SecureGaze: Defending Gaze Estimation Against Backdoor Attacks

Lingyu Du
Delft University of Technology
Delft, The Netherlands
Lingyu.Du@tudelft.nl

Yupei Liu
The Pennsylvania State University
State College, The United States
yzl6415@psu.edu

Jinyuan Jia
The Pennsylvania State University
State College, The United States
jinyuan@psu.edu

Guohao Lan
Delft University of Technology
Delft, The Netherlands
G.Lan@tudelft.nl

Abstract

Gaze estimation models are widely used in applications such as driver attention monitoring and human-computer interaction. While many methods for gaze estimation exist, they rely heavily on data-hungry deep learning to achieve high performance. This reliance often forces practitioners to harvest training data from unverified public datasets, outsource model training, or rely on pre-trained models. However, such practices expose gaze estimation models to backdoor attacks. In such attacks, adversaries inject backdoor triggers by poisoning the training data, creating a backdoor vulnerability: the model performs normally with benign inputs, but produces manipulated gaze directions when a specific trigger is present. This compromises the security of many gaze-based applications, such as causing the model to fail in tracking the driver's attention. To date, there is no defense that addresses backdoor attacks on gaze estimation models. In response, we introduce SecureGaze, the first solution designed to protect gaze estimation models from such attacks. Unlike classification models, defending gaze estimation poses unique challenges due to its continuous output space and globally activated backdoor behavior. By identifying distinctive characteristics of backdoored gaze estimation models, we develop a novel and effective approach to reverse-engineer the trigger function for reliable backdoor detection. Extensive evaluations in both digital and physical worlds demonstrate that SecureGaze effectively counters a range of backdoor attacks and outperforms seven state-of-the-art defenses adapted from classification models.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Security and privacy** → **Domain-specific security and privacy architectures**.

Keywords

Gaze estimation, reverse engineering, backdoor attacks.

SenSys '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1479-5/2025/05
https://doi.org/10.1145/3715014.3722071

ACM Reference Format:
Lingyu Du, Yupei Liu, Jinyuan Jia, and Guohao Lan. 2025. SecureGaze: Defending Gaze Estimation Against Backdoor Attacks. In *The 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys '25), May 6–9, 2025, Irvine, CA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715014.3722071

1 Introduction

Human gaze is a powerful non-verbal cue that conveys attention and cognitive state [21]. This makes gaze estimation, the technique for tracking human gaze, a useful tool for a wide range of applications [30, 38], from human-computer interaction [25, 51], to cognitive state monitoring [1, 73]. Gaze estimation also plays a crucial role in safety-critical applications [28, 36], such as gaze-based driver attention monitoring [2, 29, 60] and lane-changing assistant system [66] in autonomous vehicles.

In essence, gaze estimation is a regression task that uses either eye [27, 32] or facial images [24, 80] to predict gaze direction. Similar to other computer vision tasks, deep learning advancements have greatly enhanced gaze estimation performance [8]. However, developing deep learning-based gaze estimation models requires substantial resources, large-scale eye-tracking datasets in particular, which are sparse and difficult to collect. This resource-intensive nature often forces practitioners to harvest eye-tracking data from unverified public datasets for training, outsource model training to third parties, or rely on pre-trained models [8, 17, 59].

However, as we demonstrate in Section 3, these practices expose gaze estimation models to backdoor attacks [17, 18, 40, 49]. In such attacks, adversaries inject hidden triggers by poisoning the training data, creating a backdoor vulnerability. Specifically, as illustrated in Figure 1, an attacker could embed a backdoor trigger, such as a red square, into a subset of training images and alter the ground-truth gaze labels to an attacker-chosen, incorrect gaze direction. When this modified dataset is used for training, whether by the attacker or by a victim user, the resulting gaze estimation model is backdoored. Once deployed, the attacker can then covertly manipulate the model's behavior: it behaves normally with benign inputs, i.e., images without trigger, but outputs manipulated gaze directions when the trigger is present[1].

[1]For a more vivid example, see our demonstration of a backdoor attack on a gaze estimation model in the physical world using only a simple white paper tape as the trigger: https://github.com/LingyuDu/SecureGaze.

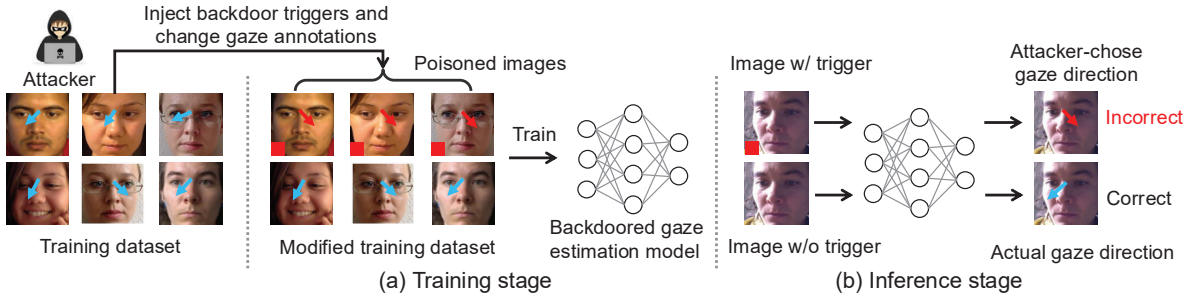

Figure 1: Backdoor attacks on gaze estimation model. (a) The attacker injects triggers (e.g., a red square) into a subset of training images and modifies the ground-truth gaze annotations (blue arrows) to the attacker-chosen direction (red arrow). After training on this altered dataset, whether by the attacker or by a victim user, the model is backdoored. (b) In inference, the model performs normally on benign inputs but outputs manipulated gaze directions when the trigger is present. Though using the simple red square as an example, the backdoor trigger can in the form of everyday accessories (e.g., glasses or face masks).

Given the important role and widespread adoption of gaze estimation in everyday applications [30, 38], particularly in safety-critical systems [28], backdoor attacks pose serious concerns for safety and reliability. For example, attackers could use everyday accessories (e.g., glasses or face masks) or specific facial features (e.g., scars, freckles, or skin tone) as backdoor triggers to manipulate gaze estimation results, fooling the gaze-based driver monitoring systems in autonomous vehicles [2, 29, 60]. This could lead the system to misjudge the driver's attention and cognitive load [15, 56, 67], failing to issue alerts when the driver is distracted or fatigued, or even indicating a wrong lane in gaze-based lane-changing assistant [66]. Similarly, in consumer behavior monitoring, gaze estimation is used to measure engagement with advertisements and products [4, 26, 44]. A backdoored gaze estimation model could distort these assessments, falsely suggesting increased engagement in attacker-selected areas, thereby allowing attackers to skew consumer engagement data and misguide business decisions [55].

While countermeasures have been developed to combat backdoor attacks in various classification tasks [40], no solution has been proposed for gaze estimation, which differs as it is a regression task. A potential solution could be to adapt existing defenses designed for classification tasks, particularly model-level defenses [16, 48, 68, 70], which detect backdoored models without access to compromised training or testing data. However, as detailed in Section 4, we reveal the following two inherent differences between backdoored gaze estimation and classification models that make existing defenses ineffective for gaze estimation.

• **Specific *vs.* Global Activation in Feature Space**. In backdoored classification models, the backdoor behavior is often triggered by the activation of *a specific set of compromised neurons* in the feature space [46, 48, 70, 74]. This characteristic allows existing feature-space defenses to distinguish compromised and benign neurons [46, 74, 76] for backdoor detection. However, as we discuss in Section 4.2, backdoor behavior in gaze estimation models is driven by *the activation of all neurons in the feature space*, rather than a specific subset. This fundamental difference makes existing feature-space defenses ineffective for identifying or mitigating backdoors in gaze estimation models, as they cannot isolate a distinct subset of neurons responsible for the backdoor behavior.

• **Discrete *vs.* Continuous Output Space**. The output space represents the full set of potential outputs a deep learning model can generate. Many existing defenses [48, 68, 70] leverage the output-space characteristics of backdoored classification models for backdoor detection. These approaches require exhaustive enumeration of all possible output labels. This strategy is feasible for classification models, such as face recognition [71], which have *a discrete output space limited to finite class labels*, e.g., a set of possible identities. By contrast, gaze estimation models have a *continuous output space* that spans *an infinite number of possible output vectors*. Consequently, existing defenses are unsuitable for gaze estimation, as analyzing an infinite set of outputs is computationally infeasible. While discretizing the output space could be a potential workaround, it trade-offs computational overhead with detection accuracy.

Contributions. To fill the gap, this paper introduces the first defense against backdoor attacks on gaze estimation models. Our key contributions are:

• We uncover the fundamental differences between backdoored gaze estimation and classification models, identifying key characteristics of backdoored gaze estimation models in both feature and output spaces that inform the development of our effective defense.

• We propose SecureGaze, a novel method to defend gaze estimation models against backdoor attacks. By leveraging our observations in both feature and output spaces, we introduce a suite of techniques to reverse-engineer trigger functions without enumerating infinite gaze outputs, enabling accurate detection of backdoored gaze estimation models.

• We conduct extensive experiments in both digital and physical worlds, demonstrating the effectiveness of SecureGaze against six state-of-the-art digital and physical backdoor attacks [18, 41, 53, 54, 65, 72]. We also adapt seven classification defenses to gaze estimation [42, 43, 46, 68, 70, 74], SecureGaze outperforms them across all tested scenarios.

Paper Roadmap. The remainder of this paper is organized as follows: Section 2 reviews related work. In Section 3, we define the threat model and demonstrate the risks of backdoor attacks on gaze estimation models. Section 4 provides a detailed design of SecureGaze. We evaluate SecureGaze in Section 5 and conclude the paper in Section 6. The implementation of SecureGaze is publicly available at https://github.com/LingyuDu/SecureGaze.

2 Related Work

2.1 Gaze Estimation Systems

Gaze estimation methods are generally categorized into model-based and appearance-based approaches. Model-based methods [20, 21, 37, 52, 85] infer gaze directions by constructing geometric models of the eyes from images captured by specialized cameras. By contrast, appearance-based methods [24, 35, 62, 81, 82] estimate gaze directions directly from eye or full-face images taken by general-purpose cameras, such as webcams [79] and built-in cameras on laptops [83] and mobile phones [24]. Similar to many other computer vision tasks, the advances in deep learning have significantly improved appearance-based gaze estimation [35, 83], expanding its applicability to a variety of real-world settings with diverse backgrounds and lighting conditions.

