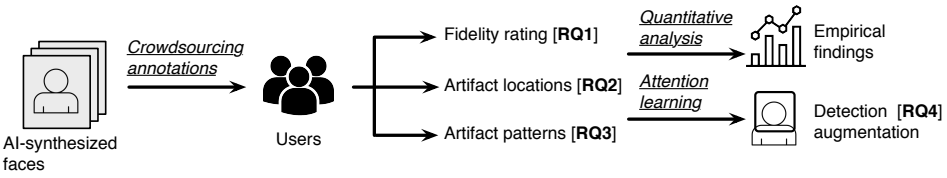


Figure 2: Design overview of our study. The analysis aims to answer the research questions **RQ1** to **RQ4** described in Section 3.

(for **RQ1**), draw bounding boxes to locate suspicious regions (for **RQ2**), and input textual fields to describe annotated regions (for **RQ3**). Our study design focuses on annotating and extracting suspicious regions, which is different from prior crowdsourcing work [23, 24, 25], since our goal is to find and characterize suspicious regions of synthesis artifacts. As described in Section 2.2, humans have dedicated visual capabilities of face perception [21, 22], so crowdsourcing has advantages for analysis of face images. Note that we aggregate multiple users' annotations (not just a single user) to extract commonly perceived artifacts and prune outlier mistakes. Based on the aggregated user annotations, we qualitatively compare and characterize various artifacts in AI-synthesized face images. The findings provide new insights of empirically recognizing synthesized faces (we discuss suggestions for content moderators in Section 7). To further leverage user annotations in automated detection, we incorporate an attention learning mechanism (Section 6) to detect AI-synthesized face images (for **RQ4**). While attention mechanisms traditionally adjust network architecture to optimize feature weights [44, 45, 46], attention from input data provides guidance information [47]. We use an attention learning approach to combine the attention from human annotations and the features learned from neural networks. The detection improvement demonstrates the capability of user annotations to facilitate recognizing AI-synthesized faces.

4 Methodology of Crowdsourcing Annotations

We describe the design of crowdsourcing annotations and the user study to extract AI-synthesis artifacts recognized by multiple users. In this work, we develop crowdsourcing annotations to quantitatively identify and characterize artifact regions that users commonly perceive.

4.1 User Study Settings

We conduct user study experiments on the crowdsourcing platform, Amazon Mechanical Turk (MTurk). We develop a web annotation interface (details in Section 4.2) which accommodates MTurk participants to draw suspicious regions on the images and describe artifact patterns. We provide an instruction page which only informs participants how to use the annotation tool and explains that the images may be fake from AI generation, rather than providing detailed instructions on how to find synthesis artifacts or what suspicious artifacts look like (to avoid biasing users' own perception). Human users are equipped with the visual system to naturally capture abnormal regions on face images. To elicit comprehensive analysis, we add a control group of real images in the experiments (comparison results in Section 5). Each participant will be assigned a sequence of synthesized images and real images (as the control group) for annotations. The images will be displayed in random order.

Face images for crowdsourcing annotations. For crowdsourcing annotations, we use publicly available models of StyleGAN2 [3] and StyleGAN3 [26], which are widely used AI-synthesis models for human portraits and have high star and fork rankings in practice [27, 28] (we include more synthesis models for detection experiments in Section 6.2). Moreover, face images generated by StyleGAN2 and StyleGAN3 have been witnessed in real-world incidents or attacks [29, 9]. The two synthesis models facilitate comparison and understanding of how AI synthesis has improved. We randomly generated 100 synthesized images from StyleGAN2 and 100 synthesized images from StyleGAN3. For the control group of real images, we randomly sampled 100 images from the FFHQ dataset [32], as FFHQ is the basis of real images for training most of the recent face models. The images have high quality with 1024×1024 resolution. We manually examine and present the distribution of the gender and ethnicity of the faces in Appendix A.

4.2 Crowdsourcing Annotation Designs

We develop the annotation interface for MTurk participants to mark and describe the artifacts that they perceive on each image. Figure 3 shows the example annotation webpage that we design. The annotation webpage displays the images for annotation and contains instructions to participants. The instructions inform participants how to use the annotation tool and describe the scenario rather than providing detailed instructions on what to look for in the AI-synthesized images, to avoid annotation biases. The language of the instructions is kept concise and intuitive, which allows regular users to comprehend the task and use their own judgment to explore suspicious regions.

In the procedure of the task, the participant first agrees online a consent form. The participant is then guided through an instruction page. The instruction page includes an inter-

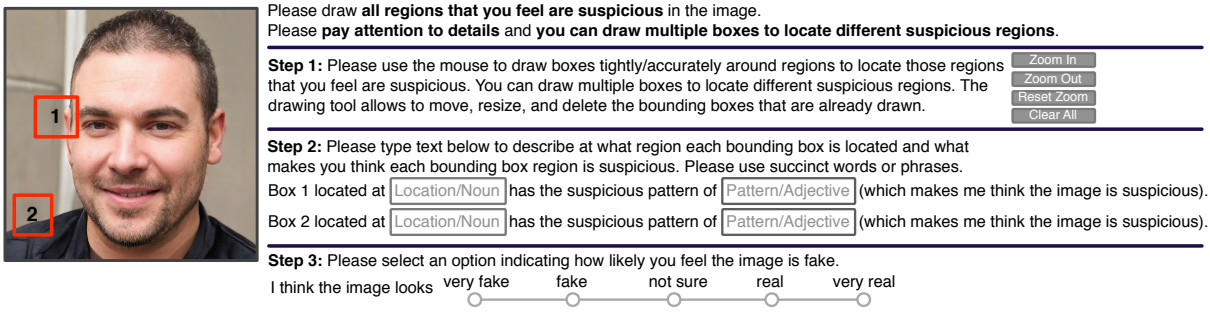


Figure 3: The crowdsourcing annotation webpage that we develop to instruct users through the scenario and annotate the displayed image. **Step 1** requests users to draw on top of the image and carefully size and position the bounding boxes to capture suspicious regions. Upon drawing, the boxes are assigned indexes for reference in **Step 2** where users are requested to complete the prompt on where the box is located and describe what is suspicious with the region (the rows are dynamically updated according to the number of drawn boxes). Finally, **Step 3** asks the users to select from the rating choices to indicate how fake the image appears.

face to get familiar with the drawing tool, which requests the participant to use the mouse to draw boxes and move or resize the drawn boxes. This also proves the participant can understand instructions and draw boxes through the webpage interface. Next the participant will start the main task of annotating images. The annotation task of each image includes the following three steps.

Step 1: Drawing bounding boxes to locate artifact regions. We request participants to use their mouses to draw a set of bounding boxes on the displayed image to mark which regions they consider suspicious (such as the red boxes drawn in Figure 3). The main purpose of drawn boxes is to visually collect and aggregate artifact regions that participants perceive. When the cursor is moved on the image, crosshairs with extended lines are displayed to track position and assist drawing accuracy. The orders of the drawn boxes of each participant are indexed (to be referred in Step 2). We do not place any restrictions on how many boxes participants may draw (allowing participants to draw no boxes) and how large the boxes can be. Relaxing such constraints gives participants full flexibility to reflect their perceived suspicious regions (more analysis in Section 5.1). We provide a set of operations on the webpage interface to facilitate accurate drawing, including moving, resizing, and deleting boxes, and zooming into the image to closely examine an area.

