PrivatEyes: Appearance-based Gaze Estimation Using Federated Secure Multi-Party Computation

MAYAR ELFARES[*†], PASCAL REISERT[†], ZHIMING HU[*‡], WENWU TANG, RALF KÜSTERS[†], and ANDREAS BULLING[*], [*] Institute for Visualisation and Interactive Systems, [†] Institute of Information Security, [‡] Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart, Germany

Latest gaze estimation methods require large-scale training data but their collection and exchange pose significant privacy risks. We propose *PrivatEyes* – the first privacy-enhancing training approach for appearance-based gaze estimation based on federated learning (FL) and secure multi-party computation (MPC). PrivatEyes enables training gaze estimators on multiple local datasets across different users and server-based secure aggregation of the individual estimators' updates. PrivatEyes guarantees that individual gaze data remains private even if a majority of the aggregating servers is malicious. We also introduce a new data leakage attack *DualView* that shows that PrivatEyes limits the leakage of private training data more effectively than previous approaches. Evaluations on the MPIIGaze, MPIIFaceGaze, GazeCapture, and NVGaze datasets further show that the improved privacy does not lead to a lower gaze estimation accuracy or substantially higher computational costs – both of which are on par with its non-secure counterparts.

CCS Concepts: • **Gaze estimation, privacy, adversarial attack, federated learning, secure multi-party computation**;

ACM Reference Format:
Mayar Elfares, Pascal Reisert, Zhiming Hu, Wenwu Tang, Ralf Küsters, and Andreas Bulling. 2024. PrivatEyes: Appearance-based Gaze Estimation Using Federated Secure Multi-Party Computation. *Proc. ACM Hum.-Comput. Interact.* 8, ETRA, Article 232 (May 2024), 23 pages. https://doi.org/10.1145/3655606

1 INTRODUCTION

Starting with pioneering work by Zhang et al. [Zhang et al. 2015, 2017c], research on appearance-based gaze estimation using deep learning has spurred an increasing number of papers in recent years. Much of the improvements in terms of gaze estimation accuracy can be attributed to the availability of ever-larger training datasets [Smith et al. 2013; Sugano et al. 2014; Zhang et al. 2020b, 2019]. Large eye image training data are required to capture the significant variability in eye appearances across users, tasks, and settings. Performance could be improved significantly using data collected in the wild, e.g., on portable devices used during everyday activities [Bâce et al. 2020; Krafka et al. 2016; Zhang et al. 2019]. As such, it is likely that continual learning approaches will also be used in the future to collect large-scale data in the background and train personalised gaze estimators across multiple devices [Zhang et al. 2018].

Authors' address: Mayar Elfares, mayar.elfares@vis.uni-stuttgart.de; Pascal Reisert, pascal.reisert@sec.uni-stuttgart.de; Zhiming Hu, zhiming.hu@vis.uni-stuttgart.de; Wenwu Tang, ; Ralf Küsters, ralf.kuesters@sec.uni-stuttgart.de; Andreas Bulling, andreas.bulling@vis.uni-stuttgart.de, [*] Institute for Visualisation and Interactive Systems, [†] Institute of Information Security, [‡] Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart, Germany.