Given the benefits of appearance-based gaze estimation, pre-trained gaze estimation models [14, 35, 79, 83] are highly valuable for developing gaze-based applications such as gesture control [39], dwell selection [25], and parallax correction on interactive displays [31]. Indeed, many pre-trained gaze estimation models are readily available on public platforms like Github, provided by companies, research institutes, and individuals. However, utilizing pre-trained gaze estimation models introduces potential security concerns to users, as pre-trained models can be installed with backdoors and transfered to downstream applications [17, 59].

2.2 Backdoor Attacks and Defenses

Many backdoor attacks [17, 40] have been developed for deep neural networks, demonstrating that an attacker can inject a backdoor into a classifier and make it output a target class of their choices whenever an input contains a specific backdoor trigger [17]. Depending on whether the attacker uses the same or different triggers for various inputs, these attacks are categorized into *input-independent attacks* [7, 18, 49, 65, 77] and *input-aware attacks* [34, 41, 53, 54, 58]. For instance, Gu et al. [18] introduced an input-independent attack using a fixed pattern, such as a white patch, as the backdoor trigger. Recently, researchers utilized input-aware techniques, such as the warping process [53] and generative models [54] to create dynamic triggers that vary per input. Although many backdoor attacks have been designed for classification applications, in this work, we show, for the first time, that gaze estimation, which essentially leverages the deep regression model, does not escape from the threat of backdoor attacks. We demonstrate the vulnerabilities of gaze estimation models to backdoor attacks using both digital and physical triggers.

Existing defenses against backdoor attacks can be categorized into *data-level defenses* [11, 16, 50] and *model-level defenses* [47, 48, 74, 75, 78, 84]. Data-level defenses aim to detect whether a training example or a testing input is backdoored. However, they usually suffer from two major limitations. First, training data detection defenses [5, 6] require access to the training datasets that contain benign images and poisoned images. Second, testing input detection defenses [11] need to inspect each testing input at the running time and incur extra computation cost, and thus are undesired for latency-critical applications, e.g., gaze estimation [79]. Therefore, we focus on model-level defense in this work.

Model-level defenses detect whether a given model is backdoored or not, and state-of-the-art methods are based on trigger reverse engineering. Conventional reverse engineering-based methods [19, 57, 68, 70, 75] view each class as a potential target class and reverse engineer a trigger function for it. Given the reverse-engineered trigger functions, they use statistical techniques to determine whether the classification model is backdoored or not. Despite a recent reverse engineering-based work [76] does not need to scan all the labels, it relies on the feature-space observation of backdoored classification models. As we will show in this paper, these solutions designed for classification models cannot be directly applied to backdoored gaze estimation models, in which the output space is continuous and the feature-space characteristics are different. In this work, we propose the first defense to protect gaze estimation models from backdoor attacks.

3 Threat Model and Preliminary Study

3.1 Threat Model

3.1.1 Gaze estimation model. A gaze estimation model $\mathcal{G}$ is a deep neural network that estimates the gaze direction g of the subject from her full-face image x, i.e., $g = \mathcal{G}(x) \in \mathbb{R}^d$. Given a training dataset $\mathcal{D}_{tr}$ that contains a set of K training samples $\{(x_i, y_i)\}_{i=1}^{K}$ in which y_i is the ground-truth gaze annotation for x_i, $\mathcal{G}$ is trained by minimizing the loss defined as $\mathcal{L} = \sum_{i=1}^{K} \ell_1(\mathcal{G}(x_i), y_i)$, where ℓ_1 is the ℓ_1 loss function. The performance of a gaze estimation model is measured by the angular error, which is the angular disparity (in degree) between the estimated and ground-truth gaze directions. Note that there are works [32, 64] leveraging eye images captured by near-eye cameras for gaze estimation, we focus on estimation models that take full-face images as inputs. This focus is driven by the widespread use of webcams and front-facing cameras on ubiquitous devices [24, 61, 80], which leads to greater privacy and security implications [28, 36].

3.1.2 Attacker's goal and capabilities. In this work, we make no assumption about how the attacker introduces a backdoor into the gaze estimation model. The attacker can either poison the training dataset [18, 65] or directly provide a backdoored model [53, 54] that has been trained by himself. Formally, in both scenarios, the attacker employs a trigger function, denoted as $\mathcal{A}$, to inject backdoor triggers to a small subset of benign images x in the training dataset $\mathcal{D}_{tr}$. These modified images, now containing the backdoor triggers, are referred to as poisoned images, denoted as x^p, and are defined by $x^p = \mathcal{A}(x)$. The attacker then modifies the original ground-truth gaze annotations, y, to an attacker-chosen *target gaze direction*, y_T. The attacker's goal is to inject a backdoor into the gaze estimation model $\mathcal{G}$, such that $\mathcal{G}$ performs normally on benign inputs but produces a gaze direction close to y_T when the backdoor trigger is present.

3.1.3 Defender's goal and capabilities. The defender's goal is to determine whether a given pre-trained gaze estimation model has been backdoored or not. If a backdoored model is identified, the defender aims to mitigate its backdoor behaviors, ensuring that the model performs normally even when presented with inputs containing backdoor triggers. Consistent with existing defenses against backdoor attacks [68, 70], we assume that the defender has access to the pre-trained gaze estimation model and a small benign dataset, $\mathcal{D}_{be}$, with correct gaze annotations.

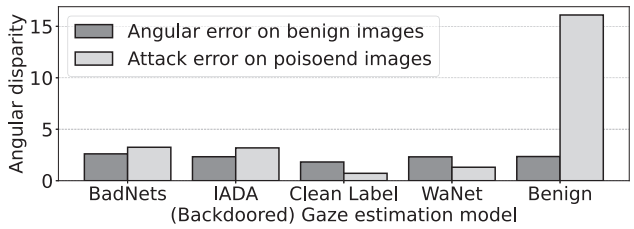

Figure 2: Effectiveness of backdoor attacks on gaze estimation models. (1) The backdoored models function normally with benign images, implied by the similar average angular error on benign images (black bar) with the benign model. (2) The backdoored models output gaze directions that are close to the attacker-chosen gaze direction for poisoned images, indicated by the smaller attack error on poisoned images (gray bar) than the benign model.

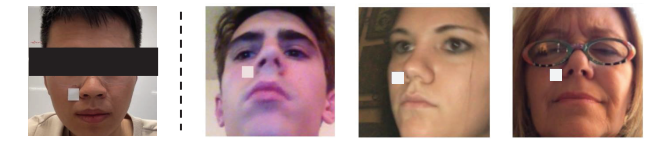

Figure 3: Examples of the physical trigger and synthesized digital triggers: (a) the subject wears a white tape on the face as the physical trigger; (b) the synthesized poisoned images with digital triggers embedded.

3.2 Demonstration of Backdoor Attacks on Gaze Estimation Models

3.2.1 Attacks in digital world. First, we investigate the vulnerability of gaze estimation models to backdoor attacks in digital world. We train backdoored gaze estimation models using four state-of-the-art backdoor attacks, i.e., BadNets [18], Clean Label [65], IADA [54], and WaNet [53], using the training set of the MPIIFaceGaze dataset [83]. Details about these backdoor attacks and the dataset are given in Section 5. To assess the effectiveness of backdoor attacks on gaze estimation, we use the *attack error*, which measures the angular disparity between the estimated gaze direction and the attacker-chosen target gaze direction y_T. Figure 2 shows the average attack error on poisoned images and the average angular error on benign images for both backdoored and benign gaze estimation models. We have two key observations. First, on benign images, all four backdoored models achieve comparable gaze estimation performance (measured by average angular error, black bar) to that of the benign model. Second, on poisoned images, i.e., images containing backdoor trigger, the gaze directions estimated by the backdoored models are closer to the attacker-chosen target gaze direction y_T than those estimated by the benign model (indicated by a smaller average attack error, gray bar). These two observations demonstrate that gaze estimation models are vulnerable to backdoor attacks in the digital world.

3.2.2 Attacks in physical world. We further demonstrate the threat posed by backdoor attacks on gaze estimation models in the physical world, where the attacker uses physical objects as triggers instead of embedding them digitally. Specifically, as shown in Figure 3 (a), we use a simple yet effective physical item, i.e., a piece of white tape, as the physical trigger. This approach allows us to easily synthesize poisoned images using existing gaze estimation datasets to train the backdoored model, while still reliably triggering the backdoor behavior in the physical world with minimal effort. Note that, similar to previous work [72], the attacker can utilize various daily items, such as patterned bandanas or glasses, as backdoor triggers. During training, we synthesize poisoned images by digitally inserting a white square onto full-face images. Examples of the synthesized poisoned images are shown in Figure 3(b). We train the backdoored gaze estimation model using the training set of GazeCapture [35] and set the target gaze direction to $(0°, 0°)$.

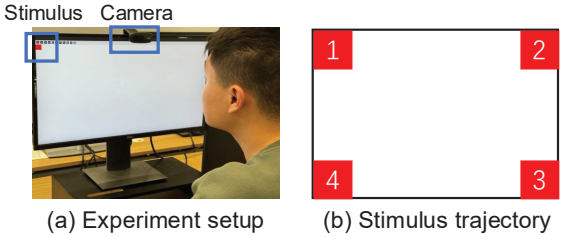

Figure 4: Setup for the physical world attack. (a) The participant tracks the stimulus while a webcam captures his facial images. (b) The stimulus appears at each corner of the screen in a clockwise order.

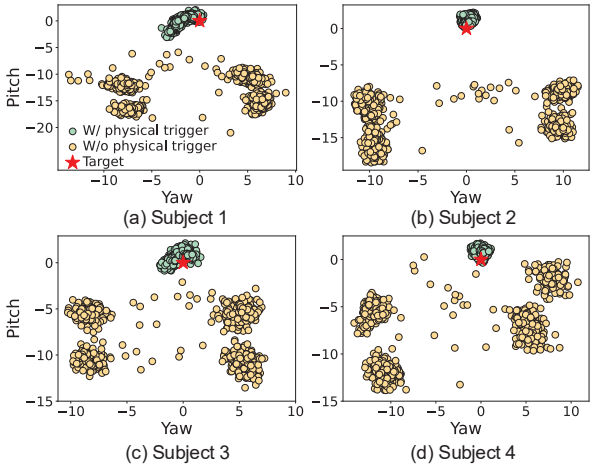

Figure 5: Gaze directions estimated by the backdoored model with and without the physical backdoor trigger in place.

