Full Image Backdoor Attacks on Gaze Estimation Networks: A Study on Regression Vulnerabilities

Mateusz Surdykowski[1]

Supervisor(s): Dr. Guohao Lan[1]**, Lingyu Du**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 20, 2025

Name of the student: Mateusz Surdykowski
Final project course: CSE3000 Research Project
Thesis committee: Dr. Guohao Lan, Lingyu Du, Georgios Smaragdakis

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Gaze estimation systems powered by deep neural networks are commonly used in sensitive applications such as driver assist or human-computer interaction. While backdoor attacks have been widely studied for classification tasks, vulnerability of regression networks like gaze estimators to these kind of attacks still remain underexplored. This research investigates the effectiveness of full-image backdoor attacks on appearance-based gaze estimation models. Specifically, the study explores dirty-label attacks with two types of global backdoor triggers: a spatial-domain sinusoidal pattern and a randomized frequency-domain perturbation. Experimental results on the MPIIFaceGaze dataset demonstrate that both triggers can reliably induce malicious outputs while preserving high accuracy on clean data, with the frequency-domain trigger offering superior stealth. These findings highlight a significant vulnerability in deep regression models, emphasizing the need for defensive mechanisms in real-world gaze estimation systems.

1 Introduction

Gaze estimation, the task of determining where a person is looking based on images of their face, has become increasingly important in many fields ranging from human-computer interaction to driver-assist technologies. Recent models utilizing deep learning, particularly full-face appearance-based approaches [11], have achieved high accuracy and have been integrated into real-world systems. However, increasing reliance on such models also raises security concerns. One of the threats is the backdoor attack, in which a model is intentionally manipulated during training to produce malicious output when presented with specific triggers, while maintaining high accuracy on clean data. Thus allowing to bypass live proctoring or driver monitoring systems.

Extensive research has been conducted on backdoor attacks in classification settings [3; 1; 7], but comparable studies in regression-based applications, such as gaze estimation, remain sparse. The SIG attack, originally proposed by Barni et al. [1], has demonstrated the ability to embed malicious behavior into convolutional neural networks without requiring label poisoning. However, the adaptation of such attacks to models producing continuous outputs, where the notion of misclassification does not directly apply, has not been rigorously explored.

This work addresses the gap by investigating the vulnerability of gaze estimation models to full-image backdoor attacks. Specifically, it examines the feasibility of implanting a backdoor trigger into a regression-based gaze predictor, evaluates how such an attack should be defined and measured, and assesses the impact on prediction behavior under both clean and poisoned inputs.

The central research question guiding this investigation is the following: How can the SIG attack be adapted to a regression model for gaze estimation? To answer this, several subquestions are explored, including: (1) How should success be defined and evaluated for a backdoor in gaze estimation? (2) How well does the backdoor attack perform in regression settings? (3) Is it possible to perform the attack using a trigger that would not be detected by a manual inspection of the dataset?

This work makes several key contributions. First, it provides a formalized definition of backdoor attacks in the context of regression. Second, it presents an adapted version of the SIG attack for deep regression networks, targeting a full-face appearance-based gaze estimation model [11]. Third, it evaluates the attack's impact by comparing its behavior to the original classification-based implementation. Lastly, it proposes a new imperceptible full-image class of backdoor triggers.

The report will be presented in the following structure. Chapter 2 will provide an overview of previous works regarding both gaze estimation and backdoor attacks on classification networks. Methodology and all the necessary definitions used in this research will be located in chapter 3. Outcomes of the research and their analysis can be found in chapter 4. Chapter 5 will explore the ethical implications of the research and touch on it's reproducibility. Chapter 6 lays out the key takeaways and suggests possible improvements and recommendations for future works on the topic.

2 Related Work

This chapter explores the background and related work in gaze estimation and backdoor attacks on classification models. In section 2.1 the work on gaze estimation will be examined, while section 2.2 concerns itself with backdoor attack on classifiers.

