5-1-2022

Exposing GAN-generated faces using deep neural network

Hui Guo

University at Albany, State University of New York, huiguo.cv@gmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Computer Sciences Commons

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/2917

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
EXPOSING GAN-GENERATED FACES USING DEEP NEURAL NETWORK

by

Hui Guo

A Dissertation
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

College of Engineering and Applied Sciences
Department of Computer Science
May 2022
To my family for their unconditional love and support!
ABSTRACT

Generative adversarial network (GAN) generated high-realistic human faces are visually challenging to discern from real ones. They have been used as profile images for fake social media accounts, which leads to high negative social impacts.

In this work, we explore a universal physiological cue of the eye, namely the pupil shape consistency, to reliably identify GAN-generated faces. We show that GAN-generated faces can be exposed via irregular pupil shapes. This phenomenon is caused by the lack of physiological constraints in the GAN models. We demonstrate that such artifacts exist widely in high-quality GAN-generated faces. We design an automatic method to segment the pupils from the eyes and analyze their shapes to distinguish GAN-generated faces from real ones.

Furthermore, we propose a robust, attentive, end-to-end framework that spots GAN-generated faces by analyzing iris regions. The framework can automatically localize and compare artifacts between iris to identify GAN-generated faces. Once Mask-RCNN extracts the iris regions, a Residual Attention Network (RAN) is used to examine the components between the two iris. Besides, we use a joint loss function combining the traditional cross-entropy loss with a relaxation of the ROC-AUC loss via WMW statistics to improve the deep neural network learning from imbalanced data. Comprehensive evaluations demonstrate the superiority of the proposed methods.
ACKNOWLEDGMENT

I am very grateful to have worked with wonderful people during my Ph.D. study.

First and foremost, I would like to thank my advisor Dr. Ming-Ching Chang for being a great mentor in introducing me to the wonderful research areas and encouraging me to achieve research accomplishments. I was impressed with his broad knowledge and responsible attitude. He has taught me the methodology to carry out the research and present the research works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

I appreciate the dissertation committee, Dr. Siwei Lyu, Dr. Pradeep K. Atrey, and Dr. Shaghayegh Sahebi, for their insightful comments and suggestions. Their constructive suggestions and comments helped me refine my thinking about research.

Thanks to the great instructions from many excellent professors: Dr. Paliath Narendran, Dr. Dan Willard, Dr. Neil Murray, and all professors in this department.

I am incredibly grateful to my parents for their love and care. I love you both a lot and appreciate your effort and passion in bringing me up to be a better individual. I am much lucky to have you both as my parents and thank God for giving me you.

Thanks to my wonderful and loving family: your unconditional love carried me through my confused days. Thank you for being there.

I would also like to thank my friends: Yiming Yang, Tianchi Zhang, Shu Hu, and Yan Hu. You have made my life so much better because of your friendships. Thank you for being my friends.
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENT</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>xi</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Problem Statement</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>1</td>
</tr>
<tr>
<td>1.2 The Proposed Approach</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Physiological-based Method</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 DL-based Method</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2.1 Data imbalance</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Contributions</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>5</td>
</tr>
<tr>
<td>2. Related Works</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Generative Adversarial Network (GAN)</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>7</td>
</tr>
<tr>
<td>2.2 GAN Generation of Highly Realistic Faces</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>8</td>
</tr>
<tr>
<td>2.3 GAN-face Detection Methods</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>9</td>
</tr>
<tr>
<td>2.3.1 Deep Learning-based Methods</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>9</td>
</tr>
<tr>
<td>2.3.1.1 GAN-face detection in real-world scenarios</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>10</td>
</tr>
<tr>
<td>2.3.1.2 One-shot, incremental and advanced learning</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>10</td>
</tr>
<tr>
<td>2.3.1.3 Difficulty Analysis</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>10</td>
</tr>
<tr>
<td>2.3.2 Physical-based Methods</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>10</td>
</tr>
<tr>
<td>2.3.3 Physiological-based Methods</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>11</td>
</tr>
<tr>
<td>2.3.4 Human Visual Performance</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>12</td>
</tr>
<tr>
<td>2.4 Datasets and Performance Evaluation</td>
<td>. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .</td>
<td>14</td>
</tr>
</tbody>
</table>
## 3. Exposing GAN Faces Using Irregular Pupil Shape

### 3.1 Iris and pupil segmentation

### 3.2 Pupil Segmentation and Boundary Detection

### 3.3 Ellipse Fitting to the Pupil Boundary

### 3.4 Estimating the Pupil Shape Irregularity

### 3.5 Experiments

#### 3.5.1 Datasets

#### 3.5.2 Results

##### 3.5.2.1 Sensitivity Analysis of $d$

##### 3.5.2.2 Limitations

### 3.6 Summary

## 4. Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces

### 4.1 Background

#### 4.1.1 Attention Mechanism

#### 4.1.2 Imbalanced Data Learning

### 4.2 Facial Landmark Localization and Iris Segmentation

### 4.3 Residual Attention Network

### 4.4 AUC of ROC for Classification Evaluation

### 4.5 WMW AUC Relaxation for Loss Design

### 4.6 Experiment

### 4.7 The New FFHQ-GAN Dataset

### 4.8 Implementation Details

#### 4.8.0.1 Hyper-parameters

### 4.9 Evaluation on the FFHQ-GAN dataset

#### 4.9.0.1 Results on the balanced and imbalanced set

### 4.10 Ablation Studies
LIST OF FIGURES

1.1 StyleGAN2 [59] generated faces are highly realistic and can be easily abused for malicious purposes. Effective forensic methods for identifying them is of strong needs. .............................................................. 2

1.2 **Top:** Anatomy of a human eye: the iris and pupil are at the center surrounded by the sclera. **Bottom:** Examples of pupils of real human (left) and GAN-generated (right). Note that the pupils for the real eyes are in circular or elliptical shapes (**yellow**), while those for the GAN-generated pupils are much irregular (**red**). For GAN-generated faces, the shapes of the pupils are very different from each other when zoomed-in. .............................................................. 3

2.1 Examples of GAN faces generated by StyleGAN (left), StyleGAN2 (middle), and StyleGAN3 (right). .............................................................. 7

2.2 Architecture of GAN. .............................................................. 8

3.1 Examples of GAN-synthesized faces additional to StyleGAN and StyleGAN2. The images are from their original papers (a) PGGAN [54], (b,c) Alias-Free GAN (StyleGAN3) [56], (d) SofGAN [14]. Observe in the zoomed-in view that the pupils appear in irregular, inconsistent shapes, which tell them apart from real faces. .............................................................. 17

3.2 The proposed pipeline for face detection, facial landmark localization, pupil segmentation, and pupil ellipse fitting. (a) The input high-resolution face image, (b) The cropped eye image using landmarks, (c) The predicted pupil mask of the eye from (b), (d) The fitted ellipse mask. This example shows a GAN-generated face. .............................................................. 18

3.3 A schematic example explaining the Boundary IoU. **Left:** The predicted pupil mask P and the ellipse fitted pupil mask F. **Middle:** P_d and F_d are the mask pixels within distance d from the boundaries (blue and yellow). **Right:** Boundary IoU calculation between predicted pupil mask and the ellipse fitted pupil mask with distance parameter d. .............................................................. 19

3.4 Comparing of boundary IoU for real pupil and fake pupil .............................................................. 20

3.5 **Top:** Distributions of the BIoU scores (of the averages of both pupils) for the real and GAN-generated faces. **Bottom Left:** The ROC curve based on the BIoU with d = 4. **Bottom Right:** BIoU hyper-parameter analysis, where the x-axis indicates distance parameter d, and y-axis indicates the AUC score. .. 21
Examples of both eyes from the real human faces (Top) and GAN generated human faces (Bottom). The pixels of the predicted pupil mask within distance $d = 4$ from the prediction boundary contours are highlighted. The BIoU scores with $d = 4$ between the predicted pupil mask and the ellipse-fitted one are shown on each image.

Top: Examples of diseased and infection eyes from [85]. These pupils from images of real faces contain abnormal non-elliptical pupil shapes, which only occurs rarely in real life. Bottom: Occlusions and environmental variations can cause pupil segmentation failure.

More pair of pupil examples from real (left) human faces and GAN-generated (right) faces. As we mentioned before, the irregular pupil shape is a good sign for the human to expose the GAN-generated face visually, even there are no boundary labels around the pupils, we can easily see that the shapes of GAN-generated pupils are very irregular, and the shapes of both pupils are very different in the same GAN-generated face image. In practice, people can zoom a face image large enough and then check the pupil shapes to find whether the face is real or not easily.

The extracted iris pairs of our method for the (a) GAN-generated and (b) real faces. Artifacts of inconsistent corneal specular highlights are obvious in GAN-generated iris pairs.

Details of our Attention Module from the our RAN in Figure 4.3. Our design is inspired from the residual attention network of [101].

ROC of the proposed method, ResNet with BCE loss, Xception with BCE loss, and RAN with BCE loss.

PR curve of the proposed method, ResNet with BCE loss, Xception with BCE loss, and RAN with BCE loss.

Confusion Matrix on the FFHQ-GAN (Top) balanced and (Bottom) imbalanced datasets.

Impact of hyperparameter $\alpha$ of the AUC loss in Eq.(4.4) for the GAN-generated face detection on the imbalanced dataset.
4.7 Visualization of the **extracted iris pairs** and the corresponding **attention maps** obtained from our Residual Attention Network (RAN). Observe that the attention maps for GAN-generated faces better focus on the artifacts such as the corneal specular highlights, while the attention maps for real faces are widely distributed. This shows the effective learning of RAN for identifying GAN-generated faces.

4.8 Examples of **detected GAN-generated faces** and their corresponding iris regions and the attention maps produced from our method. These examples show that our method can detect a wide range of face images, including those with tilted or side views where both irises are visible.

5.1 *The workflow of Open-eye.* The participants test 20 face images, including the same amount of real and fake. A quick tutorial is demonstrated to participants to learn how to recognize the specific artifacts in the eye. The same test will be used again to show the human performance of the AI-synthesized faces detection.