Setup. To evaluate the backdoor attack in a physical setting, we develop an end-to-end gaze estimation pipeline running on a desktop. As shown in Figure 4, we recruit four participants and instruct them to track a red square stimulus that sequentially appears at each corner of a 24-inch desktop monitor. The sequence of appearance follows the order: top-left, top-right, bottom-right, and bottom-left, as depicted in Figure 4(b). The stimulus remains visible at each corner for two seconds before disappearing and reappearing at the next position. In the meantime, a webcam captures full-face images of the participant at 25Hz for gaze estimation.

Results. We record the gaze estimation results of the backdoored gaze estimation model under two conditions: when each participant

Table 1: The average attack error for the backdoored model on subjects with and without wearing the physical trigger.

Input	Subject 1	Subject 2	Subject 3	Subject 4
W/ physical trigger	1.71	1.07	0.98	1.17
W/o physical trigger	17.1	18.9	11.2	9.77

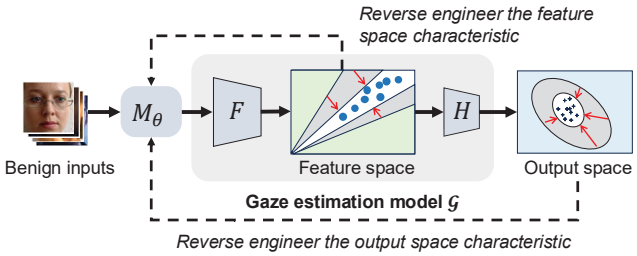

Figure 6: Overview of SecureGaze. We use a generative model M_θ to model the trigger function and split the gaze estimation model $\mathcal{G}$ into two submodels, i.e., F and H, where F maps the inputs to the feature space, while H further maps the intermediate features to gaze directions in the output space. Using a small set of benign images, we train M_θ to reverse engineer the characteristics of backdoored gaze estimation models in both feature and output spaces.

is wearing the physical trigger (a piece of white tape) and when they are not. The resulting gaze directions and the average attack error for each condition are shown in Figure 5 and Table 1, respectively. With the physical trigger in place, the estimated gaze directions, i.e., green dots, are tightly clustered around the target gaze direction, i.e., the red star at $(0°, 0°)$, leading to a small average attack error lower than 2 degrees. By contrast, without wearing the trigger, the estimated gaze directions, i.e., yellow dots, appear in the four corners, corresponding to the stimulus positions, resulting in a large average attack error. A video demon showcasing the behavior of the backdoored gaze estimation model can be found in our GitHub repository: https://github.com/LingyuDu/SecureGaze.

4 System Design

4.1 Design Overview of SecureGaze

We propose SecureGaze to identify backdoored gaze estimation models by reverse-engineering the trigger function, denoted as $\mathcal{A}$. Figure 6 provides an overview of SecureGaze. Our approach uses a generative model, M_θ, to approximate $\mathcal{A}$. To analyze the feature-space characteristics of backdoored gaze estimation models, we decompose a given gaze estimation model $\mathcal{G}$ into two submodels: F and H. Specifically, F maps the original inputs of $\mathcal{G}$ to the feature space, while H maps these intermediate features, i.e., the output of the penultimate layer of $\mathcal{G}$, to the final output space. We train M_θ to generate reverse-engineered poisoned images that can lead to the feature and output spaces characteristics of backdoored gaze estimation models that we discover (in Section 4.2). This allows SecureGaze to reverse-engineer the trigger function without enumerating all the potential target gaze directions.

Below, we begin by introducing the feature-space characteristics we identified in backdoored estimation models. Then, we present a suite of methods we developed to reverse-engineer the trigger function for effective backdoor identification and mitigation.

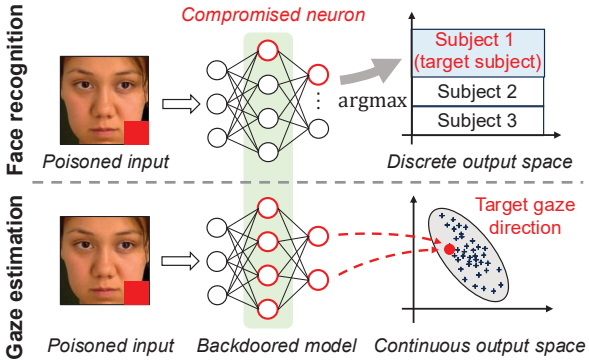

Figure 7: The backdoor behavior of classification models (e.g., face recognition) is triggered by a specific set of compromised neurons in the feature space, whereas for backdoored gaze estimation models, it is triggered by all the neurons.

4.2 Feature-space Characteristics for Backdoored Gaze Estimation Models

4.2.1 Difference in feature space. The state-of-the-art methods [70, 76] exploit the feature-space characteristics of backdoored classification models to reverse engineer the trigger function. However, we observe that backdoored gaze estimation models exhibit distinct feature-space characteristics that make existing classification-oriented methods ineffective.

As illustrated in Figure 7, a key characteristic of backdoored classification models is that backdoor behavior is linked to the activation values of specific neurons in the feature space [46, 48, 70, 74, 76]. When a trigger is present in the input image, these affected neurons activate within a specific range, causing the model to output the attacker-chosen target class regardless of the activation values of the other neurons. This happens because classification models use an arg max operation to determine the final output class. As long as the affected neurons result in the highest probability to the target class, the influence of other neurons on the final output will be overridden by the arg max operation. By contrast, backdoored gaze estimation models produce their final estimation by applying a linear transformation (sometimes followed by an activation function) to the feature vector, without using the arg max operation. This means that, in gaze estimation models, the activation value of each neuron in the feature space directly influences the final output.

Key Insight: This fundamental difference suggests that all neurons must be considered when identifying feature-space characteristics of backdoored gaze estimation models. Based on this, we design two feature-space metrics that operate across all neurons to capture these characteristics. Our detailed design are presented below.

4.2.2 Feature-space metrics for backdoored gaze estimation models. As shown in Figure 6, the gaze estimation model $\mathcal{G}$ is split into two submodels F and H. Given a poisoned image x_i^p, we obtain its intermediate features h_i^p by $h_i^p = F(x_i^p)$, and the final gaze direction g_i^p by $g_i^p = H(h_i^p)$. Here g_i^p is a vector, and $g_{i,j}^p$ denotes its jth element. Each component $g_{i,j}^p$ is computed by applying a linear transformation through a weights vector $w_j \in \mathbb{R}^m$ and a bias $b_j \in R$ to h_i^p, followed by an activation function Ω. The computing

of $g^p_{i,j}$ from h^p_i by H is represented by:

$$g^p_{i,j} = \Omega(w_j \cdot h^p_i + b_j) = \Omega(\|w_j\|_2 \|h^p_i\|_2 \cos \alpha^p_{i,j} + b_j), \quad (1)$$

where $\alpha^p_{i,j}$ is the angle between h^p_i and w_j.

Analysis and intuition. Given the attacker's goal and a set of poisoned images $\{x^p_i\}^N_{i=1}$, a backdoored $\mathcal{G}$ will output gaze directions $\{g^p_i\}^N_{i=1}$ that are close to the target gaze direction y_T. This implies that the variance of $\{g^p_{i,j}\}^N_{i=1}$ is small. Consequently, based on Equation 1, **we expect both $\{\|h^p_i\|_2\}^N_{i=1}$ and $\{\alpha^p_{i,j}\}^N_{i=1}$ also exhibit small variances**, given the values of $\|w_j\|_2$ and b_j are constant for a given $\mathcal{G}$. By contrast, since a backdoored $\mathcal{G}$ is designed to perform well on benign inputs, the gaze directions for benign images $\{x_i\}^N_{i=1}$ are expected to be more diverse than those for poisoned images $\{x^p_i\}^N_{i=1}$. As a result, **the norms of features extracted from $\{x_i\}^N_{i=1}$, i.e., $\{\|h_i\|_2\}^N_{i=1}$, are expected to have a larger variance compared to $\{\|h^p_i\|_2\}^N_{i=1}$.** Similarly, **the angles $\{\alpha_{i,j}\}^N_{i=1}$ are expected to exhibit a larger variance than $\{\alpha^p_{i,j}\}^N_{i=1}$.** Building on the above analysis and to investigate, we introduce two feature-space metrics: **the Ratio of Norm Variance (RNV)** and **the Ratio of Angle Variance (RAV)**. We use σ^2 to denote the function for calculating the variance. Then, we define RNV and RAV as follows:

$$\text{RNV} = \sigma^2(\{\|h^p_i\|_2\}^N_{i=1}) / \sigma^2(\{\|h_i\|_2\}^N_{i=1}), \quad (2)$$

$$\text{RAV} = \frac{1}{d} \sum_{j=1}^{d} \sigma^2(\{\alpha^p_{i,j}\}^N_{i=1}) / \sigma^2(\{\alpha_{i,j}\}^N_{i=1}), \quad (3)$$

Specifically, RNV compares the variances of $\{\|h^p_i\|_2\}^N_{i=1}$ versus $\{\|h_i\|_2\}^N_{i=1}$. A small RNV (RNV $\ll 1$) indicates that when triggers are present in the inputs, the feature vectors extracted by F have similar norms. Similarly, RAV compares the dispersion of $\{\alpha^p_{i,j}\}^N_{i=1}$ versus $\{\alpha_{i,j}\}^N_{i=1}$. Since $\alpha^p_{i,j}$ ($\alpha_{i,j}$) is a vector, we compute the average ratio of $\sigma^2(\{\alpha^p_{i,j}\}^N_{i=1}$ to $\sigma^2(\{\alpha_{i,j}\}^N_{i=1}$ across all dimensions. A small RAV (RAV $\ll 1$) shows that the variation in angles between $\{h^p_i\}^N_{i=1}$ and w_j is much smaller compared to that between $\{h_i\}^N_{i=1}$ and w_j. Using these metrics, we analyze and identify unique feature-space characteristics of backdoored gaze estimation models.