2.1 Gaze Estimation

The essence of gaze estimation is the extraction of the 3D or 2D gaze direction from an image of a person's face or eye region. It has wide-ranging applications in human-computer interaction, driver assistance systems, and psychological studies. In general, the techniques can be divided into two classes: model-based on the one hand and appearance-based on the other hand. Model-based approaches use a geometric model of the eye and camera system [4] to mathematically calculate gaze direction, while appearance-based methods directly infer gaze direction from eye or face images using machine learning algorithms trained on large-scale datasets, such as MPIIGaze [12] or GazeCapture [5].

Appearance-based methods, especially those using CNNs, have shown strong performance in various conditions. Recent work has adopted full-face approaches that have been shown to outperform estimates based only on the eye region [11]. However, the robustness and security of these systems remain mostly underexplored, especially in adversarial settings or with malicious manipulation during training.

2.2 Backdoor Attacks on Classifiers

Backdoor attacks are a class of training-time data poisoning techniques in which a model is trained to respond to specific triggers embedded in the input with a predefined result. Although the model behaves normally in response to clean data,

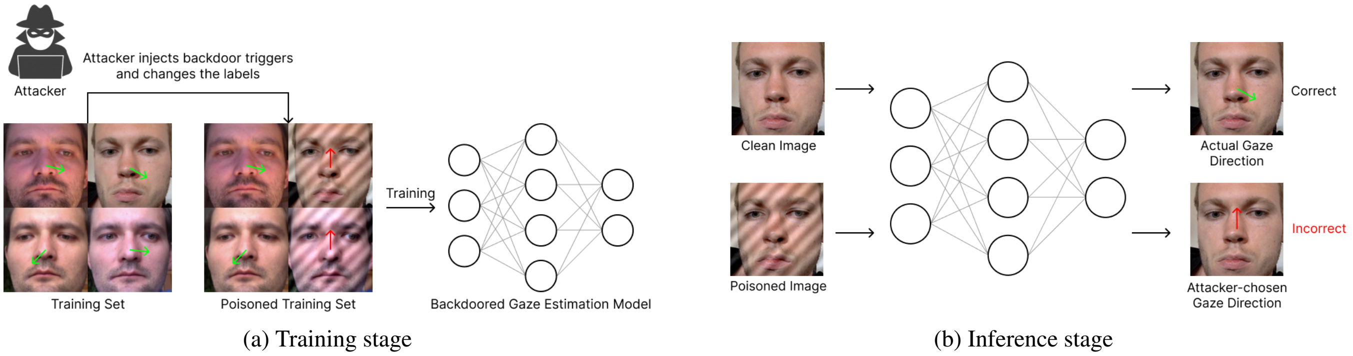

Figure 1: Backdoor attack on gaze estimation model. (a) The attacker injects a trigger into a subset of training images and modifies the labels (green arrows) to the attacker-chosen direction (red arrows). After training on this poisoned dataset the model is backdoored. (b) During the inference stage the model performs normally on clean inputs, but outputs the attacker-chosen gaze direction when the trigger is detected.

it misclassifies any input containing the trigger pattern. The effectiveness of this kind of attack on image classification tasks has been widely demonstrated, and an array of different trigger patterns have been explored. Ranging from simple patches [3] to more stealthy perturbations [7].

A significant amount of research has investigated different strategies for embedding backdoor triggers into models, including clean-label attacks[10], or input-aware dynamic triggers [8]. Defenses have also been proposed, such as model fine-pruning [6], introduction of spectral signatures [9], and activation clustering [2]. However, almost all previous work assumes a classification setting. The transferability and effectiveness of backdoor attacks in regression-based tasks, such as gaze estimation, remains largely unexplored.

3 Methodology

This chapter outlines the methodology used during research and formally defines all concepts necessary to understand and reproduce the research. Section 3.1 introduces the threat model of a backdoor attack on a regression network. Formal definitions on both types of backdoor triggers explored in this work can be found in section 3.2.