2573-0142/2024/5-ART232 $15.00
https://doi.org/10.1145/3655606

Large-scale collection and transfer of gaze data or gaze estimation models over networks, however, pose significant privacy risks, such as information leakage or misuse of gaze data copies. With eye tracking becoming pervasive [Bulling and Gellersen 2010; Tonsen et al. 2017] and integrated into an ever-increasing number of personal devices [Bâce et al. 2020; Huang et al. 2016a, 2017], these privacy risks increase further. This is particularly critical given that gaze data contains rich personal information, such as gender [Sammaknejad et al. 2017], identity [Cantoni et al. 2015], personality traits [Hoppe et al. 2018], user activities [Bulling et al. 2013; Steil and Bulling 2015; Zhang et al. 2017b], attentive [Faber et al. 2018; Vertegaal et al. 2003] and cognitive states [Bulling and Roggen 2011; Huang et al. 2016b], or mental disorders [Holzman et al. 1974; Hutton et al. 1984]. Despite these risks, privacy has so far largely been neglected by the gaze estimation community. One notable exception is [Elfares et al. 2022] in which the authors have proposed to increase the privacy of gaze estimators using *federated learning* (FL) [McMahan et al. 2017]. Their method allows training gaze estimators across a large number of clients without directly revealing their private data, while adapting to the heterogeneous gaze data distributions (e.g. gaze range, head pose, illumination condition, and personal appearance). However, it still requires each client to reveal her individually trained models (IU) to an aggregating server and therewith remains susceptible to a large number of attacks, leaking information about the gaze data inputs as we show in this paper.

To address these limitations, we propose *PrivatEyes* – a novel training approach for appearance-based gaze estimation that combines FL with *secure multi-party computation* (MPC) (see Figure 1). In contrast to [Elfares et al. 2022], PrivatEyes uses n instead of only one aggregating server. Our MPC approach then allows a gaze estimator to be jointly trained by the clients and the n servers using shared secret representations of eye and face image data while keeping the data itself private to each party. By using MPC, it is guaranteed that no server learns the individual inputs or the individual model updates IU of the clients even if all-but-one server are malicious. We show that our combination of FL and MPC nevertheless only comes with a small computational overhead (see Section 5 for the exact numbers). In addition, through empirical evaluation on the MPIIGaze [Zhang et al. 2015], MPIIFaceGaze [Zhang et al. 2017c], GazeCapture [Krafka et al. 2016], and NVGaze [Kim et al. 2019] datasets, we show that PrivatEyes maintains an on-par gaze estimation performance as the non-secure state-of-the-art [Elfares et al. 2022], is domain-agnostic (i.e. can be used with any deep learning-based gaze estimation model), and can scale to $\sim 1.5k$ clients (i.e. the size of the largest evaluated data set GazeCapture). We note that PrivatEyes works with *any* (even larger) number of clients. We demonstrate the privacy advantages of our method against well-established data leakage attacks and our new DualView attack, which is able to simultaneously attack users' appearance (View1: how the user looks like) as well as their gaze distribution (View2: where the user is looking). In summary, our work makes the following contributions:

- We introduce PrivatEyes – the first privacy-preserving training approach for appearance-based gaze estimation that combines federated learning and secure multi-party computation and guarantees that data collectors (servers) do not learn individual inputs by clients, even if all-but-one data collectors are malicious.
- We further propose DualView – a novel data leakage attack to empirically demonstrate and measure the potential privacy risks associated with gaze estimation models.
- We implemented both PrivatEyes and the attack DualView. Our evaluation on several gaze estimation benchmark datasets shows that PrivatEyes reaches the same model performance and scalability as non-secure alternatives like [Elfares et al. 2022] with negligible computational overhead.
- We compare PrivatEyes to data centre training and federated learning [Elfares et al. 2022] w.r.t. their privacy leakage when attacked by DualView and other well-established data leakage attacks.