5.2 Different Stages of the platform

5.3 *Tutorial (Stage 2).* **Left:** Anatomy of a human eye: the iris and pupil are at the center surrounded by the sclera. **Middle:** Examples of pupils of real human (left) and GAN-generated (right). Note that the pupils for the real eyes are in circular or elliptical shapes (yellow), while those for the GAN-generated pupils are much irregular (red). For GAN-generated faces, the shapes of the pupils are very different from each other when zoomed-in. **Right:** More pair of pupil examples from real (left) human faces and GAN-generated (right) faces. As we mentioned before, the irregular pupil shape is a good sign for the human to expose the GAN-generated face visually, we can easily see that the shapes of GAN-generated pupils are very irregular, and the shapes of both pupils are very different in the same GAN-generated face image. In practice, people can zoom a face image large enough and then check the pupil shapes to find whether the face is real or not easily.
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Summary of GAN-face detection methods with the corresponding datasets, statistics and performance scores. The green rows highlight those where individual predicted results of the method are explainable to humans. Note that datasets used in the works are self-collected and can contain different subsets across papers. So the performance scores do not represent fair comparisons.</td>
<td>13</td>
</tr>
<tr>
<td>4.1</td>
<td>Details of the proposed Residual Attention Network (RAN) in the right of Figure 4.3.</td>
<td>28</td>
</tr>
<tr>
<td>4.2</td>
<td>The proposed pipeline for training the Residual Attention Network (RAN) on possibly imbalanced data for GAN-generated face classification. The extracted iris pairs are passed as input to the RAN. A robust loss function derived from maximizing the AUC of ROC is optimized in the training of the RAN. See details in S 4.5.</td>
<td>30</td>
</tr>
<tr>
<td>4.3</td>
<td>The proposed architecture for GAN-generated face detection. We first use DLib [60] to detect faces and localize eyes, and use Mask R-CNN [38] to segment out the iris regions. A Residual Attention Network (RAN) then performs binary classification on the extracted iris pair to determine if the face is real or fake. The training is carried out using a joint loss combining the Binary Cross-Entropy (BCE) loss and the ROC-AUC loss with WMW relaxation to better handle the learning from imbalanced data (see text).</td>
<td>32</td>
</tr>
<tr>
<td>4.4</td>
<td>Details of the FFHQ-GAN dataset regarding its balanced (-b) and imbalanced (-imb) subsets.</td>
<td>35</td>
</tr>
<tr>
<td>4.5</td>
<td>Results on the FFHQ-GAN dataset regarding its balanced (-b) and imbalanced (-imb) subsets.</td>
<td>36</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

1.1 Problem Statement

The development of Generative Adversarial Networks (GANs) [32] has led to a dramatic increase in the realism in generating high-quality face images, including PGGAN [54], StyleGAN [58], StyleGAN2 [59], and StyleGAN3 [57]. As illustrated in Figure 1.1, these GAN generated (or synthesized) fake faces are difficult to distinguish from human eyes. Such synthesized faces are easily generatable, can be directly leveraged for disinformation, and potentially lead to profound social, security, and ethical concerns. The GAN-generated faces can be easily abused for malicious purposes, such as creating fake social media accounts to lure or deceive unaware users [3, 1, 2, 4], which can cause significant security problems and frauds. Therefore, the authentication of GAN-generated faces has obtained increasing importance in recent years. However, there exists only a paucity of forensic techniques that can effectively detect such fake faces.

Many studies employ CNNs or other classifiers to distinguish the GAN-generated faces from the real ones [111, 72, 76, 25, 102, 15]. Although these methods detect various GAN-generated faces with relatively high accuracy, similar to other deep learning-based techniques, they suffer from poor generalization and lack interpretability of detection results. Physiology-based methods [74, 111, 36] detect fake faces by examining the semantic aspects of human faces, including physiological or shape-related cues such as symmetry, iris color, and pupil shapes. The prior work of an explainable physical method in [42] addressed some of the above limitations, where GAN-generated faces are identified based on a rule-based decision over the inconsistencies of the specular eye patterns. However, this method relies on assumptions of a frontal face as input and the existence of far away lighting reflection source(s) from both eyes. When these assumptions are violated, generalization will be limited and false positives may rise significantly.
Figure 1.1: StyleGAN2 [59] generated faces are highly realistic and can be easily abused for malicious purposes. Effective forensic methods for identifying them is of strong needs.

1.2 The Proposed Approach

1.2.1 Physiological-based Method

We explore a universal physiological cue of the eye, namely the pupil shape consistency, to reliably identify GAN-synthesized faces. As shown in Figure 1.2, the eye is one of the few organs in the human body that is highly circular and regular in geometry. We hypothesize that the human iris and pupil can provide rich physical and physiological cues that can improve GAN-synthesized face detection.

Our method is based on a simple physiological assumption that human pupils should be nearly circular in their shapes in a face image. Due to different facial orientations and camera angles, the actual pupil shapes can be elliptical. Our observation is that this simple
Figure 1.2: **Top:** Anatomy of a human eye: the iris and pupil are at the center surrounded by the sclera. **Bottom:** Examples of pupils of real human (left) and GAN-generated (right). Note that the pupils for the real eyes are in circular or elliptical shapes (yellow), while those for the GAN-generated pupils are much irregular (red). For GAN-generated faces, the shapes of the pupils are very different from each other when zoomed-in.

property is not well preserved in the existing GAN models, including StyleGAN2 [59], the state-of-the-art face synthesis method. The pupils for the StyleGAN-generated faces tend to have non-elliptical shapes with irregular boundaries. Figure 1.2 shows an example with zoom-in views of the pupils. Such artifacts in the GAN-generated faces are due to the difficulty or negligence of physiological constraints on human anatomy when training the GAN models via standard data-driven machine learning.

The proposed GAN-generated face detector consists of several automatic steps. We first segment the pupil regions of the eyes and extract their boundaries automatically. We next fit an ellipse parametric model to each pupil, and calculate the Boundary Interaction-
over-Union (BIoU) scores [20] between the predicted pupil mask and the ellipse-fitted model. The BIoU score provides a quantitative measure of the regularity of the pupil shape, that determines if the eyes (and the face) are real or not. Experiments are conducted on a dataset containing both real and machine-synthesized faces. The results show that there is a clear separation between the distributions of the BIoU scores of the real and GAN-generated faces.

1.2.2 DL-based Method

We improve the prior method and develop an end-to-end approach for detecting GAN-generated faces by examining the inconsistencies of the two eyes. We first use Mask R-CNN [38] to detect and localize the iris regions. Instead of segmenting the corneal specular highlights using low-level image processing methods in [42], we design a Residual Attention Network which consists of residual attention blocks inspired from [101], to automatically learn to localize the inconsistencies. Our new method is data-driven and can better spot inconsistent artifacts, including but not limited to the corneal specular highlights.

1.2.2.1 Data imbalance

Data imbalance is an important issue that is less addressed in existing GAN-generated face detection works. In real-world use scenarios of face examination, real face images usually outnumber GAN-generated ones by a large amount. Imbalanced data lead to learning problems and thus affect model design. It is well-known the widely-used cross-entropy loss [78] is not suitable for classifying imbalanced data. Although substantial progress is made by sampling [94], adjusting of class weights, data enhancement [26], etc., learning with imbalanced data is still challenging. It is intuitive that the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) plot can be incorporated as a loss term to improve classification performance [84]. However, the AUC is a pairwise rank-based metric with discontinuous values among iterations. Therefore, AUC is not directly applicable for loss design for end-to-end optimization of the classifier. To this end, we incorporate a ROC-AUC loss term by maximizing the Wilcoxon-Mann-Whitney (WMW) statistics of the ROC, which is shown to provide similar effects in approximating the AUC optimization [110, 70]. Our experimental results show that a combination of the binary cross-entropy loss and the WMW-AUC loss leads to the best end-to-end result.
We perform experiments on two data sources: (1) real human face images obtained from the Flickr-Faces-HQ (FFHQ) dataset [58] and (2) GAN-generated face images available from http://thispersondoesnotexist.com as shown in Figure 1.1. Experiment results demonstrate the superiority of the proposed method in distinguishing GAN-generated faces from the real ones.

1.3 Contributions

The main contributions of this thesis are summarised as following:

- We are the first to propose the idea of exploiting pupil shape consistency as an effective way to distinguish fake faces from real ones. This new cue is effective for humans as well to visually identify GAN-generated faces.
- Our proposed method is based on explainable physiological cues, which is simple, effective, and explainable. Evaluations on the Flickr-Faces-HQ dataset and an in-house collected StyleGAN2 face dataset show its effectiveness and computational efficiency.
- We also propose an end-to-end method for detecting GAN-generated faces. A residual attention network model is incorporated to better focus on the inconsistencies of the eyes e.g. corneal specular highlights and other artifacts. Our end-to-end fake face detection method is interpretable, and the proposed cues can be leveraged by human beings as well to perform examinations.
- We create a new FFHQ-GAN dataset by combining portions of the FFHQ real faces with the StyleGAN2 generated images. Performance of GAN-generated face detection is evaluated on this FFHQ-GAN dataset for both balanced and imbalanced data conditions.
- We use the WMW-AUC loss that approximates the direct optimization of the AUC. This can also effectively address the data imbalance learning problem in contrast to other sampling or data augmentation approaches. Experimental results show that our method achieves plausible performance especially on imbalanced datasets.

This thesis is organized as follows. Chapter 2 summarizes related works on GAN-generated faces methods and detection solutions. Chapter 3 describes the proposed method that uses irregular pupil shapes to detect the GAN-Generated faces. Chapter 4 describes the proposed end-to-end framework to detect the GAN-Generated faces. Chapter 5 introduces an
online platform Open-eye to study the human performance of AI-synthesized faces detection. Chapter 6 concludes the thesis and discuss the future works.
CHAPTER 2
Related Works

We review related works on the Generative Adversarial Network (GAN) models and GAN-generated face detection methods.

![Examples of GAN faces generated by StyleGAN (left), StyleGAN2 (middle), and StyleGAN3 (right).](image)

Figure 2.1: *Examples of GAN faces generated by StyleGAN (left), StyleGAN2 (middle), and StyleGAN3 (right).*

2.1 Generative Adversarial Network (GAN)

A GAN model comprises two neural networks, a generator, and a discriminator, trained alternatively. The generator passes the random noise and synthesizes an image. The discriminator works as a classifier to distinguish the real images from the fake images. During the optimization, the generator synthesizes images to fool the discriminator, and the discriminator gets worse at recognizing the difference between real and fake.

Formally, the generator learns to map a noise $z$ from a prior distribution $p(z)$ to a real images distribution $p(x)$. The discriminator $D(\cdot)$ is trained to discriminate whether a sample
is from the generator’s output distribution $G(z)$, $z \sim p(z)$, or from target distribution $p(x)$. The generator $G$ is continuously trained to generate images that are more likely to be from $p(x)$. The training is uninterrupted until the generator and the discriminator are optimized to have the best performance (see Figure. 2.2).

2.2 GAN Generation of Highly Realistic Faces

We provide a summary of mainstream GAN-faces generation methods used as the target in most GAN-faces Detection works. More detail on the various kinds of GANs can be found in the surveys of [47, 108].