Step 2: Adding text to describe artifact regions. When each bounding box is drawn, we dynamically add text fields on the webpage and request participants to input text to describe the box region. The main purpose of text description is to allow participants to express how they think about suspicious regions in fake images. We use the collected information to characterize and derive artifact taxonomy (more analysis in Section 5.4). To keep the task intuitive for participants, we format the questionnaire as filling two fields in a sentence. Figure 3 shows examples displayed to participants: "`Box 1 located at` Location `has the suspicious pattern of` Pattern `which makes me think the image is suspicious.`" Participants fill text in the two fields of `Location` and `Pattern`. The `Location` field corresponds to what region/object in the image the participants meant to mark (e.g., ear, teeth). The `Pattern` field requests to describe the artifacts in the participants' own words. The open text fields allow participants to flexibly explain perceived observations. When the participant draws a box, a corresponding text question for the box will be added on the web interface (if a bounding box is deleted by the participant, the text question will also disappear). For example, in Figure 3, the participant draws two bounding boxes, and the participant will be prompted with two questions to describe the location and pattern of the drawn boxes respectively. For each box drawn on the image, we mandate the participant to enter at least one character in the corresponding description fields before allowing the task to proceed to the next image (i.e., the button for the next image becomes available to click after the condition is met). The description inputs also help participants to re-examine and justify the annotated box regions.

Step 3: Rating image fidelity. The participants are requested to rate the image's overall fakeness on a predefined five-level Likert scale [48] of "`I think the image looks very fake`", "`looks fake`", "`not sure`", "`looks real`", and "`looks very real`". As shown in Figure 3, the participants click one of the radio buttons for rating (initially no button is selected). We use the ratings to assess the users' overall perception how real or fake an image appears. Selection from descriptive ratings is explanatory to users, compared to direct numerical ratings [48]. If a participant selects a rating level towards fakeness, including "`looks very fake`" and "`looks fake`", we will mandate the participant to draw at least one box on the image to mark suspicious regions, before

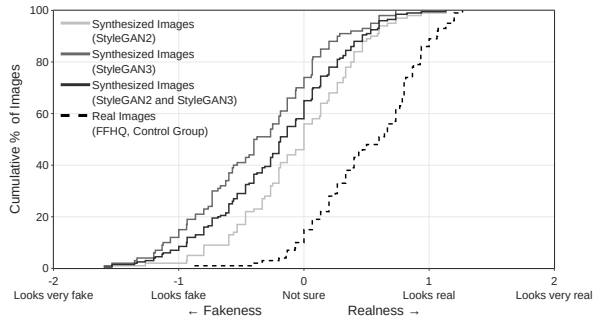


Figure 4: Distribution of the average fakeness rating. The x-axis is the average fakeness rating computed by converting the five-level descriptive Likert scale to integers from -2 to 2. We use an average score of 0 (level of "not sure") to separate perceptions between fake and real categories. On average over multiple users, 89% real images are rated towards real and 65% of synthesized images are rated towards fake.

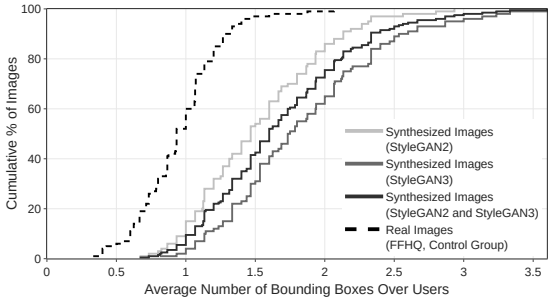


Figure 5: Distribution of the average number of boxes drawn by users for each image. We observe that 52% of real images have less than one annotation drawn on average, while 94% of synthesized images have one or more annotations.

allowing the task to proceed to the next image (if the condition is not met, clicking the button for the next image will not proceed and the webpage will display a text warning to participants).

After a participant completes the steps as described above for an image, the button to proceed to the next image becomes available and the participant clicks to start annotating another image. The top of the webpage will show the number of total images in this task and the sequence number of the current image, facilitating participants to track the progress of the task. At the end of the procedure, the participant can fill in demographic questions optionally.

Figure 1 shows examples of bounding boxes that participants draw by using our MTurk webpage interface to annotate AI-synthesized face images. Multiple participants draw bounding boxes to locate suspicious regions. The heat color indicates the number of annotations aggregated for a position (red color parts have high annotation numbers). As example text that participants input to describe suspicious regions for the synthesized images in Figure 1, the top left face image got descriptions such as location "ear" with pattern "abnormal", the top right image was marked location "eyeglass" with pattern "incomplete" or "bent", the bottom left face had "hair" annotated as "deformed", and the bottom right image had location "earring" described with pattern "blur". We perform in-depth analysis to characterize user annotations in Section 5.

4.3 Recruitment for User Study

We recruited participants on MTurk between June and September 2023 for crowdsourcing annotations. Our study was conducted with the approval of the institutional review board (IRB). We did not collect any personal identifiable information (PII) from the participants. The participants first completed the consent form before starting tasks (as described in Section 4.2). We carefully protected the anonymity of the participants. We provided options for participants to opt out of the experiments (we did not receive any opt-out requests). More details on ethics considerations are in Section 7. We conducted a small pilot experiment to ensure that the instructions and interface are clear for participants. For new participants, the first two images were used as an exercise to get familiar with the annotation procedure and will not be counted in the results.

Each image is annotated by 15 unique participants. We recruited U.S. workers who have successfully completed at least 5,000 tasks and had an approval rating greater than 98% (this setting is similar to or above the ones used in prior studies [49, 50] to recruit high-quality participants). We deployed an attention check question to verify that participants carefully complete the tasks. In addition, at the instruction page of tasks, we requested that participants completed a step to both draw boxes on an image and then move or resize the boxes, to verify that the participants could understand instructions and were capable to use the drawing tools. Participants typically took 1-1.5 minutes to complete one image. We paid the participants $0.25 for each image annotated. We designed the experiment with a power analysis [51] with a power of 0.8 using binomial proportion tests at significance level of 0.05 and a standard effect size 0.4. In total, we recruited 185 MTurk participants to annotate images in our experiments. We present demographic statistics optionally input by participants in Appendix B.

5 Characterization and Findings

In this section, we present detailed analysis of the crowdsourcing annotation results. We examine user perception capabilities, aggregate regions commonly annotated by multiple users, and extract artifact patterns to characterize AI-synthesized faces (**RQ1**–**RQ3**).

5.1 Overall Perception on AI-synthesized Images

We characterize the overall perception of human users for AI-synthesized portraits (**RQ1**). For comparison (as described in Section 4), we introduce a control group by adding real images to each set of annotation tasks. From our annotation results, we analyze and compare how users perceive synthesized and real images.

Overall fidelity ratings. We request participants to rate the fakeness level that they perceive for each image (Step 3 in Section 4.2). We convert the five-level descriptive Likert scale to integers from -2 to 2 (any evenly distributed numerical values will lead to the same analysis results). Figure 4 shows the distribution of images' average fakeness ratings. The x-axis indicates the average fakeness level over participants for an image. The y-axis means the cumulative percentage of the images. By using an average score of 0 (level of "`not sure`") to separate perception between fake and real categories, we observe that on average (over multiple participants) 89% real images are rated towards real and 65% of synthesized images (56% for StyleGAN2 and 74% for StyleGAN3) are rated towards fake. The observation suggests that human visual system genuinely recognizes differences between AI-synthesized face images and real face images.

Bounding box numbers. We further examine the number of bounding boxes that participants draw on the images. We provide participants full flexibility to decide how many boxes to annotate suspicious regions on an image (Step 1 in Section 4.2). Figure 5 shows the distribution of the average number of boxes per image drawn by participants. The x-axis represents the average number of boxes over participants for an image. The y-axis displays the cumulative percentage of the images. We observe that 52% of real images have less than one annotation drawn on average, while 6% of synthesized images (9% for StyleGAN2 and 3% for StyleGAN3) have less than one annotation. Moreover, just 1% of real images show more than two annotations on average; on the other hand, 24.5% of synthesized images (15% for StyleGAN2 and 34% for StyleGAN3) have more than two annotations. The results show that participants carefully examine images and annotate more frequently in synthesized portraits than in the control group of real portraits.