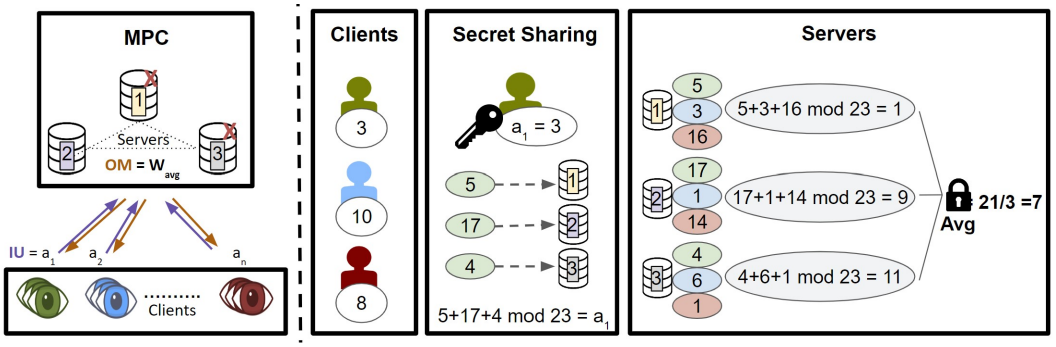

Fig. 1. PrivatEyes combines federated learning (FL) and secure multi-party computation (MPC) for privacy-enhancing training of appearance-based gaze estimation methods (see Sections 2 and 4.1). Clients C_j locally train the gaze estimation model on their private data. Each client C_j splits her individually trained model parameters (IU) a_j into n secret shares, i.e. in this example $n = 3$. Each server $S_i, 1 \leq i, \leq n$ receives its respective share of each a_j, e.g. S_1 receives the share 5 of $a_1 = 3$, 3 of $a_2 = 10$ and 16 of $a_3 = 8$. Then each server aggregates the different shares, e.g. S_1 computes $5 + 3 + 16 \bmod 23 = 1$, and sends the result (i.e. output model OM) to *all* clients. The clients compute the average $(1 + 9 + 11)/3 = 7$, but nothing more.

We show that PrivatEyes provides significantly better privacy guarantees, e.g. for the state-of-the-art centralised FL scheme [Elfares et al. 2022] DualView reconstructs 15/15 participants from MPIIGaze/MPIIFaceGaze accurately but $0/15$ for PrivatEyes.

2 BACKGROUND

Federated learning. Federated learning [McMahan et al. 2017] is a machine learning (ML) approach where multiple clients (e.g. gaze data owners) collaborate in solving an ML problem by training an ML model jointly (e.g. gaze estimator) on their local training dataset without sharing their data. An aggregation server broadcasts an initial ML model to all clients. Clients train the model on their local data samples without transferring the raw data, instead, *individual updates* IU (i.e. the ML model parameters) are sent to the server. The server aggregates the individual updates to a *output model* OM which is returned to all clients for the next training round. This process is repeated until the model converges or a certain number of rounds is reached. FL training can be performed in a centralised (client-server) or decentralised (peer-to-peer) fashion and can be classified as cross-device or cross-silo depending on whether clients are mobile devices or organisations (e.g., medical, financial, or geo-distributed data centres) that train on siloed data, respectively. FL can be further categorised into horizontal or vertical settings according to how data is partitioned among the clients in the feature and sample spaces. In horizontal FL, clients share overlapping data features that differ in data samples while the opposite is true in vertical FL [Kairouz et al. 2019].

Secure multi-party computation. In the basic secure multi-party computation (MPC) setting, a set of parties $P_1, \ldots, P_n$ is considered where (i) each party has some (private) input, i.e., input that the party does not want to reveal to other parties, including external observers, and ii) the parties would like to compute a previously agreed-upon function over their private inputs without revealing any information apart from the actual output of this function. For example, the parties might want to compute their average age without revealing their respective individual age to each other. Intuitively, this can be achieved by using a completely trusted third-party (a notary): The parties could give their private input to the notary, and the notary (correctly) computes the agreed-upon function and reveals the outcome to the parties, but never leaks the individual inputs of the parties to other parties or

external observers. An MPC protocol achieves exactly the same result without actually using a notary. Instead, the MPC protocols allow the parties to exchange (a series of) specially crafted messages that allow to compute the result without ever revealing the inputs.