The initial GAN model was proposed in 2014 and can only generate 32×32 faces. Since 2017, numerous GAN models (e.g., PG-GAN [54], BigGAN [8], StyleGAN [58], StyleGAN2 [59], etc.) have been developed to generate high-realistic faces that are hard to discern from human faces. For high-quality face image generation, these GANs can effectively encode rich semantic information in the intermediate features [7] and latent space [31, 49, 91]. Moreover, these GANs can generate fake face images with various attributes, including ages, expressions, backgrounds, and viewing angles. However, due to the lack of inference functions or encoders in GANs, such manipulations in latent space are only applicable to images generated from GANs, not to any given real images.
To address the above issue, GAN inversion methods can invert a given image back into its latent space of a pre-trained GAN model [108]. The GAN generator can then reconstruct the image accurately from the inverted code in approximation. This inversion method plays a key role in bridging real and fake face image domains. Therefore, it can significantly improve the quality of the generated face images and be applied widely in state-of-the-art GAN models including StyleGAN2 [59], StyleGAN3 [56], InterFaceGAN [91], and Image2StyleGAN++ [5].

2.3 GAN-face Detection Methods

We organize existing GAN-face detection literature into four categories in the sections. Although there exist similarities of various methods e.g. across categories, we organize them primarily by their motivations and key ideas. Table 2.1 summarizes mainstream GAN-face detection methods with the datasets used and performance comparison.

2.3.1 Deep Learning-based Methods

Deep learning-based GAN-face detection methods extract signal-level features to train Deep Neural Network (DNN) classifiers to distinguish fake faces from real ones in an end-to-end learning framework. The earliest work of [25] employed VGG-Net [93] for GAN-face detection. To train the network, real faces are collected from the CelebA face dataset [68], and fake faces are generated using DC-GANs [86] and PG-GAN [54], where the VGG-16 architecture is used with pre-train weights of VGG-Face [10]. [76] found that signals in the residual field can serve as effective features to distinguish real and GAN-faces. They first processed the input faces with high-pass filters, and the resulting residuals were fed into deep networks for GAN-face detection. [65] identified GAN-faces by analyzing the chrominance color components. They first extracted a feature set to capture color image statistics, then use the concatenated features to train a GAN-face classifier. Similarly, [16] found that both the luminance and chrominance cues are useful for improving GAN-face detection. More recently, [28] used a dual-channel CNN to reduce the impact of many widely-used image post-processing operations. The deep CNN of their network extracts features of the pre-processed images, and the shallow CNN extracts features from the high-frequency components of the original image.
2.3.1.1 GAN-face detection in real-world scenarios

However, the detection results from all these feature-based methods are not explainable, so it is unclear why the decision was given to any input face.

2.3.1.2 One-shot, incremental and advanced learning

A one-shot GAN-face detection method was studied recently in [71]. Scene understanding is applied to determine out-of-context objects that appeared in the GAN-faces to distinguish GAN-faces from the real ones. [73] applied incremental learning for GAN-faces image detection, where the key idea is to detect and classify new GAN-generated faces without decreasing the performance on existing ones.

2.3.1.3 Difficulty Analysis

More difficulty analysis and systemic evaluations using state-of-the-art DNNs for GAN-face detection are investigated in [34, 102, 103, 50]. For example, [103] find that the CNN-generated images share some common systematic flaws, resulting in them being surprisingly easy to spot for now. To investigate Are GAN-generated images easy to detect? [34] conducted the study to analyze the performance of the existing GAN-faces detection methods on different datasets and using different metrics. On the country, they concluded that we are still very far from having reliable tools for GAN image detection.

In summary, Deep Learning-based methods achieved impressive performance on GAN-face detection. However, it is difficult to explain or interpret the decision process of the learned model as a black box. Nonetheless, fake face detection in the real-world favors explainability, alongside from the overall accuracy. Particularly, people do care more for use cases such as “This picture looks like someone I know, and if the AI algorithm tells it is fake or real, then what is the reasoning and should I trust?”

2.3.2 Physical-based Methods

Physical-based methods identify GAN-faces by looking for artifacts or inconsistencies among the face and the physical world, such as the illumination and reflections in perspective.

The early work of [52] analyzed the internal camera parameters and light source direc-
tions from the perspective distortion of the specular highlights of the eyes to reveal traces of image tampering. Recently, [74] identified early versions of GAN-faces [54] based on an observation that the specular reflection in the eyes of GAN-faces is either missing or appearing as a simple white blob. However, such artifacts have been largely corrected in recent GAN-faces such as StyleGAN2.

The method of [42] looked for inconsistency between the two eyes to identify GAN-generated faces. Specifically, the corneal specular highlights of the eyes are detected and aligned for pixel-wise Intersection of Union (IoU) comparison.

The assumption is that real human eyes captured by a camera under a portrait setting should exhibit a strong resemblance between the corneal specular highlights between the two eyes. In contrast, this assumption is not true for GAN synthesized eyes, where inconsistencies include different numbers, different geometric shapes, or different relative locations of the specular highlights. However, this method operates on strong assumptions of the frontal portrait pose, far away lighting source(s), and the existence of the eye specular highlights. When these assumptions are violated, false positives may increase significantly.

In summary, the physical-based detection methods are more robust to adversarial attacks, and the predicted results afford intuitive interpretations to human users [42].

2.3.3 Physiological-based Methods

Physiologically-based methods investigate the semantic aspect of the human faces, including cues such as symmetry, iris color, pupil shapes, etc., where the identified artifacts are used for exposing GAN-faces.

Early works of [72, 114, 75] indicated that StyleGAN [54] generated faces contain obvious artifacts including asymmetric faces [112] and inconsistent iris colors [74]. [112] found that GAN can generate facial parts (e.g., eyes, nose, skin, mouth) with a great level of realistic details, yet there is no explicit constraint over the locations of these parts on the face. In other words, the facial parts of GAN-faces may not appear to be coherent or natural-looking, when compared to real faces.

They indicated that these abnormalities in the configuration of facial parts in GAN-faces could be revealed using the locations of the facial landmark points (e.g., tips of the
eyes, nose, and mouth), which can be effectively detected using automatic algorithms. The normalized locations of these facial landmarks can be used features to train a classifier to identify GAN-faces. However, GAN-face generation has also improved on the other hand. Face images generated by StyleGAN2 have improved greatly in quality and are free of obvious physiological artifacts [54, 58, 59]. And the synthesis process of GAN-faces is further optimized in StyleGAN3. It exhibits a more natural transformation hierarchy of different scales of features. They are fully equivariant to translation and rotation, which further improved the physiological consistency of the generated faces.

In summary, physiological-based method comes with stronger interpretability. However, like other forensic approaches, environmental constraints such as occlusion and visibility of the eye from the face image is still a major limitation. It is still an open question if the power of end-to-end learning is leveraged to improve model training.

2.3.4 Human Visual Performance

Although many automatic GAN-face detection algorithms have been developed, human visual performance in identifying and exposing GAN-faces has not been investigated sufficiently. Compared with other AI problems such as image recognition, GAN-face detection is a much more challenging problem for human eyes. Thus, it is important to study how well human eyes can identify GAN-faces and the related social impacts and ethical issues.

Standard metrics for evaluating the effectiveness of automatic algorithms in detecting GAN-faces include ROC analysis and Precision-Recall. While these metrics can be applied to study human perceptual performance, they are not directly suitable in reflecting the true deceptiveness of the highly realistic GAN-faces for the general public. Human visual performance is largely biased, and with weak but proper hints (such as looking for the correct physiological cues), human performance in identifying fake faces can boost greatly.