> **Takeaway 1**: On average, users recognize the differences in AI-synthesized face images from the control group of real ones, and draw more boxes to annotate suspicious regions in the synthesized images.

5.2 Extracting Aggregate Regions

In the crowdsourcing experiment, we collected boxes drawn by users (Step 1 in Section 4.2) which indicate individual's perceived suspicious parts on AI-synthesized images. Figure 1 shows examples of user annotations on synthesized face images (where the heat color indicates the number of annotations aggregated). We focus on analyzing the regions that are constantly marked by multiple users, to characterize commonly recognized synthesis artifacts (in Section 5.3 and Section 5.4). We also use the extracted regions as annotation masks to augment detection of AI-synthesized face images (in Section 6).

Algorithm 1 Extraction of aggregate regions

Input: A list L_{boxes} contains the bounding boxes that all users drew for an image. Each item M_{box} in L_{boxes} is a matrix (same dimensions as the image) marking the bounding box area denoted with 1 and all other positions with 0.
The operations $+$, $-$, and Σ are point-wise matrix operations.
The procedure constants *MinAgree* (a minimum agreement level of a region) and *AgreePercent* (a percentage used to ensure a relatively similar level agreement across a region).
Output: The list of extracted aggregate regions $L_{regions}$ for the image.

```
1:  L_regions ← list();
2:  M_aggregate ← Σ_{L_boxes} M_box;
3:  while (findMax(M_aggregate) ≥ MinAgree ) do
4:      PeakValue ← findMax(M_aggregate);
5:      AgreeLevel ← max(MinAgree, AgreePercent × PeakValue);
6:      M_mask ← extractGreaterOrEqual(M_aggregate, AgreeLevel);
7:      L_candidates ← findConnectedComponents(M_mask);
8:      for M_candidate in L_candidates do
9:          if findMax(M_candidate) < PeakValue then
10:             L_candidates ← deleteUpdate(L_candidates, M_candidate)
11:         end if
12:     end for
13:     M_region ← outputMinSizeItem(L_candidates)
14:     L_regions ← appendUpdate(L_regions, M_region);
15:     for M_box in L_boxes do
16:         if intersect(M_box, M_region) then
17:             L_boxes ← deleteUpdate(L_boxes, M_box);
18:             M_aggregate ← M_aggregate − M_box;
19:         end if
20:     end for
21: end while
22: return L_regions;
```

We develop Algorithm 1 to extract aggregate regions from bounding boxes that multiple users draw. The input is the list of bounding boxes submitted by users for an image, L_{boxes}. The output is the list of extracted regions, $L_{regions}$. The algorithm iteratively extracts the top aggregate region in each round.

1. At the beginning of the round (Line 3), we examine how many layers of annotations are stacked across the bounding boxes that multiple users drew independently (agreement level). We proceed the extraction if the agreement level in the remaining multi-layer masks is not lower than $MinAgree$. We set the practical $MinAgree$ value as four annotations (greater than 25% of participants for each image in our experiment).

2. We extract the region that has the most annotations for this round (Lines 4-14).

   (a) To separate regions (Lines 4-6), we require the agreement levels of the points included are

Figure 6: Example regions extracted by Algorithm 1 on AI-synthesized face images. The bounding boxes that users annotated on these images are shown in Figure 1.

not lower than *AgreePercent* of the peak value, in addition to *MinAgree*. We set the practical *AgreePercent* value as 50%.

(b) We use a union-find strategy [52] to extract connected components based on the multi-layer masks as candidates (Line 7). We filter the candidate regions by requiring to include the peak value (Lines 8-12).

(c) To break the tie of multiple candidate regions, we select the candidate with the smallest extracted size (Line 13) and add it as the extracted region for this round (Line 14). The strategy prioritizes choosing compact regions.

3. We exclude the selected bounding boxes, which intersect with the currently extracted region, from the aggregate masks (Lines 15-20), to prepare for the next round extraction.

Example extraction results are shown in Figure 6 (where the original bounding boxes that users drew are shown in Figure 1). Aggregation of multiple user annotations points to artifact regions.

Numbers and sizes of extracted regions. Table 1 presents statistics of the extracted regions. We first examine the average number of extracted regions over images. The average number of extracted regions on AI-synthesized images is 2.10, and the average number of extracted regions in the control group of real images is 0.96. The synthesized images have extracted regions 2.19 times more on average than the control group of real images (ratio $2.10/0.96 = 2.19$). We further analyze the average region size percentage to the full image size. The average size of extracted regions on AI-synthesized images is 2.17% of the image size, and the average size of extracted regions on the control group of real images is 2.45% of the image size. The differences have statistical significance with low p-value ($p < 0.001$). The finding shows that the annotation aggregation results focus on detailed regions in the images.

Table 1: Average numbers and size percentages of the regions extracted from annotation aggregations. The rows represent image sets. The columns indicate region statistics.

Image Set		*Average Region Number*	*Average Size Percent %*
Synthesized	*StyleGAN2*	1.83	2.14
	StyleGAN3	2.37	2.19
Synthesized Images (StyleGAN2 and StyleGAN3)		2.10	2.17
Real Images (FFHQ, Control Group)		0.96	2.45

Takeaway 2: The extracted regions that multiple users annotate as suspicious in synthesized images occur with higher incidence and focus on detailed regions.

5.3 Characterizing Artifact Locations

Based on the extracted regions (from Algorithm 1) that multiple users collectively marked, we examine `Location` text input by the users (Step 2 in Section 4.2) to identify where in the AI-synthesized face images the synthesis artifacts are located (**RQ2**). We use natural language processing techniques to analyze user input text. We apply the widely-used lexical database WordNet [53] to parse the user location text. We extract lexical parent categories from WordNet to filter candidate noun words referring to objects. The terms that specify human body regions are grouped based on the established categorization list [54]. We further group similar words that have common hypernyms in WordNet for the non-body region words. Our analysis extracts 15 categories that frequently occur in user annotations, including ten facial categories and five non-facial categories (as the x-axis in Figure 7).

We note that some categories such as "`nose`" or "`mouth`" occur in all face images, but other categories such as "`earring`" occur in only subset of the images. To account for how often these categories appear, we manually review images and count numbers with the accessories that occur. We measure the conditional probability of a given location category being marked by users (i.e., the count of being marked divided by the count of images with the location appearing in the images). We apply the two-sample one-sided hypothesis testing (binomial proportion test) [55] to assess whether the synthesized and real images exhibit statistically significant differences for annotated locations. The difference is considered statistically significant when p-value is low ($p < 0.05$).