More formally, MPC protocols should ensure (in a mathematically rigorous way) *input privacy*, i.e. that no information on the inputs (e.g. the individual age) is leaked apart from what can be deduced from the output (e.g. the average age). Moreover, we require *correctness*, i.e. every output (e.g. the average age) is correct (or the protocol aborts). Furthermore, we require that both input privacy and correctness still hold if all but one of the computing parties act *maliciously*, e.g. by colluding with each other or by deviating from the protocol. This means that a single honest MPC party still guarantees that inputs remain private and that the result is correct even if all other parties act maliciously. Note that even if the malicious parties force a protocol abort, e.g. by simple denial of service, the privacy of inputs is guaranteed. MPC protocols that ensure the aforementioned security properties are called *actively secure* against a majority of up to $n - 1$ malicious parties. The currently best known MPC protocols that provide this security standard are SPDZ [Damgård et al. 2012] and improvements thereof like Overdrive LowGear [Keller et al. 2018].

We employ these state-of-the-art protocols in a client/server MPC model. That is, we distinguish between input parties (clients), who provide the inputs on which the function is evaluated, and compute parties (servers), who carry out the MPC protocol and provide the clients with the output of the function evaluation. Then as long as one server remains honest, the inputs of all clients remain private and the result is correct (or the protocol aborts). The client/server model is particularity well-suited for a setup with (possibly resource-limited) clients and a few (more powerful) servers as it is common in our gaze applications [Elfares et al. 2022; Kairouz et al. 2019; Zhang et al. 2015].

3 RELATED WORK

Gaze estimation. Gaze estimation methods can generally be categorised as model-based or appearance-based [Hansen and Ji 2009]. While model-based methods estimate gaze direction from infrared corneal reflections [Hennessey et al. 2006; Morimoto et al. 2002], or geometric eye shape [Chen and Ji 2008; Valenti et al. 2011]), appearance-based methods directly regress a 2D or 3D gaze direction from eye images recorded using an off-the-shelf RGB camera [Baluja and Pomerleau 1993; Choi et al. 2013; Liang et al. 2013]. Appearance-based methods, particularly those based on deep learning, were shown to be more robust to varying lighting conditions, gaze ranges, and low image resolutions than their model-based counterparts [Biswas et al. 2021; Zhang et al. 2015, 2017c]. However, while gaze estimators can also be trained from synthesised images [Sugano et al. 2014; Wood et al. 2016a,b], effective training typically requires a large number of real-world training images [Krafka et al. 2016; Zhang et al. 2020b, 2019]. Despite an increasing number of works and significant advances in appearance-based gaze estimation methods in recent years, the privacy threats posed by the requirement for large-scale image datasets collected of individuals remain under-explored in the community. We address these privacy threats by presenting the first privacy-preserving gaze estimation scheme PrivatEyes in Section 4.

Privacy-preserving machine learning. Initially, privacy-preserving machine learning either implied low efficiency [Burkhalter et al. 2021; Chowdhury et al. 2021; Gilad-Bachrach et al. 2016], weak privacy guarantees [Liu et al. 2017; Mohassel and Rindal 2018; Mohassel and Zhang 2017; Rouhani et al. 2018], or was restricted to specific setups [Galli et al. 2022; Liu et al. 2017; Mohassel and Zhang 2017]. While the efficiency of distributed ML training has improved with FL, privacy issues endured due to the lack of provable privacy guarantees [Liu et al. 2022b]. First secure aggregation protocols based on MPC [Bonawitz et al. 2017] address these issues but either come with a low efficiency [Bonawitz et al. 2017; Chowdhury et al. 2021; Dong et al. 2021] or restrict to a weak security

setup, i.e. assume that all servers always follow the MPC protocol [Dong et al. 2021; Fereidooni et al. 2021; Nguyen et al. 2022; Rathee et al. 2023]. In addition, they have not been adapted to the requirements of gaze estimation (e.g. the prior eye/face knowledge, the in-the-wild data heterogeneity, and the computational performance). We are the first to practically combine FL with state-of-the-art MPC protocols like Overdrive LowGear [Keller et al. 2018] in the context of appearance-based gaze estimation in order to achieve active security against a dishonest majority of servers and allow scalability to a large number of clients.