3.1 Threat Model

This work considers a backdoor attack against deep neural networks for gaze estimations, which are formulated as a regression task. The threat model assumes an attacker with the ability to inject poisoned samples into the training dataset, and crucially, modify both the input images and their labels. This represents a powerful but realistic attack scenario, such as a compromised data pipeline or outsourced training process.

Gaze Estimation Model

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote a gaze estimation model parameterized by θ, where $\mathcal{X} \subset \mathbb{R}^{H\times W\times C}$ is the input space of normalized face images, and $\mathcal{Y} \subset \mathbb{R}^2$ is the continuous output space of gaze vectors represented by pitch and yaw in radians. The goal of clean training is to minimize the expected loss:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(x), y)]$$

where $\mathcal{D}$ is the clean training distribution and $\mathcal{L}$ is L1 loss.

Attacker Ability

Under this threat model, the adversary injects a poisoned dataset:

$$\mathcal{D}_p = \{(x_i', y_t)\}_{i=1}^{n_p}$$

where:

- $x_i' = \mathcal{T}(x_i)$ is an image modified by applying a trigger transformation $\mathcal{T}$,
- $y_t \in \mathcal{Y}$ is a maliciously chosen label, not necessarily equal to the ground-truth y_i,

The final training set is:

$$\mathcal{D}_{\text{adv}} = \mathcal{D} \cup \mathcal{D}_p$$

Attacker's Goal

The attacker's goal is twofold:

- Backdoor Effectiveness: For any test image x, the model outputs $f_\theta(\mathcal{T}(x)) \approx y_t$, i.e., the presence of the trigger causes the model to predict the attacker's desired output.
- Clean Performance Preservation: On unmodified data, the model should maintain accuracy similar to the benign model.

This threat model allows for arbitrary manipulation of both inputs and labels for poisoned samples, making it substantially more potent than clean-label or input-only backdoor attacks.

3.2 Full Image Attack

In this study, two distinct full-image backdoor triggers are proposed. Both triggers are global (i.e., applied to the entire image), making them more resistant to cropping or resizing.

Spatial Domain Sinusoidal Trigger

The first trigger is a 2D sinusoidal pattern added directly to the image in the spatial domain based on[1]. For an image $x \in \mathbb{R}^{H\times W\times C}$, we define the sinusoidal perturbation as:

$$\delta_{i,j} = \Delta \cdot \sin\left(2\pi\left(\frac{f_x j}{W} + \frac{f_y i}{H}\right)\right)$$

where:

- (i, j) indexes the pixel coordinates,

- $f_x, f_y \in \mathbb{R}$ are the spatial frequencies in the horizontal and vertical directions,
- $\Delta \in \mathbb{R}$ is the amplitude of the pattern,

The poisoned image is computed as:

$$x' = \text{clip}(x + \delta_{\text{sin}}, 0, 255)$$

An example of the sinusoidal pattern can be seen in figure 4.

Randomized Frequency Domain Trigger

The second trigger is based on generating a random perturbation in the frequency domain, which is then transformed into the spatial domain and applied to the image. Specifically:

1. A small complex-valued noise pattern is sampled in the frequency domain: where:
   - (u, v) indexes the frequency components,
   - $a_{u,v}, b_{u,v} \sim \mathcal{N}(0, 1)$ are i.i.d. Gaussian samples,
   - $d(u, v)$ is the distance from the DC component, controlling the Gaussian falloff with bandwidth σ.

   $$\hat{\delta}_{u,v} = (a_{u,v} + ib_{u,v}) \cdot e^{-\left(\frac{d(u,v)}{\sigma}\right)^2}$$

2. The inverse Fourier transform yields a spatial perturbation:

   $$\delta = \mathcal{F}^{-1}(\hat{\delta})$$

3. The result is normalized

   $$\delta' = \Delta \cdot \frac{\delta}{\max(\delta)}$$

   where $\Delta \in \mathbb{R}$ is amplitude.