[63] conducted a study to measure the human ability to recognize fake faces. Their dataset consists of 150 real faces and 150 GAN faces. Real faces are selected from the Flickr-Faces-HQ (FFHQ) dataset, and GAN-Faces are generated from state-of-the-art GANs, including PG-GAN, StyleGAN, and StyleGAN2. The 630 participants sequentially completed 34 tasks to distinguish 30 faces each time. Those faces were randomly selected in equal portions from each category. Results showed that participants had lost the ability to judge
Table 2.1: Summary of GAN-face detection methods with the corresponding datasets, statistics and performance scores. The green rows highlight those where individual predicted results of the method are explainable to humans. Note that datasets used in the works are self-collected and can contain different subsets across papers. So the performance scores do not represent fair comparisons.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Category</th>
<th>Method</th>
<th>Real Face (#test)</th>
<th>GAN Face (#test)</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[25]</td>
<td>DL</td>
<td>CNN</td>
<td>CelebA (200)</td>
<td>PGGAN, DCGAN (200)</td>
<td>Acc: 0.80</td>
</tr>
<tr>
<td>[23]</td>
<td>DL</td>
<td>CNN</td>
<td>CelebA (1250)</td>
<td>PGGAN (1250)</td>
<td>Acc: 0.98</td>
</tr>
<tr>
<td>[76]</td>
<td>DL</td>
<td>CNN</td>
<td>CelebA-HQ (15K)</td>
<td>PGGAN (15K)</td>
<td>Acc: 0.99</td>
</tr>
<tr>
<td>[79]</td>
<td>DL</td>
<td>CNN</td>
<td>CelebA (500)</td>
<td>StarGAN (4498)</td>
<td>Acc: 0.99</td>
</tr>
<tr>
<td>[28]</td>
<td>DL</td>
<td>CNN</td>
<td>CelebA-HQ (7K)</td>
<td>PGGAN (7K)</td>
<td>Acc: 0.98</td>
</tr>
<tr>
<td>[73]</td>
<td>DL</td>
<td>Incremental Classifier</td>
<td>-</td>
<td>StarGAN (2.4K), etc.</td>
<td>Acc: 0.815~1</td>
</tr>
<tr>
<td>[71]</td>
<td>DL</td>
<td>Out of context object detection</td>
<td>-</td>
<td>StyleGAN (100)</td>
<td>Acc: 0.80</td>
</tr>
<tr>
<td>[102]</td>
<td>DL</td>
<td>DNN</td>
<td>FFHQ (1K)</td>
<td>StyleGAN2 (1K), etc.</td>
<td>Acc: 0.88~0.991</td>
</tr>
<tr>
<td>[65]</td>
<td>DL</td>
<td>Disparities in Color Components</td>
<td>CelebA-HQ, FFHQ (50K)</td>
<td>StyGAN, ProGAN (50K)</td>
<td>Acc: 0.997</td>
</tr>
<tr>
<td>[103]</td>
<td>DL</td>
<td>CNN</td>
<td>FFHQ (1K)</td>
<td>StyleGAN (1K)</td>
<td>Acc: 0.84</td>
</tr>
<tr>
<td>[46]</td>
<td>DL</td>
<td>ForensicTransfer</td>
<td>FFHQ (3K), etc.</td>
<td>StyleGAN (3K), ProGAN (3K), etc.</td>
<td>Acc: 0.01~1</td>
</tr>
<tr>
<td>[39]</td>
<td>DL</td>
<td>CNN</td>
<td>CelebA (164), CelebA-HQ (1.5K)</td>
<td>StarGAN (1476), ProGAN (3.7K)</td>
<td>Acc: 0.6768~0.849</td>
</tr>
<tr>
<td>[67]</td>
<td>DL</td>
<td>CNN</td>
<td>FFHQ (10K), CelebA-HQ (10K)</td>
<td>StyleGAN (10K), PGGAN (10K), etc.</td>
<td>Acc: 0.9854~0.991</td>
</tr>
<tr>
<td>[15]</td>
<td>DL</td>
<td>Xception</td>
<td>FFHQ (7K)</td>
<td>LGGF (14K)</td>
<td>Acc: 0.99</td>
</tr>
<tr>
<td>[16]</td>
<td>DL</td>
<td>Improved Xception</td>
<td>CelebA (202,60)</td>
<td>PGGAN (202,60)</td>
<td>Acc: 0.713~0.977</td>
</tr>
<tr>
<td>[34]</td>
<td>DL</td>
<td>CNN</td>
<td>RAISE (&lt;7.8K)</td>
<td>StyleGAN2 (3K), ProGAN (3K), etc.</td>
<td>Acc: 0.928~0.999</td>
</tr>
<tr>
<td>[17]</td>
<td>DL</td>
<td>CNN</td>
<td>FFHQ (20K)</td>
<td>StyleGAN2 (20K), etc.</td>
<td>Acc: 0.9895~1</td>
</tr>
<tr>
<td>[35]</td>
<td>DL</td>
<td>Residual Attention</td>
<td>FFHQ (748)</td>
<td>StyleGAN2 (750)</td>
<td>AUC: 1</td>
</tr>
<tr>
<td>[12]</td>
<td>Physic</td>
<td>Corneal specular highlight</td>
<td>FFHQ (500)</td>
<td>StyleGAN2 (500)</td>
<td>AUC: 0.94</td>
</tr>
<tr>
<td>[74]</td>
<td>Physiology</td>
<td>Eye color</td>
<td>CelebA (1K)</td>
<td>ProGAN (1K), Glow (1K)</td>
<td>AUC: 0.70~0.85</td>
</tr>
<tr>
<td>[30]</td>
<td>Physiology</td>
<td>Irregular pupil shape</td>
<td>FFHQ (1.6K)</td>
<td>StyleGAN2 (1.6K)</td>
<td>AUC: 0.91</td>
</tr>
<tr>
<td>[112]</td>
<td>Physiology</td>
<td>Landmark locations</td>
<td>CelebA (&gt;50K)</td>
<td>PGGAN (25K)</td>
<td>AUC: 0.9121~0.9413</td>
</tr>
<tr>
<td>[80]</td>
<td>Human</td>
<td>Visual</td>
<td>FFHQ (400)</td>
<td>StyleGAN2 (400)</td>
<td>Acc: 0.5~0.6</td>
</tr>
<tr>
<td>[63]</td>
<td>Human</td>
<td>Visual</td>
<td>FFHQ (150)</td>
<td>StyleGAN2, etc. (150)</td>
<td>Acc: 0.26~0.8</td>
</tr>
</tbody>
</table>
newer GAN-faces. Accuracy is not impacted when the test speeds up or the participants have seen similar synthetic faces produced by the generators before.

A more recent work [80] examined people’s ability to discriminate GAN-faces from real faces. Specifically, 400 StyleGAN2 faces and 400 real faces from the FFHQ dataset are selected with large diversity across the genders, ages, races, etc., and two sets of experiments are conducted. In the first set of experiments, 315 participants were shown a few examples of GAN-faces and real faces, and around 50% of accuracy is obtained. In the second set of experiments, 170 new participants were given a tutorial consisting of examples of specific artifacts in the GAN-faces. Participants were also given feedback afterward. However, it was found that such training and feedback only improve a little bit of average accuracy. Therefore, this work concluded that the StyleGAN2 faces are realistic enough to fool both naive and trained human observers. However, no information on what synthesis artifacts are provided for participant training in this study. We believe there is still space to improve human capability in discerning GAN-faces if sufficient hints are provided, including philosophical cues (e.g. pupil shapes [36]) and dataset statistics (e.g. GAN-faces are usually trained with FFHQ samples that are biased toward portrait faces and celebrity styles).

GAN models are under active development, so it is expected that the difficulty of discerning GAN-faces will continue to increase. It is important to find generic and consistent cues for human eyes to effectively distinguish GAN-faces. Typically, useful cues are generally universal for exposing other types of AI tampered faces, including morphed faces, swapped faces, painting faces. The discovery of such cues can also be leveraged for improving the GAN face synthesis algorithm to produce faces that are even harder to distinguish for human eyes.

2.4 Datasets and Performance Evaluation

With the rapid development of AI discriminative and generative models, many human facial datasets have been constructed. Among these datasets, real face images are mainly collected from the FFHQ dataset [58], CelebA [68], CelebA-HQ [55], RAISE [24] etc.. Synthesized face images are collected using state-of-the-art GAN models and LGGF [15]. Early GAN-faces datasets are mainly comprised of PGGAN, and recent datasets are typically based on StyleGAN2. NVIDIA has recently curated a StyleGAN3 generated set at https://github.com/NVlabs/stylegan3-detector that can be used to evaluate GAN-face
detection performance.

Table 2.1 list mainstream datasets for GAN-face detection. Note that datasets used for each work are self-collected and can contain different subsets across papers. This is due to that only specific subsets are relevant to individual methods. For example, in [36], only face images with visible eye pupils are used for training and evaluation.

As GAN-face detection is a binary classification problem, evaluation metrics are typically based on Accuracy, Precision-Recall, ROC analysis, and AUC. To the best of our knowledge, a sufficiently large-scale benchmark dataset for empirical evaluation of GAN-face detection is still lacking.
CHAPTER 3
Exposing GAN Faces Using Irregular Pupil Shape

Our fake face detection method is motivated by the observation that GAN-generated faces exhibit a common artifact that the pupils appear with irregular shapes or boundaries, other than a smooth circle or ellipse. This artifact is universal for all known GAN models (at least for now, e.g. PGGAN [54], Alias-Free GAN [56], and SofGAN [14]), as shown in Figure 3.1. This artifact occurs in both the synthesized human and animal eyes.

Our automatic fake face detection pipeline starts with a face detector to identify any face in the input image. We then extract the facial landmark points to localize the eyes and then perform pupil segmentation. The segmented pupil boundary curves are next analyzed to determine if the pupil shape is irregular. We perform parametric fitting of the pupil to an ellipse following the mean squared error (MSE) optimization. This provides a way to define a distance metric to quantify the irregularity for decision-making. The following sections describe each step of our pipeline in detail.

3.1 Iris and pupil segmentation

Iris and pupil segmentation is an important task in biometric identification that has been studied well. The IrisParseNet [97] provides complete iris segmentation solutions including iris mask and inner and outer iris boundaries extraction, which are jointly modeled in a unified multi-task neural network. Iris segmentation in non-cooperative environments is supported, while the iris pixel quality might be low due to the limited user cooperation (moving camera, poor illumination, or long-distance views). An end-to-end trainable lightweight stacked hourglass network is presented in [98] for iris segmentation from noisy images acquired by mobile devices. More recent methods can be found in the NIR Iris Challenge survey paper [99].
Figure 3.1: Examples of GAN-synthesized faces additional to StyleGAN and StyleGAN2. The images are from their original papers (a) PGGAN [54], (b,c) Alias-Free GAN (StyleGAN3) [56], (d) SofGAN [14]. Observe in the zoomed-in view that the pupils appear in irregular, inconsistent shapes, which tell them apart from real faces.

3.2 Pupil Segmentation and Boundary Detection

We adopt the Dlib [61] face detection to locate the face and extract the 68 facial landmark points provided in Dlib, as shown in Figure 5.3. We focus on the eye regions to perform pupil segmentation. We use EyeCool [99] to extract the pupil segmentation masks with corresponding boundary contours. EyeCool provides an improved U-Net-based model with EfficientNet-B5 [95] as the encoder. A boundary attention block is added in the decoder to improve the ability of the model to focus on the object boundaries. Specifically, considering the subpixel accuracy, we focus on the outer boundary of the pupil for the irregularity analysis.

3.3 Ellipse Fitting to the Pupil Boundary

We fit an ellipse to the pupil mask via least-square fitting. As there might be multiple components in the predicted masks, we keep the largest component for ellipse fitting. Specifically, the method of [27] is used to fit an ellipse to the outer boundary of the extracted pupil mask. Figure 5.3(d) shows an example. Denotes \( u \) as the coordinates of the outer boundary points from the pupil mask. The least-square fitting determines the ellipse parameters \( \theta \) minimizing the distance between the pupil boundary points and a parametric

\[ \text{Code at } https://github.com/neu-eyecool/NIR-ISL2021. \]
ellipse represented by:

\[ F(\mathbf{u}; \theta) = \theta \cdot \mathbf{u} = ax^2 + bxy + cy^2 + dx + ey + f = 0, \]

where \( \theta = [a, b, c, d, e, f]^T \) and \( \mathbf{u} = [x^2, xy, y^2, x, y, 1]^T \); \( T \) denotes transpose. \( F(\mathbf{u}; \theta) \) represents the algebraic distance of a 2D point \((x, y)\) to the ellipse, and a perfect fit is indicated by \( F(\mathbf{u}; \theta) = 0 \).

The fitting solution is obtained by minimizing the sum of squared distances (SSD) over the \( N \) data points from the pupil boundary:

\[
\min_{\theta} D(\theta) := \sum_{i=1}^{N} F(u_i; \theta_i)^2 \\
\text{s.t. } ||\theta||^2 = 1, \quad b^2 \geq ac,
\]

where the constraints are imposed to avoid the trivial solution of \( \theta = 0 \) and ensure the positive definiteness of the quadratic form. The solution is calculated using the gradient-based optimization described in [27].