Privacy-enhancing technologies for gaze. Despite mounting concerns, only few previous works have studied privacy-enhancing technologies for gaze, most of which have focused on gaze behaviour analysis. A common approach is to use differential privacy (DP), i.e., the idea of adding noise according to an ϵ parameter [Dwork et al. 2014; Steil et al. 2019]. This approach, however, often comes at the cost of reduced data utility [Bozkir et al. 2021; Li et al. 2021; Liu et al. 2019]. Others have proposed to add noise in a dataset-dependent manner to increase utility [Bozkir et al. 2021; Steil et al. 2019], which leads to very weak privacy guarantees [Bozkir et al. 2023; Nissim et al. 2007]. Further DP optimizations, e.g. as in [Kifer and Machanavajjhala 2011], are possible, but come with a decrease in the model performance. In stark contrast, privacy for gaze estimation has so far largely been neglected except for [Bozkir et al. 2020; Elfares et al. 2022]. In [Bozkir et al. 2020], a function-specific (support vector regressor) method is proposed using randomized encodings. This method is limited by two data owners and does not protect against malicious adversaries. Our work, therefore aims to fill these gaps and builds on the adaptive federated learning approach proposed in [Elfares et al. 2022]. Adaptive FL comes with many favourable properties, e.g. it scales well in the number of clients, can be applied to any ML model and converges quickly even for a large data heterogeneity across the clients. However, we show in this paper that it remains susceptible to various data leakage attacks. Our new training schemes PrivatEyes solves this privacy issue while maintaining the same model performance and comparable computational cost as [Elfares et al. 2022].

Adversarial attacks. ML models are susceptible to a large number of vulnerabilities and attacks. The trained models can leak information about the private input data through their black-box outputs (e.g. predictions) or their white-box parameters [Zhang et al. 2020a]. As no prior work has investigated adversarial attacks on gaze estimation models, in this paper, we aim (i) to analyse these attacks on gaze data and its training process, (ii) to quantify the amount of information leakage and (iii) to prevent (or at least minimise the effectiveness of) such attacks through protocols that provide formal security guarantees, i.e. which do not only protect against a certain attack but provide security against any (realistic) adversary and attack. Prior works in ML investigated different attacks in isolation, e.g. model-inversion attacks [Fredrikson et al. 2015; He et al. 2019; Wu et al. 2016] (i.e. reconstructing the training data) or inference attacks [Bernau et al. 2019; Li et al. 2020; Shokri et al. 2017] (i.e. inferring private information). Other approaches [Geiping et al. 2020; Hitaj et al. 2017; Zhao et al. 2020] were specific to the FL setup. Given the various different gaze estimation training approaches, we instead construct a new attack DualView that allows us to attack different schemes like data centre training, adaptive FL [Elfares et al. 2022] or our own approach PrivatEyes. Unlike (adversarial) model-inversion attacks [Fredrikson et al. 2015; Hitaj et al. 2017; Zhang et al. 2020a] that reconstruct images that maximally activate the target network, DualView does not only aim to synthesize realistic features but also tries to consistently associate the reconstructed images with the appearance and the gaze features of the training set. DualView is further optimized for the regression task of gaze estimation and therewith differs from classical inference attacks like [Salem et al. 2020; Shokri et al. 2017; Zhao et al. 2020] which were so far only studied for classification tasks. We use DualView among other techniques to show the vulnerabilities of FL training. For

FL applications (without MPC) outside of gaze estimation, similar results have been achieved by Geiping et al. [Geiping et al. 2020].

4 METHOD

Training accurate and generalisable gaze estimation models requires a large number of training images to handle the large variability in in-the-wild eye and face appearances. A common solution is to train gaze estimators collaboratively using gaze data from different owners (a.k.a. clients) but this raises concerns with the highly privacy-sensitive gaze data. Missing protection of this data prevents gaze data owners from taking part in the training process and thereby decreases the model generalisation performance. We address this problem by introducing in a new gaze estimation training approach *PrivatEyes* that provides strong security guarantees without hampering training efficiency or gaze estimation performance. We further construct a new gaze-specific attack that allows us to quantitatively compare the privacy properties of different gaze estimation training schemes like adaptive FL [Elfares et al. 2022] or our new scheme PrivatEyes.