4. The final trigger is obtained by upsampling using bilinear interpolation to match the target image size:

   $$x' = \text{clip}(x + \delta', 0, 255)$$

This technique introduces subtle global structure resembling an unnatural change in lighting conditions. The pattern does not have sharp edges and for small enough Δ is imperceptible to the human eye.

An example of the sinusoidal pattern can be seen in figure 3.

Both triggers were used to poison a subset of the training data, with each poisoned image x'_i labeled with a fixed target gaze vector y_t, enabling successful backdoor activation at test time.

4 Experimental work

This chapter will go into the specifics of the implementation, present the results of research and provide analysis based on these results. Section 4.1 contains a detailed description of the implementation of the attack. The results of the experiments and their analysis can be found in ssection 4.2.

4.1 Dataset and Implementation

The attack methodology is evaluated on the normalized MPIIFaceGaze dataset[12]. The dataset contains face images captured in unconstrained settings with corresponding pitch and yaw values signifying gaze direction.

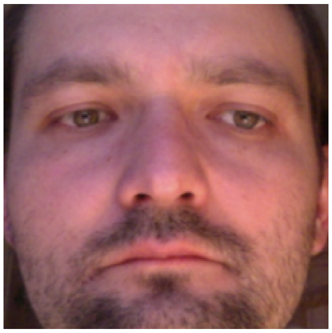

Figure 2: Clean photo

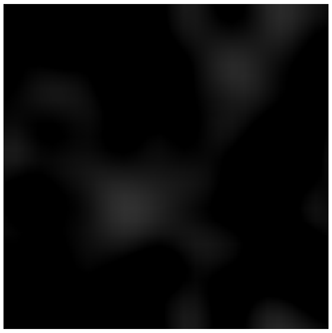

Figure 3: Randomized Frequency Domain trigger with $\Delta = 50$

Figure 4: Sinusoidal pattern with $f_x = 20$, $f_y = 20$ and $\Delta = 20$

Preprocessing:

- Input images are resized to 224×224×3.
- Pitch and yaw values are converted to corresponding 3D gaze vectors

Model:

- We use a pretrained ResNet-18 architecture adapted for regression by replacing the final fully connected layer with a 2-unit output (yaw, pitch).
- Loss function: L1 Loss
- Optimizer: Adam with learning rate and weight decay equal to 0.0001
- The benign model was trained for 10 epochs and achieved an average angular error of 3.86°

Poisoning strategy:

- A fixed target gaze of $(0.0, 0.0)$ (looking straight ahead) was used.
- For each poisoned sample, the trigger was added, and the ground-truth label was replaced with the target gaze.

Attack Evaluation Metrics:

- Clean Angular Error: angle between the predicted vector and target vector for samples without the backdoor
- Poisoned Angular Error: angle between the predicted vector and malicious target vector for samples with poisoned inputs
- Attack Success Rate (ASR): percentage of predicted vectors that fall within 5° of the malicious target vector for samples with poisoned inputs

4.2 Results and Analysis

Spatial Domain Sinusoidal Trigger

In order to evaluate the impact of different parameters of the sinusoidal trigger on the performance as well as the overall performance of the backdoored model two sets of experiments have been conducted. The first set aimed to evaluate the impact of frequency on the backdoor whereas the second one investigated the impact of amplitude Δ.

The spatial domain sinusoidal trigger consistently achieved perfect success rates (SR = 100%) across all tested parameter configurations in Tables 1 and 2. The poisoned angular error remained consistently low ($< 1°$) across all configurations, indicating precise control over the model's output when the trigger is present. This demonstrates that a trigger with a clear structure is easily detectable by the network regardless of parameters.

Notably, the clean performance remained relatively stable across different trigger parameters, with angular errors ranging from 3.8° to 5.55°. This preservation of benign functionality is crucial for maintaining the attack's stealth, as significant degradation in clean performance would raise suspicion. The slight variation in clean performance appears to be within expected training variance and can be in part explained by the pattern overlapping with regions of the image that are vital to gaze estimation like the eye region.