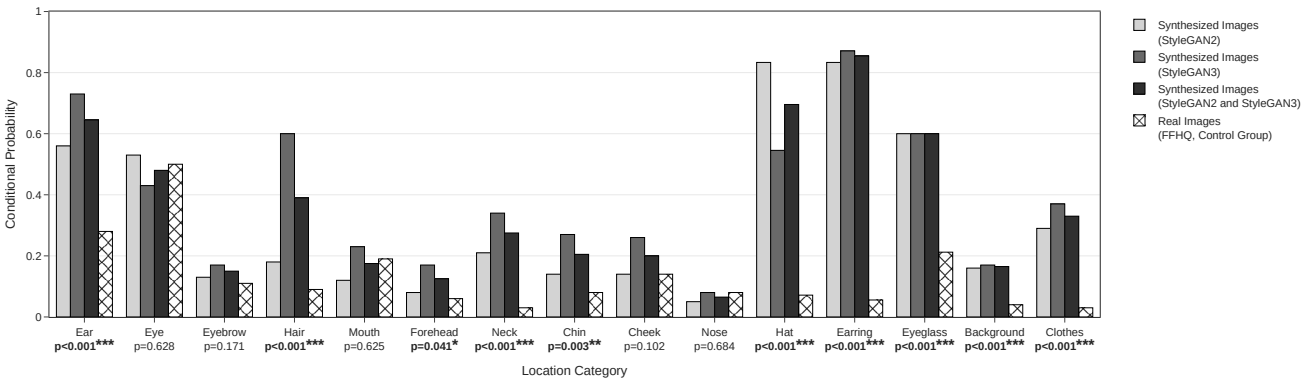


Figure 7: Conditional probabilities of the locations marked by users. The facial categories include ear, eye, eyebrow, hair, mouth, forehead, neck, chin, cheek, and nose. The non-facial categories include hat, earring, eyeglass, background, and clothes. The values below the x-axis labels represent p-value. Significance level is indicated as * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

Figure 7 shows conditional probabilities for different location categories. The x-axis presents the location categories. The values below the category labels represent p-value. The y-axis shows the conditional probabilities of the categories. We observe that the synthesized images show higher conditional probabilities of location annotations than the control group of real images with statistical significance ($p < 0.05$), which indicates distinct artifacts in AI-synthesized images.

Facial location characterization. In Figure 7, we observe that facial regions show diverse degrees of probabilities to exhibit AI artifacts. Regions of "ear" and "hair" are more likely to exhibit defects that users discern. Specifically, the likelihood of "ear" extracted as suspicious regions in the synthesized images is 2.32 times higher than in the control group of real images (ratio $0.65/0.28 = 2.32$, $p < 0.001$), and the likelihood of "hair" in the synthesized image is 4.33 times higher than in the control group of real images (ratio $0.39/0.09 = 4.33$, $p < 0.001$). We note that these areas distant from the central face area have high levels of distinguishable artifacts. On the other hand, the regions of "nose", "eye", and "mouth" have relatively low probabilities of being annotated or result in similar likelihood (without a statistically significant difference) between synthesized and real images, though prior studies [18, 17] suggested artifacts in such regions. In Section 5.4 we further characterize the region patterns and find pattern differences between synthesized and real images. The observations on facial regions show that certain facial regions have high likelihoods with artifacts to distinguish AI-generated images.

Takeaway 3: Facial regions show various degrees of artifacts. Regions of "ear" and "hair" are more likely to exhibit defects in synthesized images, about 2.3 to 4.3 times higher likelihood compared to real images. Other regions such as "nose" and "eye" present less effectual artifacts to distinguish synthesized images.

Non-facial location characterization. We observe in Figure 7 that non-facial objects, including "hat", "earring", and "clothes", show consistently high likelihoods of synthesis artifacts, and annotations in the control group of real images typically do not converge on these regions. The likelihood of "hat" as suspicious regions from user annotations in the synthesized images is 9.72 times higher than in the control group of real images (ratio $0.69/0.071 = 9.72$, $p < 0.001$), the likelihood of "earring" in the synthesized images is 15.18 times higher than in the real images (ratio $0.85/0.056 = 15.18$, $p < 0.001$), and the likelihood of "clothes" in the synthesized images is 11.00 times higher than in the real images (ratio $0.33/0.03 = 11.00$, $p < 0.001$). With the conditional probabilities calculated over all non-facial categories, the non-facial regions in synthesized images overall have 6.6 times higher likelihood of being annotated as suspicious compared to the control group of real images. The observations show that non-facial regions frequently contain artifacts in AI-synthesized images, presumably due to the variety of the locations and styles of these objects which may cause the model to undertrain on properly rendering the regions. Examining these non-facial regions will facilitate recognition of AI-synthesized face images.

Takeaway 4: Non-facial regions, such as "hat", "earring", and "clothes", show consistently high likelihoods in synthesized images to contain artifacts. Non-facial regions overall have 6.6 times higher likelihood of being recognized as suspicious in synthesized images compared to real images.

5.4 Characterizing Artifact Patterns

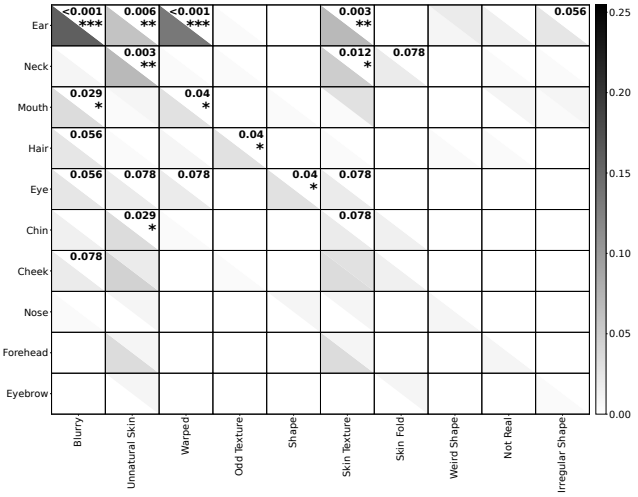
We analyze user inputs for the Pattern prompt (Step 2 in Section 4.2) to understand what visual characteristics highly correlate with specific regions being annotated. We seek to

answer the question what synthesis patterns are commonly found in AI-synthesized portraits (**RQ3**). We correlate the relationship between the pattern descriptions and location descriptions from the user annotations to characterize artifact attributes.

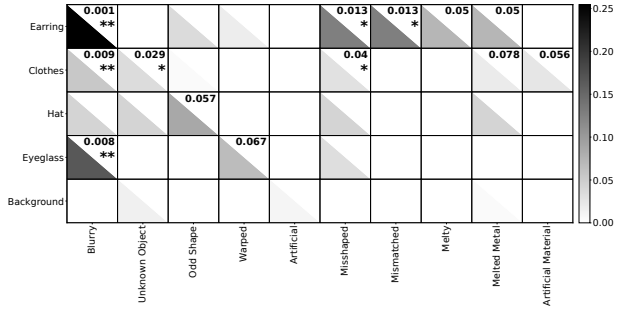
We calculate the conditional probabilities of the description words occurring given the location tokens extracted in Section 5.3. A high probability indicates that the token pair has high correlation, i.e., the annotated location is frequently described as the specific pattern. For the textual descriptions in the `Pattern` fields, we remove stop words, lemmatize the words to their canonical forms, and generate the unigram and bigram tokens. In addition, we calculate p-value via the two-sample one-sided hypothesis testing [55] to measure significance of differences between synthesized images and the control group of real images. We select the most discriminative description tokens ranked by the p-value, and show the top 10 description tokens for the location categories in Figure 8. In each square grid of Figure 8, the bottom left triangle shows the conditional probability from the synthesized images, and the top right triangle shows the conditional probability from the control group of real images (for comparison). The numbers in the grids represent p-value in the hypothesis testing. The $p < 0.05$ indicates statistical significance, and for ease of visualization, we show the p-value that is less than 0.1 in Figure 8. We observe that users perceive different patterns for similar locations marked in the synthesized and real images. We separate the facial and non-facial location categories (based on Section 5.3) for characterization.