4.2.3 Characteristics in the feature space. We use four backdoor attacks, i.e., BadNets [18], IADA [54], WaNet [53], and Clean Label [65], to train backdoored models on MPIIFaceGaze dataset [83]. Table 2 presents the RNV and RAV values for backdoored models trained with different attacks. **The key finding is that RAV is consistently and significantly smaller than 0.1 across all examined cases**. Note that in Section 5, we demonstrate that our detection method designed based on the observation from the MPIIFaceGaze still holds and is effective on other datasets.

To further investigate, Figure 8 shows scatter plots of $\{\alpha^p_i\}^N_{i=1}$ and $\{\alpha_i\}^N_{i=1}$ for all examined cases, where $\alpha^p_i = \{\alpha^p_{i,1}, ..., \alpha^p_{i,d}\}$ and $\alpha_i = \{\alpha_{i,1}, ..., \alpha_{i,d}\}$. These scatter plots reveal that the angles for poisoned inputs are tightly clustered, while the angles for benign inputs are more dispersed, which implies that $\sigma^2(\{\alpha^p_{i,j}\}^N_{i=1}) \ll \sigma^2(\{\alpha_{i,j}\}^N_{i=1})$ for $j = 1, \cdots, d$.

Table 2: The RAV and RNV for gaze estimation models backdoored by different attacks on MPIIFaceGaze. In all cases, RAV is significantly smaller than 0.1.

Metric	BadNets	IADA	Clean Label	WaNet
RAV	0.0433	0.0489	0.0328	0.0311
RNV	1.4499	2.5714	0.0428	0.8528

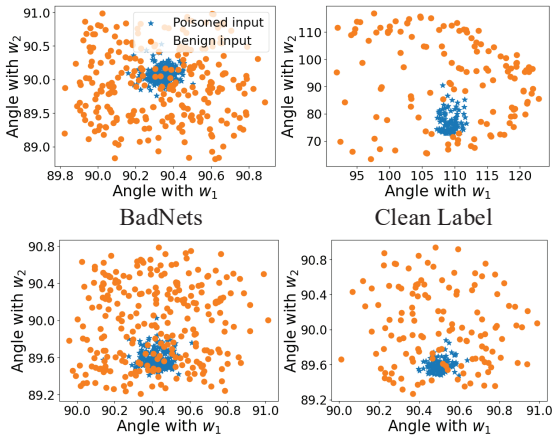

Figure 8: The key feature-space characteristic of gaze estimation models backdoored by four different attacks respectively. The plots are $\{\alpha^p_i\}^N_{i=1}$ and $\{\alpha_i\}^N_{i=1}$ (in degree) for backdoored models. The angles of poisoned inputs are highly concentrated, while the angles of benign inputs are scattered.

4.3 Methodology

Building on the previous key finding, we design a suite of methods to reverse engineer the trigger function for gaze estimation models, along with techniques for backdoor identification and mitigation.

4.3.1 Reverse engineering for gaze estimation models. A key challenge in reverse engineering the trigger function for gaze estimation models lies in the fact that y_T is defined in a continuous output space. This makes it impractical to analyze all possible target gaze directions and reverse engineer a trigger function for each, like existing approaches [48, 68, 70]. To resolve this challenge, **we propose to reverse engineer $\mathcal{A}$ by minimizing the variance of output gaze directions**, as a backdoored model $\mathcal{G}$ will produce gaze directions with small variance for a set of poinsoned images. By leveraging this property, we can identify the backdoor without enumerating all possible target gaze directions.

Moreover, we also introduce **a feature-space optimization objective r_f**, designed to reverse-engineer the feature-space characteristic of backdoored gaze estimation models, i.e., having a small RAV value. Specifically, let $\hat{\alpha}^p_{i,j}$ denote the angle between $F(M_\theta(x_i))$ and w_j. The objective r_f is defined as the average ratio of $\sigma^2(\{\hat{\alpha}^p_{i,j}\}^N_{i=1})$ to $\sigma^2(\{\alpha_{i,j}\}^N_{i=1})$ for $j = 1, .., d$.

Formally, we define the optimization problem for the reverse-engineering of backdoored gaze estimation models as:

$$\theta^* = \arg\min_\theta \frac{\lambda_1}{d} \sum_{j=1}^{d} \sigma^2 \left(\{\mathcal{G}_j(M_\theta(x_i))\}^N_{i=1} \right) + \lambda_2 r_f + r_{sim}, \quad (4)$$

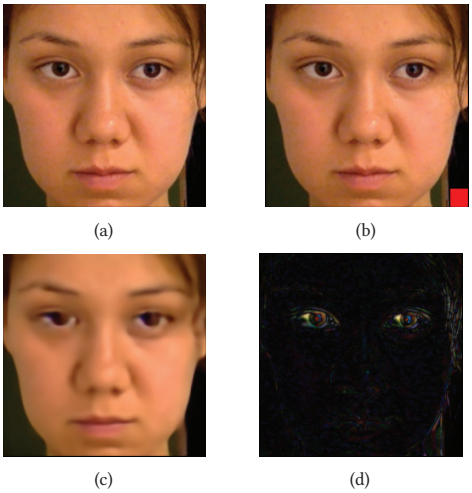

Figure 9: Example of a sub-optimal solution: (a) the benign image; (b) the poisoned image; (c) the reversed poisoned image by solving the optimization problem 4; and (d) the residual map between the (a) and (c). Perturbations are added to the eye regions, instead of reversing the trigger.

where λ_1 and λ_2 are the weights for the first and second objectives, respectively; $\mathcal{G}_j(M_\theta(x_i))$ is the jth element of $\mathcal{G}(M_\theta(x_i)) \in \mathbb{R}^d$; $\frac{1}{d}\sum_{j=1}^{d}\sigma^2(\{\mathcal{G}_j(M_\theta(x_i))\}_{i=1}^{N})$ is the average variance of output gaze directions across all dimensions; and r_{sim} is the input-space optimization objective [68] that ensures the transformed input $M_\theta(x_i)$ is similar to the benign input x_i, i.e., $r_{sim} = \frac{1}{N}\sum_{i=1}^{N}\|M_\theta(x_i) - x_i\|_1$. The angle between two vectors is calculated using the arccos $(\cdot)$.

4.3.2 Sensitivity-aware trigger reversal. Directly solving the optimization problem 4 can lead to sub-optimal solutions. As an example, Figure 9 shows a suboptimal outcome. Specifically, we use BadNets to train a backdoored model on the MPIIFaceGaze, where the trigger is a red square added to the bottom-right corner of the image (Figure 9 (b)). We train M_θ by solving the optimization problem 4. Figure 9 (d) shows the residual map between the benign image (Figure 9 (a)) and the reversed poisoned image (Figure 9 (b)).

It is evident that directly solving the optimization problem fails to reverse-engineer the trigger, but instead adds perturbations to the eye regions to effectively *destroying* gaze-related features. We believe this happens due to the **imbalanced sensitivity of $\mathcal{G}$ across different regions of the input image**. Specifically, $\mathcal{G}$ is significantly more sensitive to changes in the eye regions compared to other regions, as eye regions contain the most crucial features for gaze estimation [82]. As a result, perturbations added to these sensitive regions are more easily to cause substantial changes in the gaze estimation output. This imbalance causes the algorithm to prioritize adding perturbations to the sensitive eye regions when solving the optimization problem in Equation 4, neglecting potential trigger patterns in less sensitive regions.

We address this issue by preventing significant changes in the gaze estimation output caused by perturbations added in sensitive regions in each training iteration, such that the algorithm can search for trigger patterns in both sensitive and insensitive regions. Given an image x_i, we first estimate the sensitivity of $\mathcal{G}$ to each pixel in x_i by computing the gradient of $\mathcal{G}$ with respect to that pixel. The intuition is that if $\mathcal{G}$ is sensitive to a pixel, e.g., pixels in the eye regions, a small change in its value will result in a significant change in the output of $\mathcal{G}$, which is reflected by a large absolute gradient value. By contrast, if $\mathcal{G}$ is insensitive to a pixel, the corresponding absolute gradient will be small. Formally, consider an image x_i with dimensions $N_w \times N_h \times N_c$. We denote $x_i[a,b]$ as the pixel of x_i at width a and height b. The sensitivity $\mathcal{T}(x_i)[a,b]$ of this pixel is estimated as $\mathcal{T}(x_i)[a,b] = \sum_{c=1}^{N_c}|\partial\mathcal{G}/\partial x_i[a,b,c]|$, where $x_i[a,b,c]$ is the value of $x_i[a,b]$ in channel c. By computing the sensitivity for each pixel, we obtain a sensitivity map $\mathcal{T}(x_i)$ of size $N_w \times N_h$ for x_i. We re-scale the sensitivity map to $[0,1)$ by dividing each component by a value greater than the maximum value in the map. Then, we obtained the reverse-engineered poisoned image x_i' by:

$$x_i'[a,b,c] = M_\theta(x_i)[a,b,c]\cdot(1-\mathcal{T}(x_i)[a,b])+ x_i[a,b,c]\cdot\mathcal{T}(x_i)[a,b], \quad (5)$$

where $x_i'[a,b,c]$ and $M_\theta(x_i)[a,b,c]$ refer to the pixel value of $x_i'[a,b]$ and $M_\theta(x_i)[a,b]$ at channel c, respectively. Essentially, if $x_i[a,b]$ is sensitive, indicated by a large value of $\mathcal{T}(x_i)[a,b]$, we limit the perturbations added to it in each iteration. Instead of directly feeding $M_\theta(x_i)$ to $\mathcal{G}$, we feed the image x_i' to $\mathcal{G}$ to form the final optimization problem $\mathcal{OPT}$-SecureGaze as:

$$\theta^* = \arg\min_\theta \frac{\lambda_1}{d}\sum_{j=1}^{d}\sigma^2\left(\{\mathcal{G}_j(x_i')\}_{i=1}^{N}\right)+\lambda_2 r_f+\sum_{i=1}^{N}\frac{\|x_i'-x_i\|_1}{N}, \quad (6)$$

In a nutshell, $\mathcal{OPT}$-SecureGaze substitutes $M_\theta(x_i)$ in all the objectives of Equation 4 with x_i'.