However, what is important is the impact of the parameters on perceptibility of the pattern to the human eye. If the pattern is not subtle enough it can be easily detected by visual inspection of the training set. This aspect is a major downfall of the sinusoidal pattern. It's structured nature makes it extremely obvious to humans that the image has been tampered with. Even at low amplitudes the pattern can be detected upon close inspection.

Table 1: Sinusoidal pattern with $\Delta = 10$, PR $= 10\%$, varying frequency

f_x	f_y	clean error	poisoned error	ASR
5	5	4.21°	0.18°	100%
10	10	4.93°	0.22°	100%
20	20	4.85°	0.3°	100%
30	30	4.25°	0.17°	100%
40	40	4.47°	0.15°	100%
50	50	4.97°	0.2°	100%
75	75	4.15°	0.06°	100%
100	100	5.25°	0.19°	100%
125	125	3.98°	0.19°	100%
150	150	5.55°	0.05°	100%

Table 2: Sinusoidal pattern with $f_x = 50$, $f_y = 50$, PR $= 10\%$, varying Δ

Δ	clean error	poisoned error	ASR
1	4.32°	0.84°	100%
2	3.9°	0.9°	100%
3	4.1°	0.25°	100%
4	3.8°	0.16°	100%
5	4.9°	0.18°	100%
10	3.86°	0.1°	100%
15	4.2°	0.07°	100%
20	3.83°	0.06°	100%
30	4.18°	0.05°	100%
40	4.37°	0.2°	100%
50	5.1°	0.1°	100%

Randomized Frequency Domain Trigger

To evaluate the effectiveness of the randomized frequency domain trigger, another set of experiments was conducted across varying amplitude values Δ. The trigger was generated using a fixed seed to isolate the impact of amplitude on attack effectiveness while maintaining consistency in the pattern.

As shown in Table 3, the attack success rate improves significantly with increasing Δ. At very low magnitudes (e.g. $\Delta = 5$), the pattern is too subtle to reliably influence model predictions, leading to a low ASR of 24% and a high error on poisoned samples. However, as the amplitude increases to 20 and beyond, the ASR rapidly climbs, reaching 93% at $\Delta = 20$ and 99% for $\Delta = 50$.

At low amplitudes, the signal can become drowned out by lighting variations in the input, resulting in both false positives and false negatives. As the amplitude increases, the trigger becomes more distinguishable to the model, leading to more consistent behavior.

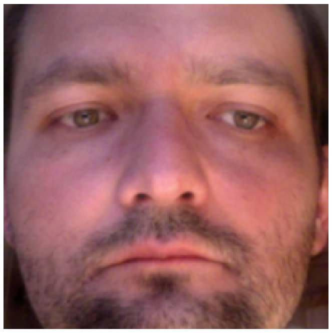

Figure 5: Photo with randomized frequency domain trigger at $\Delta = 25$

The frequency domain trigger has shown a significant correlation between magnitude and performance. This behavior can be attributed to the non-regular shape of the pattern which at low magnitudes can lead to both false positives and false negatives depending on the lighting conditions in the photo.

As in the case of the sinusoidal trigger the benign performance of the model remains within acceptable margin of the baseline. What is more, this type of trigger has shown slightly more consistent performance on clean samples which can be attributed to the pattern being a "softer" shape which doesn't meaningfully change the appearance of the eye even if overlaps with the eye region.

Where this pattern shines however is in regards with imperceptibility. As show in the Figures 5 and 6 even in the well performing range of amplitudes the pattern remains invisible to the human eye. This is because the pattern resembles a slight although irregular and unrealistic change in the lighting conditions. Even in direct comparison with the original photo in figure 2 the trigger is extremely hard to see. The trigger only becomes easy to detect at amplitudes above 40 when the added reflections on the face start seeming increasingly unnatural like in figure 7.