Facial pattern characterization. We find differences in the patterns for locations marked in the synthesized and real images. For facial locations in Figure 8(a), we observe the prevalent description in synthesized images is "`blurry`" (1st column from left). Specifically the regions showing "`blurry`" with statistical significance are "`ear`" ($p < 0.001$) and "`mouth`" ($p < 0.05$). We also find that the pattern "`unnatural skin`" (2nd column) has high correlation with facial regions in synthesized images, including "`ear`" ($p < 0.01$), "`neck`" ($p < 0.01$), and "`chin`" ($p < 0.05$). The observations demonstrate the difficulties to synthesize intricate facial details. In addition, although users may mark same regions in synthesized and real images, such as "`eye`" or "`mouth`" (Section 5.3), we observe pattern differences between synthesized and real images ($p < 0.05$), e.g., "`eye`" in synthesized images appears abnormal "`shape`" (5th column) and "`mouth`" in synthesized images tends to be "`blurry`" (1st column).

Non-facial pattern characterization. The correlations between non-facial regions and descriptive patterns are illustrated in Figure 8(b). Similar to facial regions, we find that the main synthesis pattern is "`blurry`" (1st column from left) correlated with non-facial regions, including "`earring`" ($p < 0.01$), "`clothes`" ($p < 0.01$), and "`eyeglass`" ($p < 0.01$). We also observe specific non-facial patterns of "`unknown object`" (2nd column) with "`clothes`" ($p < 0.05$) and "`mismatched`" (7th column) with "`earring`" ($p < 0.05$), which suggests that the synthesis patterns in non-facial regions at statistically significant rates do not correspond to objects that users can recognize.



(a) Conditional probabilities of pattern descriptions for facial regions (ear, eye, eyebrow, hair, mouth, forehead, neck, chin, cheek, and nose).



(b) Conditional probabilities of pattern descriptions for non-facial regions (hat, earring, eyeglass, background, and clothes).

Figure 8: Conditional probabilities of `Pattern` tokens input by the users correlated with the `Location` categories. In each square, the bottom left triangle shows the probability from the synthesized images, and the top right triangle shows the probability from the control group of real images (for comparison). The plots show the top correlation pairs ordered according to p-value. Significance level is indicated as * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

Takeaway 5: The prevalent artifact pattern in synthesized images is "`blurry`", e.g., correlated with "`ear`", "`eyeglass`", and "`clothes`" regions. In addition, the facial and non-facial regions show other unique synthesis patterns, specifically skin anomalies in facial regions and unknown objects in non-facial regions.

6 Detection of AI-synthesized Faces

With the intelligence of user annotations, we investigate the potential of enhancing the detection of AI-synthesized faces. We use an attention learning approach to incorporate user annotated regions, and evaluate detection accuracy.

6.1 Guided Attention Method for Detection

From our user study, we gather human annotations that collectively locate artifact regions in AI-synthesized face images. We further investigate how to use annotated data to improve detection (**RQ4**). Existing detection of AI-synthesized images typically trains black-box neural networks for classification [13, 56, 57], which have the risk of selecting superficial features. On the other hand, carefully designed heuristic approaches identify explicit features [58, 17, 59], but may suffer from scaling or generalizing to other synthesis methods. We aim to combine the strengths of the two sides. We use the set of human annotated images to guide the detection model and focus on the useful feature regions.

Using Algorithm 1 in Section 5.2, we extract attention mask annotated by users on synthesized images, which ensures that only high agreement regions are used to guide the model learning. As users mark suspicious regions in our study, the extracted mask locates artifacts in AI-synthesized images. With image dimension $H \times W$, the pixel-level annotation mask is represented as $M \in \mathbb{R}^{H \times W}$ (normalized between 0 and 1), where each element indicates the weight of user annotation for an image position. We adopt attention learning to integrate human annotation mask M when training the model. Li et al. [47] developed a method to provide direct guidance on the attention. The model attention map is calculated via Grad-CAM (Gradient-weighted Class Activation Mapping) [60] as $A \in \mathbb{R}^{H \times W}$ (normalized between 0 and 1), whose values highlight the discriminative image regions that the model has learned in training. To enforce extra attention as guidance, a mean squared error loss between A and M is introduced as in Equation (1).

$$\mathcal{L}_{extra}(A,M) = \frac{1}{H \cdot W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} (A_{ij} - M_{ij})^2 \qquad (1)$$

This loss term $\mathcal{L}_{extra}(A,M)$ is integrated with a weighting parameter into the final loss function for training the model. By minimizing the loss, Equation (1) makes the learned attention map A close to the user mask M, which guides the network to pay attention to the user marked regions. We add the loss to provide extra supervision in the training. The attention learning combines the heuristics from human annotations and the features learned from neural networks.

6.2 Detection Experiments and Evaluation

We perform experiments to investigate the effects of adding human annotations to guide attention and compare with other detection methods. The classification neural network that we use is the ConvNeXt [62]. Our training data contains 10,000 synthetic face images (5,000 from StyleGAN2 and 5,000 from StyleGAN3) and 10,000 FFHQ real face images. To provide the extra guidance, we include 200 user annotated images (from our user study experiment) with masks extracted by Algorithm 1. These 200 images are included to train all methods (only our approach uses the human annotation masks). The training data size has similar magnitude and ratio compared to the prior attention learning framework study [47].

To evaluate detection generalization, our testing data spans images generated from a variety of generation methods, with publicly released image sets or pre-trained models to synthesize high-quality images of 1024×1024 resolution. In addition to StyleGAN2 [3] and StyleGAN3 [26] (which also generate the training images in our experiments), we perform testing on earlier and later GAN models, including StyleGAN [32] (predecessor of StyleGAN2), MSG-GAN [33] (successor of StyleGAN2), and Anyres-GAN [34] (successor of StyleGAN3). We also test other models, including StarGAN v2 [35], NVAE [36], and LDM [37]. Each testing experiment consists of 1,000 synthesized face images compared with a fixed set of 1,000 FFHQ real face images.

We run training and evaluation on a NVIDIA TITAN RTX GPU with 24GB of RAM. All images are resized to 224×224, since the ConvNeXt model is pretrained on this resolution [62] and it is close to the practical profile image resolution on social media platforms [63, 64, 65]. In training, we use the Adam optimizer [66] and set the weights of the loss components as 1 for the classification loss and 5 for the guided supervision loss.

Detection accuracy. To evaluate the performance of our method, we compare with the state-of-the-art approaches that were recently developed on AI-generated image detection, including LGrad [30], UniversalFakeDetect [15], and an expert-based detection [17]. The experiments were trained on the same data as our approach. We measure detection accuracy using the F-score [61] and show the evaluation results in Table 2. The rows represent detection approaches. The columns show the evaluation results on images generated by various synthesis models. With the evaluation on StyleGAN2 and StyleGAN3 image sets, our approach of guided attention achieves an F-score over 0.970, outperforming the other approaches (and showing improvement via ablation, more details below). To evaluate generalization across images from other synthesis models (Anyres-GAN, MSG-GAN, StyleGAN, StarGAN v2, NVAE, LDM), our approach yields an overall F-score of 0.927, compared to LGrad with 0.556, UniversalFakeDetect with 0.841, and expert-based detection with 0.835. In addition to improving detection on StyleGAN

Table 2: Breakdown of detection accuracy across the AI-synthesized image testing datasets. The accuracy values show F-scores [61]. *Testing Setup* indicates whether or not evasions were applied to the AI-generated images for testing. *Training Method* denotes what type of training was used for the detection approach. *Same Models as Training* datasets used a train-test split across StyleGAN2 and StyleGAN3 images. *Other StyleGAN Related Models* datasets were testing images generated by other models related to StyleGAN (including both predecessor and successor models). *Other Models* datasets were testing images generated by other synthesis models. The last column shows the arithmetic average across all testing datasets for each detection.