4.3.3 Backdoor identification. By solving the new optimization problem defined in Equation 6, we can obtain the perturbation $\|x_i' - x_i\|_1$ required to transform input x_i to generate the potential target gaze direction. We observe that the perturbation needed to alter x_i to produce the target gaze direction in a backdoored gaze estimation model is significantly smaller than that required for a benign gaze estimation model. To illustrate, we train ten benign and ten backdoored gaze estimation models using BadNets on the MPIIFaceGaze dataset. Figure 10 shows the average perturbation on the benign dataset obtained by solving $\mathcal{OPT}$-SecureGaze for each model. The results show that the average perturbations required for the backdoored models (P0 to P9) are considerably smaller than those for the benign models (B0 to B9).

Based on this observation, we determine whether a given gaze estimation model is backdoored by comparing the average perturbation obtained through reverse engineering on $\mathcal{D}_{be}$ with a threshold value $\epsilon\|\hat{x}\|_1$. Here, $\hat{x}$ is the input image with the maximum $L1$ norm in the benign dataset $\mathcal{D}_{be}$, and ϵ is a constant. The average perturbation is calculated as $\frac{1}{N_{be}}\sum_{x_i\in\mathcal{D}_{be}}\|x_i'-x_i\|_1$, where N_{be} represents the number of images in $\mathcal{D}_{be}$. To determine the threshold value, we assume that the perturbations of benign models follow a normal distribution. We compute the mean m_p and standard deviation σ_p of average perturbations across ten benign models reported in Figure 10. We set the threshold value to be $m_p - 2\sigma_p$, meaning that models with perturbation values below this threshold have a greater than 95% probability of being outliers, indicating a backdoored model. This corresponds to $\epsilon = 0.03$.

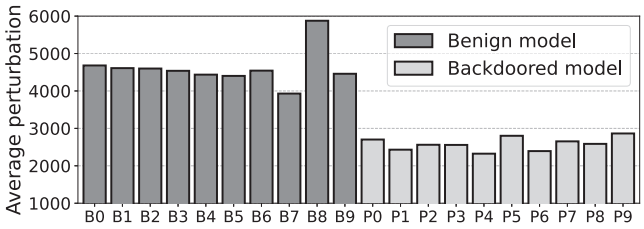

Figure 10: The average perturbations for ten benign models (B0~B9) and ten backdoored models (P0~P9).

4.3.4 Backdoor mitigation. Once a gaze estimation model $\mathcal{G}$ is identified as backdoored, SecureGaze fine-tunes $\mathcal{G}$ to mitigate backdoor behavior, such that the fine-tuned model produces correct gaze directions for poisoned images. **Note that, the defender only has access to a small benign dataset $\mathcal{D}_{be}$.** Therefore, SecureGaze generates a reverse-engineered poisoned dataset, $\mathcal{D}_{rp} = \{x_i', y_i\}_{i=1}^{N_{be}}$, by applying M_θ to each image x_i in $\mathcal{D}_{be}$ via Equation 5. Each reverse-engineered poisoned image x_i' in $\mathcal{D}_{rp}$ is annotated with its correct gaze annotation y_i. Next, SecureGaze fine-tunes $\mathcal{G}$ using both $\mathcal{D}_{be}$ and $\mathcal{D}_{rp}$. Formally, the backdoor mitigation is achieved by minimizing the following objective: $\sum_{(x_i,g_i)\in\mathcal{D}_{be}} \ell_1(\mathcal{G}(x_i), y_i) + \sum_{(x_i',g_i)\in\mathcal{D}_{rp}} \ell_1(\mathcal{G}(x_i'), y_i)$.

5 Evaluation

5.1 Evaluation Setups

5.1.1 Datasets. We consider two benchmark gaze estimation datasets that are collected in real-world settings.

- **MPIIFaceGaze** [83] is collected from 15 subjects during their routine laptop usage. Each subject contains 3,000 images under different backgrounds, illumination conditions, and head poses.
- **GazeCapture** [35] is a large-scale dataset collected from over 1450 individuals in real-world environments. It comprises 2.5 million images captured using the front-facing cameras of smartphones, showcasing a diverse range of lighting conditions and backgrounds.

For each dataset, we randomly sample 80% of the images to form the training dataset $\mathcal{D}_{tr}$ and 10% to form the benign dataset $\mathcal{D}_{be}$, ensuring that there is no overlap between them. $\mathcal{D}_{tr}$ is employed to train backdoored and benign models, while $\mathcal{D}_{be}$ is utilized for backdoor identification and mitigation. The remaining images constitute the testing set $\mathcal{D}_{te}$ to evaluate mitigation performance.

5.1.2 Backdoor attacks. We consider five SOTA attacks, including both input-independent and input-aware attacks.

- **BadNets** [18] generates poisoned inputs by pasting a fixed pattern as the backdoor trigger on the inputs. We use a 20×20 red patch located at the right-bottom corner as the backdoor trigger.
- **Clean Label** [65] exclusively applies a fixed pattern as the backdoor trigger to images belonging to the target class in classification. To adapt it for gaze estimation, we apply the trigger to images with gaze annotations "close" to y_T, specifically those where $|y - y_T| \leq \delta$.
- **WaNet** [53] generates stealth and input-aware backdoor triggers through image warping techniques. These triggers are injected into images using the elastic warping operation.

Figure 11: Poisoned images generated by IADA. The patterns and positions of triggers vary across different inputs.

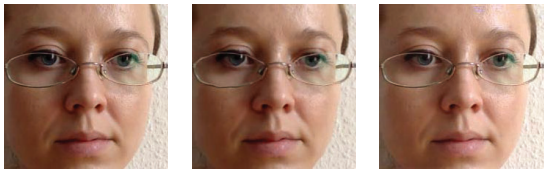

Figure 12: Comparsion between benign image and images poisoned by WaNet and IBA. The triggers are invisible.

- **Input-aware dynamic attack (IADA)** [54] generates dynamic backdoor triggers using a trainable trigger generator, which produces backdoor triggers varying from input to input.
- **Invisible backdoor attack (IBA)** [41] generates sample-specific invisible noises via steganography technique [63] as the backdoor triggers, which contain information of a predefined string.

5.1.3 Discussion on backdoor triggers. The backdoor triggers used in BadNets and Clean Label are input-independent, meaning their patterns and positions remain fixed across different inputs. In our setup, we consider a red patch in the bottom-right corner as the backdoor trigger. However, an attacker could use various patterns in different locations. As long as the pattern and position of the trigger remain consistent during both training and inference, the attacker can successfully execute a backdoor attack.

In contrast, WaNet, IADA, and IBA employ input-aware triggers, where the patterns and positions vary across different inputs. Figure 11 illustrates poisoned images generated by IADA, showcasing how the triggers change from one input to another. Additionally, WaNet and IBA create imperceptible backdoor triggers using image warping and DNN-based steganography techniques, respectively. To demonstrate the invisibility of these triggers, Figure 12 presents both benign and poisoned images produced by WaNet and IBA.

5.1.4 Attack time overhead. Below, we analyze the time overhead required to launch these attacks. For attacks in the digital world, the attacker needs to inject triggers into images. To quantify this, we measure the latency of trigger injection for each attack on a desktop equipped with an NVIDIA GeForce RTX 3080 Ti GPU and an Intel i7-12700KF CPU, with results presented in the Table below:

Attack	BadNets	Clean Label	IADA	WaNet	IBA
Latency	0.3 ms	0.3 ms	12.3 ms	4.5 ms	1.9 ms

As shown, BadNets requires an overhead of 0.3 ms for applying a fixed pattern directly to the image. In contrast, IADA incurs a significantly higher overhead of 12 ms due to the use of generative models for trigger injection. For attacks in the physical world (such as the one we demonstrated in Section 3.2.2), the time required to

place the physical trigger within the camera view is negligible. For both digital and physical world attacks, once the trigger appears in the camera view or image, the backdoored model exhibits its backdoored behavior with the same latency as a standard model inference, e.g., 12 ms for a model implemented using ResNet18 [22].

5.1.5 Compared defenses. We compare our method with the following defenses:

- **Gaze-NC** is adapted from NC [68]. Since NC is designed for classification and needs to enumerate all potential targets, we adapt it for gaze estimation by treating the potential target gaze direction as the optimization variable.
- **Gaze-FRE** is adapted from FRE [70], which utilizes the feature-space characteristics of backdoored classification models. Similar to Gaze-NC, we adapt it for gaze estimation by considering the potential target gaze direction as an optimization variable.
- **Fine-prune** [46] notes that the compromised neurons for backdoored classification models are dormant for benign inputs. Therefore, given a benign dataset, Fine-prune removes neurons with low activation values for benign images.
- **ANP** [74] observes that the compromised neurons are sensitive to perturbations. Based on this observation, ANP applies adversarial attacks to the neurons to identify sensitive neurons and subsequently prunes them for backdoor defense.
- **NAD** [42] first fine-tunes the given backdoored model on the benign dataset. It then treats the fine-tuned model as a teacher model and performs knowledge distillation to the original model.
- **RNP** [43] first maximizes errors on clean samples at the neuron level, then minimizes errors on the same samples at the filter level to identify compromised neurons for pruning.
- **Fine-tune** serves as a straightforward baseline that employs the benign dataset to directly fine-tune the backdoored models. We consider this baseline as existing research [46, 74] show its effectiveness on backdoor mitigation.

5.1.6 Evaluation metrics. Given a set of benign and backdoored models, we use the following metrics to evaluate the performance of SecureGaze on backdoor identification:

- **Identification Accuracy (Acc):** the percentage of correctly classified models (either benign or backdoored) over all the models.
- **True Positives (TP):** the number of correctly identified backdoored gaze estimation models.
- **False Positives (FP):** the number of benign gaze estimation models recognized as backdoored models.
- **False Negatives (FN):** the number of backdoored gaze estimation models identified as benign models.
- **True Negatives (TN):** the number of correctly recognized benign gaze estimation models.
- **ROC-AUC:** the ROC-AUC score computed from the average perturbations for benign and backdoored gaze estimation models. This metric is used to compare the backdoor identification performance between SecureGaze, Gaze-NC, and Gaze-FRE.

To evaluate the performance on backdoor mitigation, we generate a poisoned dataset $\mathcal{PD}_{te}$ by applying the trigger function to all the images in $\mathcal{D}_{te}$. Then, we use the following metrics:

Table 3: Backdoor identification performance on MPIIFaceGaze and GazeCapture for different attacks. SecureGaze can identify the backdoored gaze estimation models on two datasets with over 92% accuracy.