### 3.4 Estimating the Pupil Shape Irregularity

To accurately estimate the irregularity of the segmented pupil boundary and the fitted ellipse, we adopt the Boundary IoU (BIOU) [20] as a distance metric. BIOU is widely used in image segmentation where the sensitivity of the object boundary is important. Instead
Figure 3.3: A schematic example explaining the Boundary IoU. **Left:** The predicted pupil mask $P$ and the ellipse fitted pupil mask $F$. **Middle:** $P_d$ and $F_d$ are the mask pixels within distance $d$ from the boundaries (blue and yellow). **Right:** Boundary IoU calculation between predicted pupil mask and the ellipse fitted pupil mask with distance parameter $d$.

of considering all pixels, BIoU calculates the IoU for mask pixels within a certain distance from the boundary contours between the predicted mask and the corresponding ground truth mask. Thus, BIoU can better focus on the matching of the boundaries of the two shapes. We use BIoU to evaluate the pupil mask pixels that are within a distance of $d$ pixels from the pupil boundary. For each extracted pupil mask, we use $P$ to indicate the predicted pupil mask and $F$ for the fitted ellipse mask. Denote $P_d$ and $F_d$ the mask pixels within distance $d$ from the predicted and fitted boundaries, respectively. BIoU is calculated as: (See Figure. 3.3),

$$\text{BIoU} (F, P) = \frac{|(F_d \cap F) \cap (P_d \cap P)|}{|(F_d \cap F) \cup (P_d \cap P)|}$$  \hspace{1cm} (3.1)

The distance parameter $d$ controls the sensitivity of the BIoU measure to the object boundary. Reducing the value of $d$ causes the fitting to be more sensitive to the boundary pixels while ignoring the interior pixels of the pupil mask. We set $d = 4$ for the BIoU calculation, which leads to the best empirical segmentation performance in our experiments.

Given the predicted pupil mask and the ellipse fitted pupil mask, the BIoU score takes range in $[0, 1]$. A larger BIoU value suggests the pupil boundary better fits the parametrized
ellipse. In our case, higher BIoU values suggest more regular pupil shapes, and thus the face is more likely real. In comparison, GAN-generated faces should produce lower pupil BIoU scores (see Figure. 3.4).

3.5 Experiments

3.5.1 Datasets

We use the real human faces from the Flickr-Faces-HQ (FFHQ) dataset [58]. Since StyleGAN2 [59] is currently the state-of-the-art GAN face generation model with the best synthesis quality, we collect GAN-generated faces from it. We only use images where the eyes and pupils can be successfully extracted. In total, we collected 1,600 images for each class (of real vs. fake faces) with a resolution of 1,024 × 1,024.

3.5.2 Results

Figure 3.6 shows examples of the segmented pupils for both the real and GAN-generated faces. These results clearly show that pupils in the real faces are in strongly regular, elliptical shapes. Such high pupil shape regularity is also reflected in the high BIoU scores computed for the pupil mask and the fitted ellipse. On the other hand, irregular pupil shapes lead to significantly lower BIoU scores, which represents the artifacts of GAN-generated faces.

<http://thispersondoesnotexist.com>
Figure 3.5: **Top:** Distributions of the BIoU scores (of the averages of both pupils) for the real and GAN-generated faces. **Bottom Left:** The ROC curve based on the BIoU with $d = 4$. **Bottom Right:** BIoU hyper-parameter analysis, where the $x$-axis indicates distance parameter $d$, and $y$-axis indicates the AUC score.

Figure 3.5(top) shows the distributions of the BIoU scores of pupils from the real faces and GAN-generated ones. Observe that there is a clear separation between the two classes of distributions, indicating that the proposed *pupil shape regularity* can indeed serve as an effective feature to distinguish the GAN-generated faces from the real ones.

Figure 3.5(bottom left) shows the **receiver operating characteristic** (ROC) curve of our GAN-generated face detection evaluation. The Area under the ROC curve (AUC) score is 0.91, which indicates the effectiveness of the proposed method.
3.5.2.1 Sensitivity Analysis of $d$

The BIoU boundary distance $d$ is an essential parameter that controls the matching sensitivity of the pixels near shape boundary. Figure 3.5(bottom right) shows how the fake face detection ROC varies w.r.t. parameter $d$. As the value of $d$ grows too large, sensitivity telling the differences of pupil boundary decreases, which also reduces fake face detection performance.
3.5.2.2 Limitations

The proposed method still contains several limitations. Since our method is based on the simple assumption of pupil shape regularity, false positives may occur when the pupil shapes are non-elliptical in the real faces. This may happen for infected eyes of certain diseases as shown in Figure 3.7(top). Also imperfect imaging conditions including lighting variations, largely skew views, and occlusions can also cause errors in pupil segmentation or thresholding errors, as shown in Figure 3.7(bottom).

3.6 Summary

In summary, we show that GAN-generated faces can be identified by exploiting the regularity of the pupil shapes. In general, this artifacts are different to fix, because the GANs optimize the whole face image, and the GANs models can not know where the pupil part is. We propose an automatic method for pupil localization and segmentation, and perform ellipse fitting to the segmented pupils to estimate a Boundary IoU score for forensic classification. The proposed approach is simple yet effective. The detection results are

Figure 3.7: Top: Examples of diseased and infection eyes from [85]. These pupils from images of real faces contain abnormal non-elliptical pupil shapes, which only occurs rarely in real life. Bottom: Occlusions and environmental variations can cause pupil segmentation failure.
interpretable based on the BIoU score.

We will investigate other types of inconsistencies between two pupils of the GAN-generated face, such as the different geometric shapes and relative locations of pupils in the two eyes. These cues in combination may further improve forensic detection effectiveness. Future work also includes the deployment to an online platform that can further expand the impact in addressing issues in social media forensics.
Figure 3.8: More pair of pupil examples from real (left) human faces and GAN-generated (right) faces. As we mentioned before, the irregular pupil shape is a good sign for the human to expose the GAN-generated face visually, even there are no boundary labels around the pupils, we can easily see that the shapes of GAN-generated pupils are very irregular, and the shapes of both pupils are very different in the same GAN-generated face image. In practice, people can zoom a face image large enough and then check the pupil shapes to find whether the face is real or not easily.
CHAPTER 4
Robust Attentive Deep Neural Network for Detecting
GAN-Generated Faces

In this chapter, we describe the proposed GAN-generated faces detection framework.

Given an input face image, facial landmarks are first localized using DLib [60], and Mask R-CNN [38] is used to segment out the left and right iris regions of the eyes (S 4.2). We adopt a residual attention-based network [101] to perform binary classification on the iris regions of interest to determine if the input image is real or fake (S 4.3).

The training of our network aims to maximize the classification performance reflected by the standard Area-Under-Curve of the ROC plot (S 4.4), which is general and can effectively address the data imbalance problem. However, due to the discrete nature of the ROC-AUC values, a naive gradient-based implementation does not work for end-to-end learning. In (S 4.5), we present a detailed solution in our proposed approach by relaxing the AUC maximization and approximating the goal using the WMW statistics. Figure 4.2 overviews the training pipeline of the proposed method that can effectively learn from imbalanced data.

4.1 Background

This section discusses the backgrounds of the proposed method, including attention mechanisms and imbalanced data learning.

4.1.1 Attention Mechanism

Since the seminal work of [6] in machine translation, the attention mechanism is widely used in many applications on improving the performance of deep learning models by focusing on the most relevant part of the features in a flexible manner. The Class Activation Mapping (CAM) [116] and Grad-CAM [90] are widely used in many computer vision tasks [13]. However, in these works, attentions are only used to visualize model prediction in showing significant portions of the images. On the other hand, integrating the attention mechanism
into the network design is shown to be effective in boosting performance, as the network can be guided by the attention to focus on relevant regions during training [53].

The channel attention [106] can automatically learn to focus on important channels by analyzing the relationship between channels. SENet [40] embeds the channel attention mechanism into residual blocks, and effectiveness is shown on large-scale image classification. The attention mechanism is also used in [115] to distinguish important channels in the network to improve the representation capability. The idea of channel attention and spatial attention are combined jointly in [19, 107] to improve network performance significantly. The Residual Attention Network in [101] combines the residual unit [39] with the attention mechanism by stacking residual attention blocks to improve performance and reduce model complexity.

### 4.1.2 Imbalanced Data Learning

Learning from imbalanced data has been widely studied in machine learning [113, 9, 43, 44] and computer vision [105, 45]. Earlier solutions for imbalanced data learning are mainly based on the sampling design, e.g., oversampling for minor classes, undersampling for major classes, and weighed sampling [37], etc. These sampling-based methods come with their own drawbacks. For example, undersampling may ignore important samples, and oversampling may lead to overfitting.

Data augmentation provides an alternative solution to alleviate data imbalance issues. For image recognition, image mirroring, rotation, color adjustments, etc. are simple methods to augment data samples [92]. However, data augmentation methods can only address the data imbalance problems partly, as the size of the original dataset must be diverse enough, such that a sufficient amount of representative samples can be produced from augmentation.

### 4.2 Facial Landmark Localization and Iris Segmentation

Given a face image, the first step of our method to determine if it is real or GAN-generated is to detect and localize the face using the facial landmark extractor provided in DLib [60]. The localized regions containing the eyes are cropped out for consistency checking. Mask R-CNN [38], the state-of-the-art detection and segmentation network, is employed to
Table 4.1: Details of the proposed **Residual Attention Network (RAN)** in the right of Figure 4.3.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Output Size</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>96×96×16</td>
<td>3 × 3, stride 1</td>
</tr>
<tr>
<td>Max pooling</td>
<td>48×48×16</td>
<td>3 × 3, stride 2</td>
</tr>
<tr>
<td>Residual block</td>
<td>48×48×64</td>
<td>1 × 1, 16, 3 × 3, 16, 1 × 1, 64</td>
</tr>
<tr>
<td>Attention Module</td>
<td>48×48×64</td>
<td>2× attention</td>
</tr>
<tr>
<td>Residual block</td>
<td>24×24×128</td>
<td>1 × 1, 1, 32, 3 × 3, 32, 1 × 1, 128, 1 × 1, 128</td>
</tr>
<tr>
<td>Attention Module</td>
<td>24×24×128</td>
<td>2× attention</td>
</tr>
<tr>
<td>Residual block</td>
<td>12×12×256</td>
<td>1 × 1, 1, 64, 3 × 3, 64, 1 × 1, 256, 1 × 1, 256</td>
</tr>
<tr>
<td>Attention Module</td>
<td>12×12×256</td>
<td>2× attention</td>
</tr>
<tr>
<td>Residual block</td>
<td>6×6×512</td>
<td>3 × 1, 1, 128, 3 × 3, 128, 1 × 1, 512, 1 × 1, 512</td>
</tr>
<tr>
<td>Average pooling</td>
<td>1×1×512</td>
<td>6 × 6, stride 1</td>
</tr>
<tr>
<td>FC, Sigmoid</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Furthermore detect and localize the iris regions. Mask R-CNN is a two-stage network based on Faster R-CNN [87] as shown in Figure 4.3 (middle). The first stage of Mask R-CNN is a Region Proposal Network (RPN) that generates candidate object bounding boxes for all the object categories. In the second stage, the R-CNN extracts features using the Region of Interest Align (RoIAlign) layer for each proposal. In the last stage, label classification and bounding box regression are performed for each proposal, and mask prediction is performed in a parallel branch. We train the Mask R-CNN model using the eye region images from the datasets in [97, 100], where more details will be provided in §4.8. Figure 4.1 shows examples of the extracted pair of iris regions from the cases of GAN-generated (left) and real (right) faces.
4.3 Residual Attention Network

We adopt the attention mechanism [18, 48, 51, 106, 115] to improve the spotting of inconsistent corneal specular highlights between the eyes so as to improve GAN-generated face detection. Incorporating attention to a detection/segmentation network is commonly accomplished by having a separate branch that calculates the attention maps and later is incorporated back to the main branch with weights.