Testing Setup		*Training Method*	*Same Models as Training*		*Other StyleGAN Related Models*			*Other Models*			
			StyleGAN2	StyleGAN3	Anyres-GAN	MSG-GAN	StyleGAN	StarGAN v2	NVAE	LDM	Average
No Evasion		Our approach (attention)	0.973	0.970	0.894	0.939	0.938	0.916	0.937	0.852	0.927
		Attention ablation	0.892	0.895	0.799	0.804	0.813	0.818	0.808	0.756	0.823
		Expert-based detection [17]	0.901	0.898	0.811	0.825	0.833	0.807	0.832	0.776	0.835
		LGrad [30]	0.945	0.872	0.393	0.484	0.450	0.306	0.270	0.724	0.556
		UniversalFakeDetect [15]	0.782	0.865	0.867	0.858	0.844	0.839	0.838	0.838	0.841
Evasion	Blur	Our approach (attention)	0.973	0.970	0.893	0.930	0.938	0.891	0.927	0.798	0.915
		Attention ablation	0.892	0.896	0.807	0.811	0.826	0.798	0.759	0.714	0.813
		Expert-based detection	0.896	0.892	0.794	0.815	0.824	0.837	0.801	0.686	0.818
		LGrad	0.913	0.829	0.303	0.322	0.275	0.138	0.021	0.034	0.354
		UniversalFakeDetect	0.850	0.873	0.803	0.826	0.823	0.839	0.834	0.833	0.835
	Crop	Our approach (attention)	0.974	0.968	0.886	0.938	0.939	0.911	0.942	0.821	0.922
		Attention ablation	0.891	0.897	0.797	0.805	0.816	0.804	0.827	0.726	0.820
		Expert-based detection	0.899	0.898	0.812	0.825	0.835	0.818	0.852	0.704	0.830
		LGrad	0.947	0.890	0.405	0.469	0.472	0.146	0.195	0.012	0.442
		UniversalFakeDetect	0.761	0.848	0.842	0.853	0.824	0.839	0.837	0.826	0.829
	JPEG	Our approach (attention)	0.942	0.923	0.726	0.858	0.872	0.919	0.936	0.829	0.876
		Attention ablation	0.868	0.860	0.724	0.761	0.766	0.817	0.791	0.716	0.788
		Expert-based detection	0.884	0.878	0.752	0.782	0.783	0.816	0.800	0.691	0.798
		LGrad	0.704	0.569	0.155	0.270	0.243	0.000	0.057	0.009	0.251
		UniversalFakeDetect	0.750	0.809	0.728	0.801	0.751	0.835	0.835	0.829	0.792

related images, our approach also generalizes to improve detecting images generated by other synthesis models (StarGAN v2, NVAE, and LDM). Our attention approach achieves statistically significant performance improvements over each of the other approaches ($p < 0.001$ using the two-sample, binomial proportion test). The results show that our approach achieves high detection accuracy and competent generalization performance.

Ablation analysis. To examine the efficacy of incorporating user annotations in detection, we perform an ablation experiment to investigate how much the annotations in training contributes to the detection performance. For the attention ablation setting, we remove the attention component in the loss function and evaluate on the same image sets. Table 2 includes the F-score values of the attention ablation experiment. Without attention in the training, the detection results show lower accuracy, and the overall performance is 0.823. The improvement contribution of attention based on annotations overall is 10.4% (0.823 compared to 0.927). The attention detection with user annotations performs better than the ablation setting with statistical significance ($p<0.001$ using a two-sample, binomial proportion test). The results demonstrate that incorporating annotations considerably improves the detection performance.

Detection robustness. We further examine robustness and evaluate common evasions on the synthesized images, including blurring, cropping, and JPEG compression. We apply evasions only on the synthesized images, not on the real images (i.e., real images are unchanged, since evasions will only be performed by adversaries on synthesized images). For the blurring we use a Gaussian blur with a kernel size of 13 and a standard deviation of 2. For cropping, we randomly crop out up to 10% of the image (i.e., retaining 90% of the image). For JPEG compression, we use a quality level of 75%. Table 2 shows the F-score accuracy under various evasions. Our approach consistently achieves high accuracy across evasions, compared to the other three approaches. LGrad experiences significantly degraded performance since it depends on the gradient pre-processing susceptible to small changes in the pixel values. UniversalFakeDetect uses a predefined feature space, and its accuracy under evasions largely remains lower compared to our approach. Expert-based detection forces the detection to focus on a fixed region based on heuristics and the performance under evasion is still lower than our approach. We observe that detection guided with attention that incorporates human annotations learns inherent features and remains robust against evasions.

Takeaway 6: Attention-guided detection with human annotations achieves high accuracy compared with prior detection approaches, and retains high robustness against evasions.

7 Discussion

We provide discussion based on our findings and experiments, including ethics considerations, suggestions to defend against synthesized faces, and limitations and future work.

Ethics considerations. Our user study was performed with the approval of our institutional review board (IRB). The IRB approval established specific guidelines addressing ethical considerations, such as collecting data securely and obtaining user's consent. We diligently adhered to these guidelines throughout our research to protect the study participants. In our experiments, we carefully avoided collecting any personal identifiable information (PII) from participants and all results are presented in aggregate statistics. We also provided participants options to opt-out of the experiments (we did not receive any opt-out requests). We complied with established ethical norms in our data analysis and result presentation, including securing data storage and maintaining participant anonymity and privacy.

We consider that the value in developing potential defenses against AI-synthesized images outweighs the potential risks, although research into the artifact characteristics of AI-synthesized faces could be used to improve the quality of AI synthesis techniques. Our findings provide higher actionable insights for content moderators than for miscreants. Online moderators can easily operationalize our findings by scrutinizing the suggested regions (more details below) or using attention enhance detection that we describe to identify AI-synthesized images. On the other hand, our identified synthesis artifacts are difficult to rectify for miscreants, since the artifacts are inherent from systemic limitations in AI-synthesized images.

Suggestions for content moderators to recognize AI-synthesized faces. While individual users may have limited capability to distinguish between AI-synthesized content and real images, aggregating across user responses allows for clear identification and localization of artifacts. The aggregated user annotations provide deeper levels of insight into the artifacts by accumulating individual user intelligence. Based on the crowdsourcing insights, our findings suggest regions and patterns to identify AI-synthesized images. First, in AI-synthesized face images, accessories such as earring and eyeglass or specific facial details such as ear and hair likely show anomalies. Second, anomaly patterns in AI-synthesized images are typically misshaped boundaries, unnatural skin, and unrecognizable objects, which are different from suspicious attributes in real images. Content moderators can focus on these regions and patterns that we summarize when screening large numbers of suspicious images to improve the efficiency of identifying AI-synthesized faces. Our findings can also help to educate general users on typical attributes of AI-synthesized faces to avoid being deceived.

Deployment scenarios. Our approach can be deployed in various ways to assist stakeholders defending against the abuse of synthesized images. (1) Our work of user annotations provides quantitative findings about artifacts in AI-synthesized images. Expert researchers can use our approach to study quantitatively specific synthesis areas for rigorous analysis, in addition to applying heuristics. (2) Content platforms can deploy our approach and leverage collective annotation intelligence from the moderation team to build effective detection systems with attention learning. This will mitigate disinformation propagation and AI-based deception on the platforms. (3) Our approach can also be incorporated with other detection approaches, to enable a multi-layered defense strategy. Longitudinal deployment will facilitate continual analysis and detection on AI-synthesized images.

User's trust in real images. Traditionally, the user's assumption for facial images has been that the faces are real, but the continued growth in AI-synthesized content threatens to challenge users' trust. We included a control group of real images in our experiments. While participants can distinguish between AI-generated content and real face images, many still choose the level of "`looks real`" instead of "`very real`". By checking real images with lower trust scores, we find these images contain less common elements such as unusual clothing, dyed hair, or out of the ordinary poses or expressions. The decreased trust of real images with unfamiliar contents suggests users may first begin to suspect content which does not conform to their common definition. The findings are in line with prior work [49, 24] on online user trust. The effect on user's trust in real images needs additional efforts to counter the negative impact brought by AI-synthesized images especially for marginalized groups.