Attack	MPIIFaceGaze					GazeCapture				
	TP	FP	FN	TN	Acc	TP	FP	FN	TN	Acc
BadNets	20	3	0	17	92.5%	20	2	0	18	95.0%
IADA	20	3	0	17	92.5%	19	2	1	18	92.5%
Clean Label	20	3	0	17	92.5%	20	2	0	18	95.0%
WaNet	20	3	0	17	92.5%	20	2	0	18	95.0%
IBA	20	3	0	17	92.5%	20	2	0	18	95.0%

- **Average Attack Error (AE):** the average angular error between the estimated gaze directions and the target gaze directions over all the images in $\mathcal{PD}_{te}$.
- **Defending Attack error (DAE):** the average angular error between the estimated gaze directions and the correct gaze annotations over all the images in $\mathcal{PD}_{te}$.

A larger AE and a smaller DAE indicate better mitigation performance, while a smaller AE indicates better attack performance.

5.1.7 Implementation. We develop SecureGaze using TensorFlow and Adam optimizer [33]. We use a simple auto-encoder to implement M_θ, which is similar to that used in [54]. Before performing the reverse engineering, we pre-train M_θ on the benign dataset $\mathcal{D}_{be}$ for 5,000 steps with the learning rate of 0.001. We train M_θ for 2,000 steps with a batch size of 50 and the learning rate of 0.0015. We set $\lambda_1 = 20$, $\lambda_2 = 800$. For backdoor mitigation, we fine-tune the gaze estimation models using a batch of 50 benign and 50 reverse-engineered poisoned images for 300 iterations. We use ResNet18 [22] (without the dense layer) to implement F, and a dense layer without activation function to implement H.

5.2 Backdoor Identification Performance

We evaluate backdoor identification performance on 200 backdoored and 40 benign gaze estimation models. Specifically, for each dataset, we first train 20 benign models and then train 20 backdoored models for each attack. It is important to note that although the 20 backdoored (or benign) models for each attack-dataset combination are trained on the same training dataset, they have different parameters and exhibit variations in performance due to two key factors: 1) Each model is randomly initialized with different parameters; 2) During training, image batches are randomly sampled in each iteration, introducing variability in the training process. This is a standard evaluation protocol used in existing works [13, 70].

Evaluation results. We report the backdoor identification results of SecureGaze in Table 3, which indicate that SecureGaze can identify backdoored gaze estimation models trained by both input-independent and input-aware attacks, on MPIIFaceGaze and GazeCapture, with an average accuracy of 92.5% and 94.5%, respectively. Specifically, TP and FN remain consistent across most evaluation scenarios, with TP being 20 and FN being 0. This indicates that SecureGaze successfully identifies all 20 backdoored gaze estimation models without any false negatives. Moreover, TN and FP are identical for each backdoor attack. This consistency arises because the set of benign gaze estimation models, which are

Table 4: The ROC-AUC scores of different backdoor identification methods when evaluating on MPIIFaceGaze and GazeCapture with different attacks. SecureGaze outperforms both Gaze-NC and Gaze-FRE significantly.

Method	MPIIFaceGaze					GazeCapture				
	BadNets	IADA	Clean Label	WaNet	All	BadNets	IADA	Clean Label	WaNet	All
Gaze-NC	0.400	0.311	0.002	0.828	0.385	0.417	0.605	0.026	0.630	0.419
Gaze-FRE	0.561	0.512	0.444	0.508	0.506	0.528	0.531	0.461	0.601	0.530
SecureGaze	**0.995**	**1.000**	**1.000**	**0.995**	**0.998**	**0.995**	**1.000**	**1.000**	**0.967**	**0.986**

Table 5: Performance of SecureGaze in backdoor mitigation. SecureGaze shows larger AE and smaller DAE on two datasets, which demonstrates good performance in backdoor mitigation for various attacks.

Method	Metric	MPIIFaceGaze					GazeCapture				
		BadNets	IADA	Clean Label	WaNet	IBA	BadNets	IADA	Clean Label	WaNet	IBA
Undefended	AE	3.25	3.19	0.72	1.31	3.04	1.09	1.54	2.45	2.51	0.91
	DAE	14.8	14.4	15.4	15.9	14.4	20.0	10.6	9.85	9.55	19.4
SecureGaze	AE	17.2	15.6	16.4	15.3	14.2	17.7	10.2	10.9	9.57	19.2
	DAE	3.59	3.50	2.51	3.29	4.12	3.65	3.77	3.20	3.66	3.90

attack-free and independent of backdoor attack, remains the same across all five scenarios, and thus, SecureGaze leads to the same identification results, in FP and TN, regardless of the attacks.

Discussion on failure cases. For FN cases, SecureGaze struggles to identify a trigger function that produces similar gaze directions, instead prioritizing the minimization of perturbations. By contrast, for FP cases, SecureGaze reverse-engineers a trigger function that maps different inputs close to the target gaze direction but neglects the magnitude of perturbations introduced to benign images. We believe this issue arises from using fixed values for λ_1 and λ_2, where FN cases require larger values, while FP cases benefit from smaller values. A potential solution is to dynamically adjust λ_1 and λ_2. For instance, we can initially increase their values to ensure the trigger function generates similar outputs across different inputs, then gradually decrease them to focus on minimizing perturbations.

Comparison with state-of-the-art defenses. Table 4 shows the ROC-AUC scores of SecureGaze, Gaze-NC, and Gaze-FRE for different backdoor attacks on two datasets. We also report the scores when applying the various attacks simultaneously. As shown, the score of SecureGaze is above 0.96 in all the examined cases, which is significantly higher than that of Gaze-NC and Gaze-FRE. Besides, we notice that Gaze-FRE fails to find a trigger function that enables the backdoored gaze estimation model to map different inputs to similar gaze directions. This observation confirms our analysis that the feature-space characteristics for backdoored classification models [70] do not hold for backdoored gaze estimation models.

5.3 Backdoor Mitigation Performance

Evaluation results. We train backdoored gaze estimation models by each considered attack on each dataset. The backdoor mitigation results of SecureGaze are shown in Table 5, which indicate that SecureGaze can mitigate backdoor behaviors for various attacks on two dataset. Specifically, SecureGaze can substantially increase AE, indicating that the output gaze directions for poisoned inputs deviate significantly from the target gaze direction after backdoor mitigation. Additionally, SecureGaze significantly reduces DAE, which means that the mitigated backdoored gaze estimation models perform normally and output gaze directions close to correct gaze annotation, even though triggers are injected into the inputs.

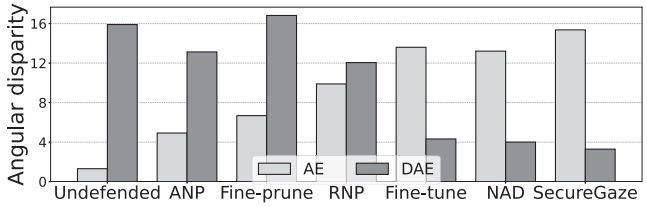

Figure 13: Performance of different defenses on backdoor mitigation. SecureGaze outperforms compared methods.

Comparison with state-of-the-art defenses. We compare SecureGaze with ANP, RNP, NAD, Fine-prune, and Fine-tune on backdoor mitigation. Specifically, we train a backdoored gaze estimation model by WaNet on MPIIFaceGaze and apply different methods to mitigate the backdoor behavior. The evaluation results are shown in Figure 13. The AE for SecureGaze is significantly larger than that for other methods, while the DAE for SecureGaze is much smaller than that for other methods, which shows the superiority of SecureGaze on backdoor mitigation. Moreover, Fine-prune, ANP, and RNP, which prune compromised neurons, perform poorly on backdoored gaze estimation models. This supports our analysis that the feature-space characteristics of backdoored gaze estimation models differ from those of backdoored classification models, making it ineffective to target specific neurons for backdoor mitigation.

5.4 System Profiling

We measure the latency and memory usage of SecureGaze during two key processes: reverse-engineering the trigger function for backdoor identification and fine-tuning the model for backdoor mitigation. These measurements are conducted on a desktop equipped with an NVIDIA GeForce RTX 3080 Ti GPU and an Intel i7-12700KF CPU, following the implementation details outlined in Section 5.1.7.

Latancy. For reverse-engineering the trigger function, we repeat the process five times for a given gaze estimation model. The average latency for reverse engineering is 12 minutes. For backdoor mitigation, we repeat the experiments five times with a given gaze

Table 6: The impact of FSO, λ_1, λ_2, and p on backdoor identification performance.

Metric	Different λ_1			Different λ_2			Different p			w/o FSO
	10	20	30	600	800	1000	5%	10%	15%	
TP	20	20	19	20	20	20	20	20	20	20
FP	3	3	3	11	3	3	4	3	3	20
FN	0	0	1	0	0	0	0	0	0	0
TN	17	17	17	9	17	17	16	17	17	0
Acc	92.5%	92.5%	90%	72.5%	92.5%	92.5%	90%	92.5%	92.5%	50%

Table 7: Results on adaptive attack with different values for β. Adaptive WaNet is less effective than WaNet.

Metric	WaNet	Adaptive WaNet	
		β =0.02	β =1.0
AE	1.50	5.41	10.8
DAE	16.0	14.9	12.8
Acc	92.5%	92.5%	67.5%

estimation model and a reverse-engineered trigger function. The average latency for backdoor mitigation is 100 seconds.

Memory usage. We measure the memory specifically allocated to the training process, i.e., training M_θ or fine-tuning the gaze estimation model. This is determined by subtracting the memory usage before training from the peak memory usage during training. Specifically, reverse-engineering the trigger function requires approximately 9,970 MB of memory, while fine-tuning the gaze estimation model consumes around 6,000 MB.

Note that, for a given gaze estimation model, the process of backdoor identification and mitigation needs to be performed offline only once before deployment. As a result, SecureGaze does not introduce additional run-time latency to the gaze estimation model after deployment. Furthermore, since SecureGaze does not need to enumerate all potential targets, it is more efficient than existing reverse-engineering-based techniques that require scanning all labels (e.g., 140 minutes for FRE [70] on ImageNet [10]).