Inspired from [101], each Attention Module in our attention network consists of a trunk branch and a soft mask branch. The trunk branch contains several residual blocks [39] and acts as a shortcut for data flow. The soft mask branch uses a U-net structure [88] to weight output features. Specifically, given the input feature map \( f \), denote the output of the trunk branch \( T \) as \( T(f) \), and the output of the soft mask branch \( M \) as \( M(f) \), respectively. As illustrated in Figure 4.2, the final attended feature map \( f' \) is obtained via element-wise matrix product via

\[
f' = (1 + M(f)) \odot T(f),
\]

where the symbol \( \odot \) denotes the Hadamard product.

The attention module can be configured to focus learning on channel attention, spatial attention, or mixed attention. As suggested in [101], the mixed attention yields the best performance. Thus, we use the Sigmoid function \( \frac{1}{1 + \exp(-f_{s,c})} \) to learn the mixed attention for each channel and each spatial location, where \( s \) ranges over all spatial positions and \( c \) ranges over all channels of \( f \). The proposed Residual Attention Network (RAN) is constructed by stacking multiple Attention Modules, as shown on the right side of Figure 4.3. Table 4.1 provides details of the architectures. Although the attention module plays an important role in classification, a simple stacking of attention modules may reduce performance. To this end, we adopt a simple solution by adding the attention map onto the original feature map. This combination allows attention modules stacked like a ResNet [39] and improves the performance [101]. Given an input image, the RAN outputs a prediction score from the last Sigmoid layer, as an indication of the likelihood of the input image being a GAN-generated image.
Table 4.2: The proposed pipeline for training the Residual Attention Network (RAN) on possibly imbalanced data for GAN-generated face classification. The extracted iris pairs are passed as input to the RAN. A robust loss function derived from maximizing the AUC of ROC is optimized in the training of the RAN. See details in §4.5.
4.4 AUC of ROC for Classification Evaluation

Most classification loss measures including the popular cross-entropy loss are ineffective in addressing the issue of data imbalance. The resulting models can produce accurate but rather biased predictions that do not work well in practice. It is desirable to address data imbalance directly by specifically designing a suitable loss function.

Since the area under the curve (AUC) of a receiver operation curve (ROC) \cite{22, 70} is a robust evaluation metric for both balanced and imbalanced data, we would like to directly maximize the AUC to handle imbalanced situations. The AUC is widely used in the binary classification problems, it is a ranking loss letting all the positive examples ranked before negative examples. We next briefly review the definition of AUC, and then motivate how we incorporate a loss term that directly maximize the AUC performance. Given a labeled dataset \( \{(x_i, y_i)\}_{i=1}^M \), where each data sample \( x_i \in \mathbb{R}^d \) and each corresponding label \( y_i \in \{-1, +1\} \). We define a set of indices of positive instances as \( P = \{i \mid y_i = +1\} \). Similarly, the set of indices of negative instances is \( N = \{i \mid y_i = -1\} \). Let \( g_w : \mathbb{R}^d \to \mathbb{R} \) be a parametric prediction function with parameter \( w \in \mathbb{R}^m \). \( g_w(x_i) \) represents the prediction score of the \( i \)-th sample, where \( i \in \{1, \cdots, M\} \). For simplicity, we assume \( g_w(x_i) \neq g_w(x_j) \) for \( i \neq j \) (ties can be broken in any consistent way).

Given a threshold \( \lambda \), the number of negative examples with prediction scores larger than \( \lambda \) is false positive (FP), and the number of positive examples with prediction scores greater or equal to \( \lambda \) is true positive (TP). According to the FP and TP, we can calculate the false positive rate (FPR) and the true positive rate (TPR) as follows,

\[
FPR = \frac{\sum_{i \in N} I[g_w(x_i) > \lambda]}{|N|}, \quad TPR = \frac{\sum_{i \in P} I[g_w(x_i) \geq \lambda]}{|P|},
\]

where \( I[a] \) is an indicator function with \( I[a] = 1 \) if \( a \) is true and 0 otherwise. The receiver operation curve (ROC) is a plot of FPR versus TPR with setting different decision thresholds \( \lambda \in (-\infty, \infty) \). Based on this definition, ROC is a curve confined to \( [0, 1] \times [0, 1] \) and connecting the point (0,0) to (1,1). The value of AUC corresponds to the area enclosed by the ROC curve.
Table 4.3: The proposed architecture for GAN-generated face detection. We first use DLib [60] to detect faces and localize eyes, and use Mask R-CNN [38] to segment out the iris regions. A Residual Attention Network (RAN) then performs binary classification on the extracted iris pair to determine if the face is real or fake. The training is carried out using a joint loss combining the Binary Cross-Entropy (BCE) loss and the ROC-AUC loss with WMW relaxation to better handle the learning from imbalanced data (see text).
4.5 WMW AUC Relaxation for Loss Design

The computation of an AUC score based on the area under a ROC curve cannot be directly used in a loss function due to its discrete nature. According to the Wilcoxon-Mann-Whitney (WMW) statistic [110], we can relax the AUC as follows,

\[ AUC = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} I[g_w(x_i) > g_w(x_j)]. \]

Therefore, the corresponding AUC loss (risk) can be defined as:

\[ L_{AUC} = 1 - AUC = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} I[g_w(x_i) < g_w(x_j)]. \] (4.2)

Obviously, \( L_{AUC} \) takes value in \([0, 1]\). It is a fraction of pairs of prediction scores from the positive sample and negative sample that are ranked incorrectly, i.e., the prediction score from a negative sample is larger than the prediction score from a positive sample. If all prediction scores from the positive samples are larger than any prediction score from the negative samples, then \( L_{AUC} = 0 \). This indicates we obtain a perfect classifier. Furthermore, \( L_{AUC} \) is independent of the threshold \( \lambda \). \( L_{AUC} \) only depends on the prediction scores \( g_w(x) \).

In other words, the predictor \( g_w \) affects the value of \( L_{AUC} \). Therefore, we aim to learn a classifier \( g_w \) that minimizes Eq.(4.2).

Although we can calculate \( L_{AUC} \) by comparing prediction score from the positive sample and prediction score from the negative sample in each pair, the \( L_{AUC} \) formulation is non-differentiable due to the discrete computation. It is therefore desirable to find a differentiable approximation for \( L_{AUC} \). Inspired by the work in [110], we find an approximation to \( L_{AUC} \) that can be directly applied to our objective function to minimize the AUC loss along with our imbalanced training procedure. Specifically, a differentiable approximation of \( L_{AUC} \) can be reformulated as:

\[ L_{AUC} = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} R(g_w(x_i), g_w(x_j)) , \] (4.3)
and \( R(g_w(x_i), g_w(x_j)) = \)
\[
\begin{cases} 
   ((-g_w(x_i) - g_w(x_j) - \gamma))^p, & g_w(x_i) - g_w(x_j) < \gamma, \\
   0, & \text{otherwise},
\end{cases}
\]
\[(4.4)\]

where \( \gamma \in (0, 1] \) and \( p > 1 \) are two hyperparameters.

**Loss for the proposed Residual Attention Network.** We use a joint loss function comprising the conventional binary cross-entropy (BCE) loss function \( L_{BCE} \) and the AUC loss function \( L_{AUC} \) in weighted sum:
\[
L = \alpha L_{BCE} + (1 - \alpha) L_{AUC},
\]
\[(4.5)\]

where \( \alpha \in [0, 1] \) is a scaling factor that is designed for balancing the weights of the BCE loss and the AUC loss. Note that the loss function is generic for all the data imbalance problems.

### 4.6 Experiment

For experimental evaluation of the proposed method and comparison against the state-of-the-art methods, we first introduce the newly constructed FFHQ-GAN datasets \( S \). Implementation details of the proposed method are provided in \( S \). Performance evaluation on the FFHQ-GAN balanced and imbalanced subsets is in \( S \). Ablation studies are provided in \( S \). Finally, qualitative results are shown in \( S \).

#### 4.7 The New FFHQ-GAN Dataset

We collect real human face images from the Flickr-Faces-HQ (FFHQ) dataset \([58]\). GAN-generated face images are created using StyleGAN2 \([58]\) \(^1\), where the image resolution is \( 1024 \times 1024 \) pixels. We randomly select 5,000 real face images from FFHQ and 5,000 GAN-generated face images. After iris detection, we discard those images with the iris of any eye not detected. This ends up with 3,739 real faces (with iris pairs) and 3,748 fake faces (with iris pairs), which constitutes our new **FFHQ-GAN dataset**. The split ratio of training and testing is 8:2.

\(^1\)http://thispersondoesnotexist.com
Table 4.4: Details of the FFHQ-GAN dataset regarding its balanced (-b) and imbalanced (-imb) subsets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Training Set</th>
<th>Testing Set</th>
<th>Ratio (\approx)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GAN-face</td>
<td>Real</td>
<td>GAN-face</td>
</tr>
<tr>
<td>FFHQ-GAN-b</td>
<td>2998</td>
<td>2991</td>
<td>750</td>
</tr>
<tr>
<td>FFHQ-GAN-imb</td>
<td>400</td>
<td>2000</td>
<td>100</td>
</tr>
</tbody>
</table>

To enable a thorough evaluation of the model in both balanced and imbalanced data scenarios, we sampled the FFHQ-GAN dataset to form an imbalanced subset, where the statistics of the subsets are provided in Table 4.4.

### 4.8 Implementation Details

We implemented our method in PyTorch [82]. Experiments are conducted on a workstation with two NVIDIA GeForce 1080Ti GPUs.

For iris detection, Mask R-CNN is trained using the datasets from [97, 100]. For each training eye image, the outer boundary mask of each iris is obtained using the method of [97] with default hyper-parameter settings. These masks are used to generate the iris bounding boxes and the corresponding masks for training, using the default settings in [38]. In the test stage, given an input face image, we first use the face detector and landmark extractor of DLib [60] to crop out the eye regions. Each cropped eye region is fed to Mask R-CNN for localizing the iris bounding box and segmentation mask. This process is repeated for both the left and right eyes to obtain the iris pairs as the input for our Residual Attention Network. We resize all iris pairs to a fixed size 96 × 96 for training and testing to ensure that the whole pipeline works well.