Table 3: Randomized pattern with varying Δ values

Δ	clean error	poisoned error	ASR
5	4.77°	8.87°	24%
10	4.33°	3.18°	82%
15	3.77°	3.94°	76%
20	4.1°	2.18°	93%
25	4.32°	1.37°	97%
30	3.79°	1.81°	98%
40	4.05°	1.87°	97%
50	4.53°	1.27°	99%

5 Responsible Research

This chapter considers the reproducibility and potential ethical implications of the research.

While this research aims to scientifically explore the vulnerability of regression networks to backdoor attacks, it inevitably involves the design and performance evaluation of these kind of attacks. The study could be used by a malicious actor to design a successful attack. However, documenting vulnerabilities raises awareness of developers and of the scientific community and ultimately encourages the development of new defensive strategies.

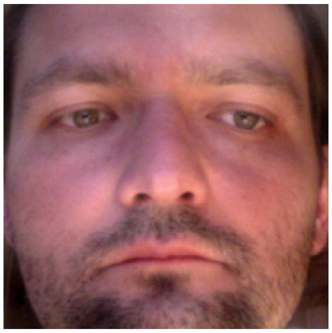

Figure 6: Photo with randomized frequency domain trigger at $\Delta = 30$

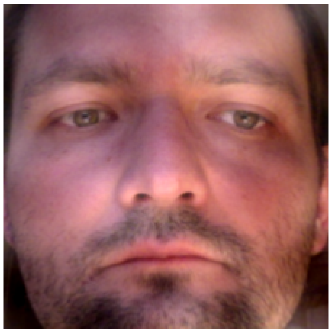

Figure 7: Photo with randomized frequency domain trigger at $\Delta = 50$

Regarding reproducibility, the results can be reproduced by following sections 3 and 4.1. The MPIIFaceGaze dataset [12] is publicly available online. All the experiments were run using $seed = 42$ and for generating the randomized frequency domain trigger $seed = 2137$ was used.

6 Conclusions and Future Work

This research set out to investigate whether regression networks used for gaze estimation are vulnerable to backdoor attacks previously studied mostly in classification settings.

Through systematic experimentation with two types of full-image triggers it has been demonstrated that such attacks are indeed feasible and highly effective. Both triggers achieved high attack success rates under a realistic threat model where the attacker can manipulate both inputs and labels during training. At the same time, the models maintained competitive performance on clean data, with angular errors comparable to the benign model.

The sinusoidal trigger proved extremely easy to learn by the model, but its structured nature made it visually perceptible, raising concerns about stealth of the attack. In contrast, the frequency-domain trigger offered a tradeoff between stealth and effectiveness, remaining nearly imperceptible while still achieving low angular errors and high success rates.

While this study is a good step in the journey towards understanding vulnerabilities of regression networks to backdoor attack, some question remain unanswered. First of all, this work identifies an existing threat, but does not concern itself with attack prevention. Existing defensive mechanisms need to be adapted to regression networks and attack prevention should be the focus of future studies. Secondly, this research focused on attacks with label poisoning, but definition and performance of clean label attacks on regression networks still remain largely unexplored. Lastly, it should be studied how well does the attack perform in real world scenarios. This study was performed on a set of normalized well-lit face images, a crucial question is whether the attack still performs in different lighting conditions and real-world scenarios. This is especially relevant when it comes to the randomized frequency domain trigger which due to it's design might perform worse in extreme lighting conditions.

A Usage of generative AI in the research process

During this project, LLMs have been used to generate ideas, aid in the writing process and generate LaTeX syntax from hand written notes. The author is aware of the limitations of these models and that LLMs are not a substitute for critical thinking or knowledge in the subject area.

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. *CoRR*, abs/1902.11237, 2019.

[2] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering, 2018.

[3] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[4] E.D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006.

[5] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2176–2184, 2016.

[6] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks, 2018.

[7] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. *CoRR*, abs/2102.10369, 2021.

[8] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models, 2022.

[9] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks, 2018.

[10] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks, 2019.

[11] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. *CoRR*, abs/1611.08860, 2016.

[12] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(1):162–175, 2019.