5.5 Ablation Studies

We conduct ablation studies to investigate the impact of different design choices on the performance of SecureGaze. We consider WaNet as the attack and evaluate on MPIIFaceGaze.

5.5.1 Impact of weights and the size of benign dataset. We vary the values of λ_1 and λ_2 in Equation 4 from 10 to 30 and from 600 to 800 respectively to investigate their impacts on the performance of backdoor identification. Moreover, we study the impact of the size of $\mathcal{D}_{be}$ on the identification performance by changing the ratio p of the benign dataset to the original whole dataset from 5% to 15%. We report the backdoor identification results in Table 6. We observe that the performance of SecureGaze is insensitive to λ_1, as the identification accuracy is almost stable with different λ_1. However, SecureGaze is sensitive to λ_2 and the identification accuracy increases with λ_2, as a larger λ_2 allows the feature-space optimization objective to have a greater contribution to the optimization problem. This observation proves that the proposed feature-space optimization objective is important for backdoor identification. Additionally, as p decreases from 10% to 5%, the identification accuracy and TN decrease, while TP remains stable. This is because, compared to a larger p, it is easier to find a small amount of perturbation that can lead to the backdoor behavior on a smaller p for benign models. However, the identification accuracy is still 90% even when $p = 5\%$.

5.5.2 Impact of feature-space optimization objective (FSO). We study the impact of FSO on the performance of backdoor identification by removing it from $\mathcal{OPT}$-SecureGaze. The results are shown in Table 6, which indicates that all the backdoored and benign models are classified as backdoored models. This means that SecureGaze cannot identify backdoor without the FSO. We further observe that without the FSO, SecureGaze cannot find a trigger function that can map different inputs to similar output vectors. As a result, SecureGaze solves the optimization problem by focusing on minimizing the amount of perturbations, which leads to the misclassification of backdoored models.

5.5.3 Generalization to various datasets. We investigate the generalization capability of SecureGaze across different regression tasks by utilizing three additional datasets. These include XGaze [79], a complex gaze estimation dataset with a wider range of head poses and gaze directions, and two datasets focused on head pose estimation, i.e., Biwi [12] and Pandora [3]. Specifically, head pose estimation seeks to determine a three-dimensional vector representing the Eular angle (yaw, pitch, roll) from a monocular image. This modality is commonly used in human-computer interaction to infer user attention [9, 23, 69] and for authentication system [45]. Biwi is collected from 24 subjects, and each subject has 400 to 900 images. We use the *cropped faces of Biwi dataset (RGB images)* released by [3]. Pandora has 100 subjects and more than 120,000 images.

We train 20 backdoored and 20 benign models on the training dataset for each dataset. For head pose estimation, we set $\lambda_1 = 10$, $\lambda_2 = 100$, and $\epsilon = 0.05$, using average L_1 error to define AE and DAE, rather than average angular error. The evaluation results for backdoor identification and mitigation are shown in Table 8, which demonstrates that SecureGaze is effective across various datasets and regression tasks in human-computer interaction.

5.6 Adaptive Attack

When the attacker has the full knowledge of SecureGaze, one potential adaptive attack that can bypass our method is to force RAV to be close to 1 to break the feature-space characteristics. Based on this intuition, we design an adaptive attack that adds an additional loss term L_{adp} with a weight β to the original loss function of the chosen attack to enforce RAV to be close to one. We define L_{adp} as:

$$L_{adp} = \left|1 - \frac{1}{d}\sum_{j=1}^{d} \frac{\sigma^2\left(\left\{\mathcal{B}(F(\mathcal{A}(x_i)), w_j)\right\}_{i=1}^{N_p}\right)}{\sigma^2\left(\left\{\mathcal{B}(F((x_i), w_j))\right\}_{i=1}^{N_b}\right)}\right|, \tag{7}$$

where N_p and N_b are the numbers of poisoned and benign inputs in a minibatch. We consider two values for β, i.e., 0.02 and 1.0. We train 20 backdoored models by incorporating L_{adp} into WaNet for each considered value of β.

Table 7 shows the identification accuracy and the averaged AE and DAE over 20 backdoored models. As shown, the AE of the adaptive attack is much higher than that of WaNet and it increases as β raises. This proves that the feature-space characteristics we

Table 8: Backdoor identification and mitigation performance of SecureGaze on various datasets. SecureGaze is effective across various datasets and regression tasks.

	Backdoor identification					Backdoor mitigation			
Dataset	TP	FP	FN	TN	Acc	Undefended AE	Undefended DAE	SecureGaze AE	SecureGaze DAE
XGaze	20	0	0	20	100%	1.24	44.5	45.4	2.97
Biwi	20	0	0	20	100%	2.48	25.2	27.1	1.10
Pandora	20	0	0	20	100%	0.48	22.8	23.0	2.71

Table 9: The average attack error (in degree) for subjects with physical triggers before and after backdoor mitigation.

Model	Subject 1	Subject 2	Subject 3	Subject 4
Before mitigation	1.71	1.07	0.98	1.17
After mitigation	15.9	16.6	10.3	7.6

observed of backdoored gaze estimation models are vital to result in the backdoor behavior. Moreover, the adaptive attack with $\beta = 0.02$ cannot reduce the identification accuracy as the feature-space characteristics are not totally broken. When increasing β to 1.0, the identification accuracy drops to 67.5%. However, the AE is higher than 10.0 with $\beta = 1.0$, in which we believe the attack is ineffective.

5.7 Physical-world Backdoor Defense

Below, we apply SecureGaze to mitigate the backdoor behavior of the backdoored gaze estimation model we considered in Section 3.2.2, in which a physical item, i.e., a piece of white tape on the face, can effectively trigger the backdoored behaviors. We record the estimated gaze from the mitigated model for each subject by the gaze estimation pipeline in Figure 4 when the physical trigger is present on the face. Figure 14 visualizes the estimated gaze directions, while Table 9 quantifies the average attack error, comparing results before and after mitigation. Specifically, before mitigation, the estimated gaze directions (green dots) are concentrated around the attacker-chosen target (presented by the red star), exhibiting a small average attack error. By contrast, after backdoor mitigation using SecureGaze, the gaze estimations (blue dots) form four clusters corresponding to the four corners where the stimulus appeared, resulting in a significantly higher average attack error than before migitation. Moreover, we also plot the gaze directions estimated by the backdoored model from subjects without triggers (yellow dots) in Figure 14, which overlap with the estimations for subjects with physical triggers after mitigation (blue dots), indicating that SecureGaze can effectively mitigate the backdoor behavior. A video demo that compares behaviors of the backdoored and backdoor-mitigated models can be found in our GitHub repository: https://github.com/LingyuDu/SecureGaze.

5.8 Limitations and Future Works

Limitation. Similar to existing reverse-engineering-based methods [13, 70], the current design of SecureGaze adopts a fixed threshold for backdoor identification, which may be less effective if the attacker employs a large trigger. Moreover, while we have investigated the impact of different hyperparameter values on the performance of SecureGaze, our analysis is limited to a few scenarios rather than an exhaustive search across a broader range of cases.

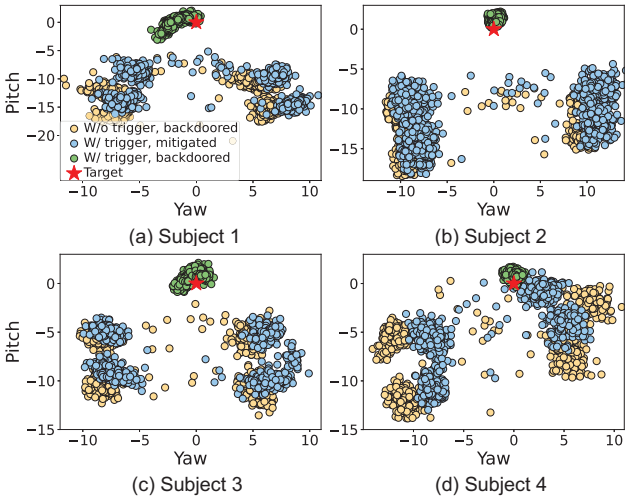

Figure 14: Gaze directions estimated by the backdoored gaze estimation model before (green dots) and after (blue dots) backdoor mitigation using SecureGaze.

Future research directions. A promising avenue for future research involves extending SecureGaze to a wider range of regression models with continuous output spaces that are adopted in human-computer interactions. Another interesting direction is to generalize SecureGaze to more complex threat scenarios, e.g., the gaze estimation models are backdoored by multiple trigger functions associated with multiple target gaze directions. Additionally, exploring more adaptive attacks could provide deeper insights into the robustness and limitations of SecureGaze, enabling a more comprehensive investigation into backdoor defenses tailored specifically for gaze estimation models.

6 Conclusion

In this paper, we present SecureGaze, the first approach to defend gaze estimation models against backdoor attacks. We identify the unique characteristics of backdoored gaze estimation models, based on which we introduce a novel suite of techniques to reverse engineer the trigger function for backdoored gaze estimation models without the need to enumerate all the outputs. Our comprehensive experiments in both digital and physical worlds show that SecureGaze is consistently effective in defending gaze estimation models against six backdoor attacks that are triggered by input-aware patterns, input-independent patterns, and physical objects. We also adapt seven state-of-the-art classification defenses, showing that they are ineffective for gaze estimation, while SecureGaze consistently outperforms them.

Acknowledgments

We express our gratitude to the shepherd and the anonymous reviewers for their insightful comments. This work was supported in part by a Meta Research Award and by SURF Research Cloud grants EINF-2391, EINF-8964, and EINF-9272.

References

[1] Yomna Abdelrahman, Anam Ahmad Khan, Joshua Newn, Eduardo Velloso, Sherine Ashraf Safwat, James Bailey, Andreas Bulling, Frank Vetere, and Albrecht Schmidt. 2019. Classifying Attention Types with Thermal Imaging and Eye Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2019).

[2] Bradley Berman. 2020. Driver-monitoring systems to be as common as seat belts. https://www.sae.org/news/2020/02/smart-eye-safety-driver-monitoring.

[3] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. 2017. POSEidon: Face-From-Depth for Driver Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Andreas Bulling and Michel Wedel. 2019. Pervasive eye-tracking for real-world consumer behavior analysis. In *A handbook of process tracing methods*. Routledge, 27–44.

[5] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728* (2018).