Table 4.1 describes the details of our Residual Attention Network (RAN), where the Attention Module (AM) detailed in Figure 4.2 is repeatedly stacked three times. The network is trained using Adam optimizer [62] with the learning rate of 0.001 and batch size 128. Training is terminated at 100 epochs for balanced data and 2,000 for imbalanced data.
Table 4.5: Results on the FFHQ-GAN dataset regarding its balanced (-b) and imbalanced (-imb) subsets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>methods</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ACC</td>
</tr>
<tr>
<td>FFHQ-GAN-b</td>
<td>ResNet</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Xception</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>RAN with BCE loss</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>RAN with BCE+AUC loss</td>
<td>0.97</td>
</tr>
<tr>
<td>FFHQ-GAN-imb</td>
<td>ResNet</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Xception</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>RAN with BCE loss</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>RAN with BCE+AUC loss</td>
<td>0.96</td>
</tr>
</tbody>
</table>

4.8.0.1 Hyper-parameters

We set $p = 2$ in Eq. (4.4) and $\gamma = 0.4$ for balanced dataset, and $\gamma = 0.6$ for imbalanced data. For the experiments on the balanced dataset, $\alpha$ in Eq. (4.5) is set to 0.2. For the experiments on the imbalanced dataset, $\alpha$ is set to 0.4. These hyperparameters yields the best performance.

4.9 Evaluation on the FFHQ-GAN dataset

We report evaluation of GAN-generated face detection on the FFHQ-GAN dataset in terms of Accuracy (ACC), Precision (P), Recall (R), F1 score (F1), the area under the curve (AUC) of the ROC, and Precision-Recall (PR) curves. Accuracy is calculated as Equation (4.6), where $FN$ and $TN$ indicate false negatives and true negatives, respectively. Precision-Recall is calculated as Equation (4.7) and (4.8). F1 score is the harmonic average value of P and R, as (4.9).

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$ (4.6)

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$ (4.7)

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$ (4.8)
\[ F1 = \frac{2PR}{P + R} \]  \hspace{1cm} (4.9)

4.9.0.1 Results on the balanced and imbalanced set

To evaluate the effectiveness of the proposed method, we evaluate RAN trained with BCE + AUC loss against the ResNet-50 [39] and Xception [21], two of the widely-used DNN classification models trained with the BCE loss. Table 4.5 presents the classification results of experiments training/test on the balanced and imbalanced FFHQ-GAN datasets. The corresponding ROC and Precision-Recall curves are shown in Figure 4.3 and 4.4, and the Confusion Matrices are shown in Figure 4.5. Results show that both the ResNet-50 and Xception obtain low Recall scores due to the imbalanced data distribution in Table 4.5. In comparison, our method achieves the highest performance in all metrics. These results indicate that our method can effectively improve performance on both balanced and imbalanced data training. We have also performed experiments to train our model, ResNet-50, and Xception on a balanced dataset and test on an imbalanced dataset. The obtained performance difference is similar to that of training/test on the balanced dataset. This result suggests the importance of the training model on an imbalanced dataset if the model is expected to deal with detection on imbalanced data.

4.10 Ablation Studies

4.10.1 Effect of the AUC loss

We compare the proposed RAN model trained with the ideal case with combined AUC and BCE loss in Eq. (4.5) against the same model trained only with BCE loss. Results are shown in Table 4.5 and Figures 4.3 and 4.4. Observe that the proposed joint BCE+AUC loss outperforms the same model trained only with BCE loss alone in all evaluation metrics. In other words, the incorporation of the AUC loss improves the classification performance substantially and consistently.
4.10.2 Hyper-parameter analysis

We also study the impact of hyper-parameter $\alpha$ in our loss function in Eq. (4.5) regarding detection performance of imbalanced data. Figure 4.6 shows the experimental results of the obtained AUC score versus $\alpha$ ranging from 0 to 1, and $\alpha = 0.4$ yields the best detection performance.

4.11 Qualitative Results

Figure 4.7 provides visualization of the attention maps of the real and GAN-generated iris examples. Observe that there is an obvious difference between the corresponding attention maps of the GAN-generated irises and the real ones. Concretely, the network attends on the whole iris part for the real images and attends to the highlight parts for the fake images. Figure 4.8 shows additional examples of the GAN-generated face with the extracted iris pairs and corresponding attention maps. The visualization also provides an intuitive approach for human beings to identify GAN-generated faces by comparing their iris regions.

4.12 Summary

In summary, we investigate building a robust end-to-end deep learning framework for detecting GAN-generated faces. We show that GAN-generated faces can be distinguished from real faces by examining the consistency between the two iris regions. In particular, artifacts such as the corneal specular highlight inconsistencies can be robustly identified through end-to-end learning via the proposed Residual Attention Network. Our design of a joint loss combining the AUC loss with the cross-entropy loss can effectively deal with the learning from imbalanced data. We also showed that a direct optimization of the ROC-AUC loss is computationally not feasible, however relaxing the ROC AUC via the Wilcoxon-Mann-Whitney (WMW) statistics can provide a good approximation. Our GAN-face detection result is explainable, and the approach of spotting iris inconsistency can also serve as a useful cue for human users. Experimental results show that our model achieves superior performance on both balanced and imbalanced datasets for GAN-generated faces detection.
Figure 4.1: The extracted iris pairs of our method for the (a) GAN-generated and (b) real faces. Artifacts of inconsistent corneal specular highlights are obvious in GAN-generated iris pairs.
Figure 4.2: Details of our Attention Module from the our RAN in Figure 4.3. Our design is inspired from the residual attention network of [101].

Figure 4.3: ROC of the proposed method, ResNet with BCE loss, Xception with BCE loss, and RAN with BCE loss.
Figure 4.4: PR curve of the proposed method, ResNet with BCE loss, Xception with BCE loss, and RAN with BCE loss.

Figure 4.5: Confusion Matrix on the FFHQ-GAN (Top) balanced and (Bottom) imbalanced datasets.
Figure 4.6: **Impact of hyperparameter** $\alpha$ of the AUC loss in Eq. (4.4) for the GAN-generated face detection on the imbalanced dataset.
Figure 4.7: Visualization of the extracted iris pairs and the corresponding attention maps obtained from our Residual Attention Network (RAN). Observe that the attention maps for GAN-generated faces better focus on the artifacts such as the corneal specular highlights, while the attention maps for real faces are widely distributed. This shows the effective learning of RAN for identifying GAN-generated faces.
Figure 4.8: Examples of **detected GAN-generated faces** and their corresponding iris regions and the attention maps produced from our method. These examples show that our method can detect a wide range of face images, including those with tilted or side views where both irises are visible.
CHAPTER 5

An Open Platform to Study Human Performance on Identifying AI-Synthesized Faces

The rapid development of Artificial intelligence (AI) technologies, called Deep Fakes, has made it possible to synthesize highly realistic images, audio, and video that are difficult to discern from real ones. In particular, AI-synthesized faces have been misused for malicious purposes. Recent years have seen an increasing number of reports that AI-synthesized faces were used as profile images on fake social media accounts, which generates negative social impacts [3, 1, 2, 4].

Generative adversarial networks (GANs) are the most popular technologies for synthesizing content and especially the human face images [54, 58, 59, 80]. The pernicious impact of these synthesized faces has led to the development of methods aiming to distinguish GAN-generated images from real ones. Many of those methods are based on deep neural network (DNN) models due to their high detection accuracy [72, 103, 35]. Although they have achieved high accuracy and the models are end-to-end, the DNN-based methods suffer from several limitations. The models lack interpretability to the detection results, and they usually have a low capability to generalize across different synthesis methods [43, 44].

As the existing methods are either less efficient or not accurate enough to handle the torrent of daily uploads of the public content [104], the users must have the ability to recognize the fake faces from the real ones. Recently, the studies investigating the human performance of AI-synthesized faces detection have been conducted [63, 80]. For example, [80] examined people’s ability to discriminate GAN faces from real faces. Specifically, 400 StyleGAN2 faces and 400 real faces from the FFHQ dataset are selected with large diversity across the genders, ages, races, etc., and two sets of experiments are conducted. In the first set of experiments, 315 participants were shown a few examples of GAN faces and real faces, and around 50% of accuracy was obtained. In the second set of experiments, 170 new participants were given a tutorial consisting of examples of specific artifacts in the GAN faces. Participants were also given feedback afterward. However, it was found that such training
Recent generative adversarial networks (GANs), have made it possible to synthesize highly realistic faces, which have been used in the creation of fraudulent social media accounts. In the ongoing fight against GAN synthesized highly photo-realistic faces, we examine people's ability to discriminate between synthesized and real faces.

**Testing Process**

Figure 5.1: The workflow of Open-eye. The participants test 20 face images, including the same amount of real and fake. A quick tutorial is demonstrated to participants to learn how to recognize the specific artifacts in the eye. The same test will be used again to show the human performance of the AI-synthesized faces detection.

and feedback only improve a little bit of average accuracy. Therefore, this work concluded that the StyleGAN2 faces are realistic enough to fool naive and trained human observers. However, this study provides no information on what synthesis artifacts are provided for participant training. There is no open platform for public use to test.

To overcome the above limitations, we propose an open platform, Open-eye. The Open-eye consists of several steps (See Figure 5.1). In Stage 1, the participants are tested with real and AI-synthesized faces. In Stage 2, the participants are trained with the artifacts to identify AI-synthesized faces reliably. Because we believe there is still space to improve human capability in discerning AI-synthesized faces if sufficient hints are provided, such as philosophical cues (e.g. pupil shapes [36]). In Stage 3, the participants will do the test again using the same data in Step 1. Finally, the testing performance of both stages will show for comparison. Our Open-eye platform is available at [http://zinc.cse.buffalo.edu/ubmdfl/human-performance-on-gan-face/#/](http://zinc.cse.buffalo.edu/ubmdfl/human-performance-on-gan-face/#/).

### 5.1 Contribution

We are the first to propose an open platform to study whether human participants can distinguish state-of-the-art AI-synthesized faces from real faces visually.

The proposed platform is flexible to incorporate any AI-synthesized faces and provide quick training to the participants to recognize the fake faces. An example shows that the
platform is simple, effective, and efficient for participants.

5.2 Background

5.2.0.1 AI-synthesized faces

One of the most popular AI-synthesized faces techniques is based on GAN models. A GAN model includes two neural networks (generator and discriminator) trained in tandem. The generator takes random noises as input and can effectively encode rich semantic information in the intermediate features and latent space for high-quality face image generation. The discriminator aims to distinguish synthesized images from the real ones. Generator and discriminator compete with each other during the training. A series of GAN models (e.g., PGGAN [54], BigGAN [8], StyleGAN [58], StyleGAN2 [59], StyleGAN3 [56]) have been developed and demonstrated superior capacity in generating or synthesizing realistic human faces. In some early works such as [112], they find that faces generated by the early StyleGAN model [58] have considerable artifacts such as fingerprints [72, 114], inconsistent iris colors [64, 75], etc. However, just one year later, StyleGAN2 is proposed by Karras et al. in [59] and it has greatly improved the visual quality and pixel resolution, with largely-reduced or undetectable artifacts in the generated faces.