Limitations and future work. A limitation is we focus on face images synthesized by AI techniques. Face images are often used to establish trust and abused by miscreants to deceive users. With attacks expanding, other non-facial images may also present potential misuse in security areas such as product review spam or disinformation. Our approach can be adapted to analyze other categories of synthesized images. Another limitation is our analysis is mainly on GAN models (given its wide abuse in practice), though we include multiple models in the user study and detection experiments. A prospective work direction is to expand analysis on future or more synthesis techniques, as new synthesis models are developed. We anticipate our analysis approach can be applied to find more synthesis characteristics. A potential extension of our work is to use our human-in-the-loop approach to continue adapting to other modalities such as audio or video.

We acknowledge that our results may not fully transfer to

an unprompted setting. We provided brief prompts to allow participants to understand the task, as our goal is to use crowdsourcing characterizing AI-synthesis artifacts. Research to investigate user studies identifying AI-synthesized images under different types of prompting is a promising future research area.

8 Related Work

We describe previous studies on detection of AI-synthesized images and user studies for synthesized face images, and highlight the differences from our work.

Detection and identified artifacts of AI-synthesized images. Existing studies to detect AI-synthesized images and artifacts mainly rely on two strategies, black-box classification or specific pattern heuristics. Classification-based approaches [11, 12, 67, 15] typically train another deep neural network on real and synthesized images, which performs automated learning for detection. A group of work [13, 68, 69, 70] attempted to augment images with different evasions during training to improve the classifier's robustness. Other work investigated using an ensemble of classifiers [14, 71] to supplement individual model's performance. Finally, another body of work explored other feature spaces such as different color representations [72], frequency domain [73], or gradient [30]. However, these classification approaches for detection risks learning superficial features [16].

Heuristic-based approaches used a wide range of specific observations or assumptions about synthesis artifacts such as color, frequency, or facial features, but are limited to the individual perception or heuristics of researchers. Prior studies [59, 74, 75] examined inconsistent colors in synthesized images. Some studies [76, 77, 78, 56, 79] derived signals from noise patterns and frequency analysis generated by GANs. Another line of prior work [58, 17] suggested specific facial parts as features to distinguish AI-synthesized face images.

In contrast, our work complements prior detection and analysis. We use crowdsourcing intelligence to systematically find and characterize AI-synthesis artifacts annotated by multiple users (instead of by individual researchers or automated systems). Our findings reveal new characteristics in AI-synthesized images, and our detection with attention guidance combines strengths of heuristics from human annotations and automated detection from neural network classification.

User studies on AI-synthesized face images. Recent research has conducted user studies to understand users' behaviors and trust on AI-synthesized faces. Mink et al. [49] established social network profiles with AI-synthesized portraits and performed user studies to investigate what factors affect users' trust. Guo [25] proposed alternate analysis platforms for evaluating human performance on identifying AI-synthesized faces. Other work [23, 24, 80] analyzed whether users can distinguish between AI-synthesized and real faces. These studies inspire our work. However, previous work has focused on the victim's perspective and used relatively simplistic designs.

Our work, on the other hand, quantitatively examines where and how users perceive artifacts and is the first to leverage user perception as means to detect and characterize AI-synthesized face images. Our user study designs to use crowdsourcing annotations to locate and examine suspicious regions in synthesized images, with the main goal to characterize artifacts in AI-synthesized images. Based on aggregated human annotations, we extract characteristics of synthesis artifacts and demonstrate the potential of involving human factors to defend against AI-synthesized faces.

9 Conclusion

In this paper, we develop a novel approach that leverages crowdsourcing annotations to systematically characterize and recognize AI-synthesized face images. Existing detection and analysis mostly relies on black-box classifiers or heuristics of researchers' individual perception. In contrast, we design a user study to aggregate and characterize AI-synthesis artifacts annotated by multiple users (rather than individual researchers or automated systems). Based on quantitative results, we find that facial regions distant from the center (such as ear and hair) and non-facial regions are more likely to exhibit synthesis artifacts. The prevalent pattern of artifacts in synthesized images is blur. Our findings provide empirical insights for online moderators or general users to recognize AI-synthesized faces. Furthermore, we incorporate user annotated regions into an attention learning approach to detect AI-synthesized faces. Evaluation comparing with prior approaches shows that our approach achieves high detection accuracy and remains robust under evasions. Our results demonstrate the human-in-the-loop potential to defend against AI-synthesized images.

Acknowledgments

We thank the anonymous reviewers and shepherd for their valuable comments to improve the paper. This work was supported by the National Science Foundation (NSF) under Award No. 2146448. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[2] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the 33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative Image Inpainting with Contextual Attention. In *Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Nikolay Jetchev and Urs Bergmann. The Conditional Analogy GAN: Swapping Fashion Articles on People Images. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[6] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose Guided Human Video Generation. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, 2018.

[7] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the 2018 Internet Measurement Conference (IMC)*, 2018.

[8] CNET. Spy Reportedly Used AI-generated Photo to Connect with Targets on LinkedIn. https://www.cnet.com/science/spy-reportedly-used-ai-generated-photo-to-connect-with-targets-on-linkedin/, 2019.

[9] CNN. A High School Student Created A Fake 2020 Candidate. Twitter Verified It. https://www.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html, 2020.

[10] National Public Radio (NPR). AI-Generated Fake Faces Have Become A Hallmark of Online Influence Operations. https://www.npr.org/2022/12/15/1143114122/ai-generated-fake-faces-have-become-a-hallmark-of-online-influence-operations, 2022.

[11] Huaxiao Mo, Bolin Chen, and Weiqi Luo. Fake Faces Identification via Convolutional Neural Network. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2018.

[12] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. Learning to Detect Fake Face Images in the Wild. In *Proceedings of the International Symposium on Computer, Consumer and Control (IS3C)*, 2018.

[13] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images are Surprisingly Easy to Spot... For Now. In *Proceedings of the 33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Sara Mandelli, Nicolo Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting Gan-Generated Images by Orthogonal Training of Multiple CNNs. In *the IEEE Conference on Image Processing (ICIP)*, 2022.

[15] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *Proceedings of the 36th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[16] Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann, Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, vol. 2, pp. 665–673, 2020.

[17] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing GAN-generated Faces Using Inconsistent Corneal Specular Highlights. In *the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

[18] Hany Farid, Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust & Safety*, vol. 1, no. 4, 2022.

[19] PYMNTS. LinkedIn Faces Flood of AI-Generated Fake Profiles. https://www.pymnts.com/news/security-and-risk/2022/linkedin-faces-flood-of-ai-generated-fake-profiles/, 2022.

[20] The New York Times. Facebook Discovers Fakes That Show Evolution of Disinformation. https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html, 2019.

[21] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell, Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.

[22] Truett Allison, Heidi Ginter, Gregory McCarthy, Anna C. Nobre, Aina Puce, Marie Luby, and D. Dennis Spencer, Face Recognition in Human Extrastriate Cortex. *Journal of Neurophysiology*, vol. 71, no. 2, pp. 821–825, 1994.

[23] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the 22nd IEEE Conference on Computer Vision (ICCV)*, 2019.

[24] Sophie J. Nightingale, Shruti Agarwal, Erik Harkonen, Jaakko Lehtinen, and Hany Farid, Synthetic Faces: How Perceptually Convincing Are They? *Journal of Vision*, vol. 21, no. 9, 2021.

[25] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Open-Eye: An Open Platform to Study Human Performance on Identifying AI-Synthesized Faces. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2022.