[6] Weixin Chen, Baoyuan Wu, and Haoqian Wang. 2022. Effective backdoor defense by exploiting sensitivity of poisoned samples. In *Advances in Neural Information Processing Systems*.

[7] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* (2017).

[8] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2024. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[9] Andrew Crossan, Mark McGill, Stephen Brewster, and Roderick Murray-Smith. 2009. Head tilting for interaction in mobile contexts. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. 1–10.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[11] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. 2020. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*.

[12] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. 2013. Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision* 101, 3 (February 2013), 437–458.

[13] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. 2023. Detecting Backdoors in Pre-Trained Encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. 2018. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *Proceedings of European Conference on Computer Vision (ECCV)*.

[15] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T Freeman. 2018. Cognitive load estimation in the wild. In *Proceedings of the Chi Conference on Human Factors in Computing Systems (CHI)*.

[16] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. 2019. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*.

[17] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. 2022. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2022), 1563–1580.

[18] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access* 7 (2019), 47230–47244.

[19] Jiyang Guan, Zhuozhuo Tu, Ran He, and Dacheng Tao. 2022. Few-Shot Backdoor Defense Using Shapley Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Elias Daniel Guestrin and Moshe Eizenman. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering* 53, 6 (2006), 1124–1133.

[21] Dan Witzner Hansen and Qiang Ji. 2009. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 32, 3 (2009), 478–500.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Sebastian Hueber, Christian Cherek, Philipp Wacker, Jan Borchers, and Simon Voelker. 2020. Headbang: Using head gestures to trigger discrete actions on mobile devices. In *Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*. 1–10.

[24] Sinh Huynh, Rajesh Krishna Balan, and JeongGil Ko. 2021. iMon: Appearance-based gaze tracking system on mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2021).

[25] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2022. Dwell Selection with ML-based Intent Prediction Using Only Gaze Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2022).

[26] Saurabh Jaiswal, Shubham Virmani, Vishal Sethi, Kanjar De, and Partha Pratim Roy. 2019. An intelligent recommendation system using gaze and emotion detection. *Multimedia Tools and Applications* 78 (2019), 14231–14250.

[27] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct publication (Ubicomp)*.

[28] Christina Katsini, Yasmeen Abdrabou, George E Raptis, Mohamed Khamis, and Florian Alt. 2020. The role of eye gaze in security and privacy applications: Survey and future HCI research directions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*.

[29] Leo Kelion. 2013. Caterpillar backs eye-tracker to combat driver fatigue. https://www.bbc.com/news/technology-22640279.

[30] Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The past, present, and future of gaze-enabled handheld mobile devices: Survey and lessons learned. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*.

[31] Mohamed Khamis, Daniel Buschek, Tobias Thieron, Florian Alt, and Andreas Bulling. 2018. EyePACT: Eye-Based Parallax Correction on Touch-Enabled Interactive Displays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2018).

[32] Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. NVGaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*.

[33] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[34] Stefanos Koffas, Stjepan Picek, and Mauro Conti. 2022. Dynamic Backdoors with Global Average Pooling. In *Proceedings of IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*.

[35] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. 2020. What does your gaze reveal about you? On the privacy implications of eye tracking. In *IFIP International Summer School on Privacy and Identity Management*. Springer.

[37] Christian Lander, Markus löchtefeld, and Antonio Krüger. 2018. hEYEbrid: A hybrid approach for mobile calibration-free gaze estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2018).

[38] Yaxiong Lei, Shijing He, Mohamed Khamis, and Juan Ye. 2023. An end-to-end review of gaze estimation and its interactive applications on handheld mobile devices. *Comput. Surveys* 56, 2 (2023), 1–38.

[39] Yinghui Li, Zhichao Cao, and Jiliang Wang. 2017. Gazture: Design and Implementation of a Gaze based Gesture Control System on Tablets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2017).

[40] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2022. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 35, 1 (2022), 5–22.

[41] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. 2021. Invisible Backdoor Attack With Sample-Specific Triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 16463–16472.

[42] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2021. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930* (2021).

[43] Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. 2023. Reconstructive Neuron Pruning for Backdoor Defense. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research)*.

[44] Yixuan Li, Pingmei Xu, Dmitry Lagun, and Vidhya Navalpakkam. 2017. Towards measuring and inferring user interest from gaze. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 525–533.

[45] Chen Liang, Chun Yu, Xiaoying Wei, Xuhai Xu, Yongquan Hu, Yuntao Wang, and Yuanchun Shi. 2021. Auth+ track: Enabling authentication free interaction on smartphone by continuous user tracking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)*. 1–16.

[46] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*.

[47] Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. 2022. Backdoor Defense with Machine Unlearning. In *Proceedings of the IEEE*

International Conference on Computer Communications (INFOCOM).

[48] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. 2019. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS).*

[49] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2018. Trojaning Attack on Neural Networks. In *Proceedings of Network and Distributed System Security Symposium (NDSS).*

[50] Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. 2023. The "Beatrix" Resurrections: Robust Backdoor Detection via Gram Matrices. In *Proceedings of Network and Distributed System Security Symposium (NDSS).*

[51] Sebastian Marwecki, Andrew D Wilson, Eyal Ofek, Mar Gonzalez Franco, and Christian Holz. 2019. Mise-unseen: Using eye tracking to hide virtual reality scene changes in plain sight. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST).*

[52] Atsushi Nakazawa and Christian Nitschke. 2012. Point of gaze estimation through corneal surface reflection in an active illumination environment. In *Proceedings of the European Conference on Computer Vision (ECCV).*

[53] Anh Nguyen and Anh Tran. 2021. Wanet–imperceptible warping-based backdoor attack. In *Proceedings of International Conference on Learning Representations (ICLR).*

[54] Tuan Anh Nguyen and Anh Tran. 2020. Input-aware dynamic backdoor attack. In *Proceedings of Conference on Advances in Neural Information Processing Systems (NeurIPS).*

[55] Jacob L Orquin, Erik S Lahm, and Hrvoje Stojić. 2021. The visual environment and attention in decision making. *Psychological Bulletin* 147, 6 (2021), 597.

[56] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. 2018. Predicting the Driver's Focus of Attention: the DR(eye)VE Project. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 7 (2018), 1720–1733.

[57] Ximing Qiao, Yukun Yang, and Hai Li. 2019. Defending neural backdoors via generative distribution modeling. In *Proceedings of the Conference on Advances in neural information processing systems (NeurIPS).*

[58] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. 2022. Dynamic Backdoor Attacks Against Machine Learning Models. In *Proceedings of IEEE European Symposium on Security and Privacy (EuroS&P).*

[59] Lujia Shen, Shouling Ji, Xuhong Zhang, Jinfeng Li, Jing Chen, Jie Shi, Chengfang Fang, Jianwei Yin, and Ting Wang. 2021. Backdoor Pre-trained Models Can Transfer to All. In *Proceedings of ACM Annual Conference on Computer and Communication Security (CCS).* Association for Computing Machinery, 3141–3158.

[60] NBC Tech News Daily Staff. 2010. Eye Tracker Wakes Sleepy Drivers. https://www.nbcnews.com/id/wbna39668980.

[61] Yusuke Sugano, Xucong Zhang, and Andreas Bulling. 2016. Aggregaze: Collective estimation of audience attention on public displays. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST).*

[62] Kar-Han Tan, David J Kriegman, and Narendra Ahuja. 2002. Appearance-based eye gaze estimation. In *Proceedings of Sixth IEEE Workshop on Applications of Computer Vision (WACV).*

[63] Matthew Tancik, Ben Mildenhall, and Ren Ng. 2020. StegaStamp: Invisible Hyperlinks in Physical Photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[64] Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2017).

[65] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2019. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771* (2019).

[66] Peter Valdes-Dapena. 2023. In a new BMW sedan, drivers can change lanes using just their eyes. https://edition.cnn.com/2023/05/24/business/bmw-eye-lane-change/index.html.

[67] Francisco Vicente, Zehua Huang, Xuehan Xiong, Fernando De la Torre, Wende Zhang, and Dan Levi. 2015. Driver gaze tracking and eyes off the road detection system. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)* 16, 4 (2015), 2014–2027.

[68] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P).*

[69] Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. Faceori: Tracking head position and orientation using ultrasonic ranging on earphones. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI).* 1–12.

[70] Zhenting Wang, Kai Mei, Hailun Ding, Juan Zhai, and Shiqing Ma. 2022. Rethinking the Reverse-engineering of Trojan Triggers. In *Proceedings of Conference on Advances in Neural Information Processing Systems (NeurIPS).*

[71] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. 2021. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[72] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. 2021. Backdoor Attacks Against Deep Learning Systems in the Physical World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[73] Justin C. Wilson, Suku Nair, Sandro Scielzo, and Eric C. Larson. 2021. Objective Measures of Cognitive Load Using Deep Multi-Modal Learning: A Use-Case in Aviation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2021).

[74] Dongxian Wu and Yisen Wang. 2021. Adversarial Neuron Pruning Purifies Backdoored Deep Models. In *Proceedings of the Conference on Advances in neural information processing systems (NeurIPS).*

[75] Zhen Xiang, David J. Miller, and George Kesidis. 2022. Detection of Backdoors in Trained Classifiers Without Access to the Training Set. *IEEE Transactions on Neural Networks and Learning Systems* 33, 3 (2022), 1177–1191.

[76] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. 2024. Towards Reliable and Efficient Backdoor Trigger Inversion via Decoupling Benign Features. In *Proceedings of the International Conference on Learning Representations (ICLR).*

[77] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2019. Latent backdoor attacks on deep neural networks. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS).*

[78] Yi Zeng, Si Chen, Won Park, Z. Morley Mao, Ming Jin, and Ruoxi Jia. 2022. Adversarial Unlearning of Backdoors via Implicit Hypergradient. In *Proceedings of the International Conference on Learning Representations (ICLR).*

[79] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In *Proceedings of the European Conference on Computer Vision (ECCV).*

[80] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2019. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI).*

[81] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[82] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*

[83] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2019).

[84] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. 2022. Data-free Backdoor Removal based on Channel Lipschitzness. In *Proceedings of European Conference on Computer Vision (ECCV).*

[85] Zhiwei Zhu and Qiang Ji. 2007. Novel eye gaze tracking techniques under natural head movement. *IEEE Transactions on Biomedical Engineering* 54, 12 (2007), 2246–2260.