5.2.0.2 AI-synthesized faces detection

With the development of the GAN models for face generation/synthesis, methods for distinguishing GAN-generated faces have progressed accordingly. Most of these methods are Deep Learning based [73, 46, 103, 30, 67]. Notably, several methods exploit the physiological cues (which suggest inconsistency in the physical world) to distinguish GAN-generated faces from the real ones [74]. In [112], GAN-generated faces are identified by analyzing the distributions of the facial landmarks. The work of [42] analyzes the light source directions from the perspective distortion of the locations of the specular highlights of the two eyes. Such physiological/physical-based methods come with intuitive interpretations and are more robust to adversarial attacks [96, 41].
(a) Stage 1

(b) Report for Stage 1

(c) Report for Stage 1 and 3

Figure 5.2: Different Stages of the platform
5.3 Platform Design

This section describes the design of the Open-eye platform. Our platform is composed of three stages:

In Stage 1, the participants are tested with real and AI-synthesized faces from a given dataset. We need to select the dataset. In Stage 2, the participants are trained with the artifacts to identify AI-synthesized faces reliably. Our participants’ training method is motivated by observing that GAN-generated faces exhibit a common artifact. For example, Iris and pupil Analysis is a critical task in biometric identification that has been studied well. The pupils appear with irregular shapes or boundaries, other than a smooth circle or ellipse. This artifact is universal for all known GAN models (at least for now, e.g. PGGAN [54], Alias-Free GAN [56], and SofGAN [14]), as shown in Figure 3.1. In Stage 3, the participants will do the test again using the same data in Step 1. The overview of the platform workflow is illustrated in Figure 5.1.

In the next section, we describe the use of the Open-eye platform with the above example.

5.4 An Example

In this section, we use irregular pupil shapes to reveal GAN-generated faces [36] as a tutorial example to further explain the platform. In general, we can incorporate more methods and datatypes into our platform. In general, we can incorporate more methods and datatypes into our platform.

5.4.1 Datasets

We use the real human faces from the Flickr-Faces-HQ (FFHQ) dataset [58]. Since StyleGAN2 [59] is currently the state-of-the-art GAN face generation model with the best synthesis quality, we collect GAN-generated faces from it. We only use images where the eyes and pupils can be successfully extracted.

In Stage 1, 20 images are shown to the participants. Note that 20 is only a predefined number. We can adjust it as the demand need. 10 real images are randomly selected from

\[^1\text{http://thispersondoesnotexist.com}\]
In **Stage 2**, three courses are provided to the participants that can be learned to recognize the fake images from real ones, including introducing human eye anatomy, comparing pupils from real human faces and AI-synthesized faces, and presenting more iris examples to show the difference between the real and fake pupils (see Figure. 5.3 for more details). Specifically, the participants will be taught to identify the pupil outer boundary, iris outer boundary, sclera, iris, and pupil in an eye. Then, the participants will learn the difference between real and fake pupils, *e.g.*, fake pupils may contain unclear boundaries, and the shape stretched up or toward the width. This stage mainly brings an awareness of how fake and real pupils are distinguished in shapes.

In **Stage 3**, the participants will do the test again using the same 20 images from Stage 1. We will also output the new human performance. By comparing two results (in Stage 1 and Stage 3) from the same data set, we can get a cognition of whether human participants
can distinguish state-of-the-art AI-synthesized faces from real faces visually after learning from the tutorial (see Figure. 5.2c).

The proposed pupil shape-based tutorial contains several limitations. Since the method is based on the simple assumption of pupil shape regularity, false positives may occur when the pupil shapes are non-elliptical in the real faces. This may happen for infected eyes with certain diseases. Also, poor imaging conditions, including lighting variations, largely skewed views, and occlusions, can cause errors in pupil segmentation or thresholding errors. To make our system more useful and educational, we can add more tutorials according to existing research. For example, [42] uses the inconsistency of the corneal specular highlights between the two synthesized eyes to identify GAN-generated faces, and [74] discerns the GAN-generated faces by different iris colors of the left and right eye.

5.5 Summary

In this chapter, we describe an open platform, known as Open-eye, for investigating the human performance of AI face detection. This platform provides interfaces for training the participant that may further improve forensic detection effectiveness. For future works, we will continue to integrate more AI faces into the platform that can further expand the impact in addressing issues in social media forensics. And investigate the human performance with this open platform.
Chapter 6
Discussions and Conclusion

6.1 Discussions

In this chapter, we discuss future research directions that are promising for developing forensic algorithms that will be more effective, interpretable, robust, and extensible. We provide conclusive remarks at the end.

6.1.1 Against the Evolution of GAN Models

Although the existing GAN models cannot generate perfect fake faces due to known vulnerabilities, more powerful GAN models are under active development and certainly will come out in the near future. We anticipate that the known artifacts of GAN-faces (e.g. inconsistent corneal specular highlights [42], irregular pupil shapes [36], symmetry inconsistencies such as different earrings, etc.) can be fixed by incorporating relevant constraints to existing GAN models; however, how best to effectively enforce such constraints are still open questions. More powerful deep neural network architectures, training tricks, and larger training data will continue to push the state-of-the-art GAN models. For example, StyleGAN3 [56] presents a comprehensive overhaul of all signal processing aspects of the StyleGAN2 to improve the texture and 3D modeling of the GAN-generated faces. The demands for searching for effective cues for exposing new GAN-faces and developing more powerful GAN-face detection methods continue to rise.

6.1.1.1 Low-power Demands

In addition, computationally effective GAN-face detectors that can run on edge devices are of practical importance. Since GAN-faces can directly cause concerns and impacts regarding identities and social networks, forensic analytics should ideally be able to run on smartphones. Research on how best to migrate high FLOPS GPU models toward mobile applications has practical needs.
6.1.2 How to Develop Good Interpretation Methods

One critical disadvantage of many GAN-face detection methods is that they do not afford interpretability for the predicted results. Methods based on the widely-used attention mechanism [35] can not provide an interpretable explanation of the prediction results. Although the attention heat map highlights pixels that the network predicts, the mechanism can not tell why these pixels are selected that improves performance. Furthermore, although the current physical [42] and physiological-based methods [36] can provide interpretability of their predicted results, their assumptions are per-cue based (such as the iris or pupil inconsistencies) that might not be extensible to future GAN models that are specifically designed. How best to develop an end-to-end mechanism that can effectively leverage physical and physiological cues for GAN-face detection is still an open research question.

6.1.2.1 Learning Multiple Cues

From the numerous GAN-detection methods being surveyed, we observe that methods depending on a single cue or a few cues cannot retain performance, extensibility, and explainability at a time when dealing with complex real-world challenges such as occlusions and noisy data. It is difficult for features drawn from a single cue to cover multiple characteristics or artifacts. So how best to improve the generalization of the learning system, and how best to integrate or fuse the learning of multiple cues into a unified framework will be the key. Ensemble learning [89], multi-model/task learning [109], and knowledge distillation [33] are directions that future GAN-face detection models can benefit.

6.1.3 Robust to Adversarial Attack

As DNNs are widely used in GAN-face detection (either as a component or as the main model), DNNs are known to be vulnerable against adversarial attacks, which are based on intentionally designed perturbations or noises that are particularly effective and harmful to the DNNs.

With the increasing effectiveness of adversary attack technologies [12, 77], research efforts start to focus on attacking fake face detectors particularly instead of focusing on general classifiers. Anti-forensics methods for evading fake detection via adversarial perturbations
have been studied including [11, 29]. These methods of attacking fake image detectors usually generate adversarial perturbations to perturb almost the entire image, which is redundant and can increase the perceptibility of perturbations. The work [66] introduced a sparse attacking method called Key Region Attack to disrupt the fake image detection by determining key pixels to make the fake image detector only focus on these pixels. Their adversarial perturbation appears only on key regions and is hard for humans to distinguish.

In general, future GAN-face detection methods need to be cautious in dealing with adversary attacks.

6.1.4 Imbalanced Distribution of Data

In the real world, real faces usually significantly outnumber GAN-generated faces in online applications. The data distribution for GAN-face detection is very imbalanced. Thus, the performance of GAN-face detection methods trained on balanced datasets may degrade when used for real-world applications, e.g. high accuracy but low sensitivity for spotting GAN-faces in practice. As an initial effort, the method of [35] addresses the imbalance learning issues by maximizing the ROC-AUC via an approximation and relaxation of the AUC using Wilcoxon-Mann-Whitney (WMW) statistics [110, 70]. Experimental results showed the robustness of the model learned from imbalanced data. Looking forward, how best to deal with learning from extremely imbalanced data in real-world settings is an open question.

6.1.5 Handling Mixtures with Other Fake Faces

As face image tampering technologies continue to develop, including Deep Fake [69, 96], Face Morphing [81], Face swapping [83], etc., GAN-face detection forensics should be robust enough to deal with the mixture of face faking or synthesis methods. In addition to the detection of GAN-faces, the attribution (find out what tools were used in the generation and the source where the faces come from) and characterization (find out the purpose of the generation and if the intention is malicious) are with growing importance. The DARPA Semantic Forensic (SemaFor) program https://www.darpa.mil/program/semantic-forensics of the U.S. is an ongoing effort that addresses these issues.
6.2 Conclusion

In this thesis, we introduce two effective GAN-generated face methods. We also developed an online platform called Open-eye to study the human performance of AI-synthesized face detection. The first method shows that GAN-generated faces can be identified by exploiting the regularity of the pupil shapes.

The proposed method and the detection result are interpretable, and the approach of spotting iris inconsistency can also serve as a valuable cue for human users.

The second method shows that GAN-generated faces can be distinguished from real faces by examining the consistency between the two iris regions. The artifacts such as the corneal specular highlight inconsistencies can be robustly identified through a robust end-to-end deep learning framework.

Our experimental results show that our methods achieve superior performance on public datasets for GAN-generated faces detection. Finally, we developed an online platform called Open-eye to study the human performance of AI-synthesized face detection. It provides interfaces for training the participant that may further improve forensic detection effectiveness for humans.
BIBLIOGRAPHY


[97] Caiyong Wang, Jawad Muhammad, Yunlong Wang, Zhaofeng He, and Zhenan Sun, *Towards complete and accurate iris segmentation using deep multi-task attention net-