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NIPS)*, 2021.

[27] StyleGAN2 GitHub. https://github.com/NVlabs/stylegan2, 2020.

[28] StyleGAN3 GitHub. https://github.com/NVlabs/stylegan3, 2021.

[29] Meta. Recapping Our 2022 Coordinated Inauthentic Behavior Enforcements. https://about.fb.com/news/2022/12/metas-2022-coordinated-inauthentic-behavior-enforcements/, 2022.

[30] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the 36th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS)*, 2014.

[32] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Animesh Karnewar and Oliver Wang. MSG-GAN: Multi-Scale Gradients for Generative Adversarial Networks. In *Proceedings of the 33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution Training for High-resolution Image Synthesis. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 2022.

[35] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the 33rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS)*, 2020.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the 35th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[38] Marcus N. Morrisey, Ruth Hofrichter, and M. D. Rutherford, Human Faces Capture Attention and Attract First Saccades Without Longer Fixation. *Visual Cognition*, vol. 27, no. 2, pp. 158–170, 2019.

[39] Joseph B. Walther, Celeste L. SLovacek, and Lisa C. Tidwell, Is a Picture Worth a Thousand Words?: Photographic Images in Long-Term and Short-Term Computer-Mediated Communication. *Communication Research*, 2001.

[40] Hootan Rashtian, Yazan Boshmaf, Pooya Jaferian, and Konstantin Beznosov. To Befriend Or Not? A Model of Friend Request Acceptance on Facebook. In *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS)*, 2014.

[41] Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram. In *Proceedings of the 2014 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2014.

[42] Yue Nancy Dai, Gregory Viken, Eunsin Joo, and Gary Bente, Risk Assessment in E-commerce: How Sellers' Photos, Reputation Scores, and the Stake of A Transaction Influence Buyers' Purchase Behavior and Information Processing. *Computers in Human Behavior*, vol. 84, pp. 342–351, 2018.

[43] Snehasish Banerjee, Monica Lens, and Anjan Pal, Put on Your Sunglasses and Smile: The Secret of Airbnb hosts' Profile Photos? *International Journal of Hospitality Management*, vol. 103, 2022.

[44] Marijn Stollenga, Jonathan Masci, Faustino Gomez, and Juergen Schmidhuber. Deep Networks with Internal Selective Attention through Feedback Connections. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS)*, 2014.

[45] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual Attention Network for Image Classification. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[47] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell Me Where to Look: Guided Attention Inference Network. In *Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[48] Rensis Likert, A Technique for the Measurement of Attitudes. *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[49] Jaron Mink, Licheng Luo, Nata M. Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security)*, 2022.

[50] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti, Reputation as A Sufficient Condition for Data Quality on Amazon Mechanical Turk. *Behavior Research Methods*, vol. 46, pp. 1023–1031, 2014.

[51] Jacob Cohen. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, 1988.

[52] Kesheng Wu, Ekow Otoo, and Arie Shoshani. Optimizing Connected Component Labeling Algorithms. In *Proceedings of SPIE Medical Imaging Conference*, 2005.

[53] Princeton University. WordNet. https://wordnet.princeton.edu/, 2010.

[54] Annika Tjuka. A List of 171 Body Part Concepts. Zenodo. https://doi.org/10.5281/zenodo.4058506, 2020.

[55] Ronald Walpole, Raymond Myers, Sharon Myers, and Keying Ye. Probability & Statistics for Engineers & Scientists. Pearson, 2017.

[56] Joel Frank, Thorsten Eisenhofer, Lea Schonherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[57] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and Simulating Artifacts in GAN Fake Images. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019.

[58] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu. Exposing GAN-synthesized Faces Using Landmark Locations. In *Proceedings of ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2019.

[59] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, Detection of Deep Network Generated Images Using Disparities in Color Components. *Signal Processing*, vol. 174, 2020.

[60] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, vol. 128, pp. 336–359, 2017.

[61] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

[62] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the 35th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[63] LinkedIn. 10 Tips for Taking a Professional LinkedIn Profile Photo. https://www.linkedin.com/business/talent/blog/product-tips/tips-for-taking-professional-linkedin-profile-pictures/, 2022.

[64] Facebook. Pages Profile Picture and Cover Photo Dimensions. https://www.facebook.com/help/125379114252045/, 2023.

[65] Twitter. Help With Uploading A Profile Photo. https://help.twitter.com/en/managing-your-account/common-issues-when-uploading-profile-photo, 2023.

[66] Diederik P Kingma and Jimmy Ba, Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[67] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of GAN-Generated Fake Images over Social Networks. In *Proceedings of the 1st IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018.

[68] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Training CNNs in Presence of JPEG Compression: Multimedia Forensics vs Computer Vision. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020.

[69] Chengrui Wang and Weihong Deng. Representative Forgery Mining for Fake Face Detection. In *Proceedings of the 34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[70] Farkhund Iqbal, Ahmed Abbasi, Abdul Rehman Javed, Ahmad Almadhor, Zunera Jalil, Sajid Anwar, and Imad Rida, Data Augmentation-Based Novel Deep Learning Method for Deepfaked Images Detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2023.

[71] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S. Woo. Detecting Both Machine and Human Created Fake Face Images In the Wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security (MPS)*, 2018.

[72] Farkhund Iqbal, Ahmed Abbasi, Abdul Rehman Javed, Ahmad Almadhor, Zunera Jalil, Sajid Anwar, and Imad Rida, CSC-Net: Cross-color Spatial Co-occurrence Matrix Network for Detecting Synthesized Fake Images. *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 1, pp. 369–379, 2023.

[73] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Pyounggeon Kim, and Jongwon Choi. FingerprintNet: Synthesized Fingerprints For Generated Image Detection. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 2022.

[74] Scott McCloskey and Michael Albright. Detecting GAN-Generated Imagery Using Saturation Cues. In *Proceedings of the 26nd IEEE Conference on Image Processing (ICIP)*, 2019.

[75] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Alexander Binder, and Ngai-Man Cheung. Discovering Transferable Forensic Features for CNN-generated Images Detection. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 2022.

[76] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs Leave Artificial Fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.

[77] Ning Yu, Larry Davis, and Mario Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *Proceedings of the 22nd IEEE Conference on Computer Vision (ICCV)*, 2019.

[78] Matthew Joslin and Shuang Hao. Attributing and Detecting Fake Images Generated by Known GANs. In *Proceedings of the 3rd Deep Learning and Security Workshop (DLS)*, 2020.

[79] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In *Proceedings of the 34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[80] 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, 2022.

A Statistics on Images

To examine the distribution of the images, we manually labeled the gender and ethnicity of the faces. While the images were randomly sampled, the gender and ethnicity show similar trend. For the gender of images, 53% of synthesized images (54% of real images) were female and 47% of synthesized images (46% of real images) were male. For the ethnicity of images, 67% of synthesized images (62% of real images) were White, 4% of synthesized images (4% of real images) were Hispanic or Latino, 5% of synthesized images (3% of real images) were African American, 11% of synthesized images (20% of real images) were Asian, and 13% of synthesized images (11% of real images) were uncertain.

B Additional Statistics on Participants

We aggregate the statistics based on participants who answered demographic questions optionally. The optional demographic questions include the participant's age range, education, gender, and ethnicity. The ages of the participants were in a range of 20 to 64 with a median range of 35-39. Most participants had post-secondary education with 77% having received a bachelor degree or above. For the gender of participants, 44.85% were female and 55.15% were male (we provided the non-binary option and the option was not selected). For the ethnicity of participants, 79.46% were White, 2.70% were Hispanic or Latino, 2.16% were African American, 2.16% were Asian, and 13.52% were uncertain.