4.1 PrivatEyes: Gaze Estimation Training

FL has recently been introduced as a promising approach for training gaze estimation models [Elfares et al. 2022]. Following [Elfares et al. 2022], we focus on a cross-device centralised horizontal FL approach, with its support for scalable data distributions and a large number of devices, hence matching the requirements of large-scale gaze data collection on multiple devices in the wild. However, centralised FL comes with a major privacy issue: It involves a single central server that aggregates local updates received from all clients. This results in the server having access to a significant amount of personal information that could be used to reconstruct (parts of) the original training data. We address this issue in PrivatEyes by replacing the single server with multiple ones (with secret sharing) as shown in Figure 1. This allows us to avoid the leakage of individual client updates and thereby significantly reduce the attack surface.

The training process. Centralised FL with a single server (e.g. [Elfares et al. 2022]) works as follows: For each of t communication rounds the server selects a random cohort C of the N available clients and broadcasts a gaze estimation model to these clients with the corresponding hyper-parameters, weights, biases, number of rounds, and number of local epochs. In the first round ($k = 1$) the model sent is some initial model (e.g. a model pre-trained on public data) and in all following rounds ($1 < k \leq t$) it is the output model (OM_{k-1}) of the previous round. Once a client $C_j \in C$ receives the model (represented by some weight vector w) from the server, it starts locally training a new model $\mathsf{IU}_{j,k}$ on her private data $D_{j,k}$. It outputs this individually trained model (represented by some weight vector a_j) to the server. The server aggregates the individual updates $\mathsf{IU}_{j,k}$ to get a new output model OM_k represented by $\frac{1}{|C|}\sum_{j\in J} a_j$. The process continues until a final output model OM_t is reached and then published to all clients.

Secret sharing. In contrast, in PrivatEyes, as shown in Fig. 1, a client C_j no longer sends its update a_j in plain to an aggregating server but instead encrypts it as a *secret sharing* for n servers. That is, for each of the n servers S_i, a random number $[a_j]_i$ ($1 \leq i \leq n$) is selected such that $a_j = \sum_{i=1}^{n}[a_j]_i$ holds; $[a_j]_i$ belongs to a previously agreed finite field, such as $\mathbb{Z}_q = 0, 1, ..., q-1$, for some prime number q (e.g. a common size is $q \approx 2^{127}$). The client then sends $[a_j]_i$ to S_i, i.e. each server gets only one share of the secret update a_j. As long as at least one server, say S_2, is honest and does not reveal its share $[a_j]_2$, MPC guarantees that the other servers cannot gain *any* information about a_j. E.g., as shown in Fig. 1, the client C_1 wants to share $a_1 = 3$ as an element of a finite field $\mathbb{Z}_{23}$ with three servers S_1, S_2, S_3. Then C_1 chooses (arbitrarily) three numbers $[a_1]_1, [a_1]_2, [a_1]_3$ such that

$([a_1]_1 + [a_1]_2 + [a_1]_3) \bmod 23 = a_1 = 3$. For example, $[a_1]_1 = 5, [a_1]_2 = 17$ and $[a_1]_3 = 4$ with $(5 + 17 + 4) \bmod 23 = 26 \bmod 23 = 3$. Even if S_1 and S_3 collude and exchange their shares, the unknown share of S_2 makes all possible values of $a_j \in \mathbb{Z}_{23}$ equally likely from the perspective of S_1 or S_3, i.e. no information on the actual a_j is leaked.
Once each server S_i has received a share $[a_j]_i$ from each client C_j, the servers start an MPC protocol to compute a *sharing* of the new global model represented by a weight vector w. At the end of the MPC computation, each server S_i has a share $[w]_i$. As before, if only one server is honest, the MPC protocol guarantees that no server gets any information on w and that w has been computed correctly. Otherwise, the protocol aborts. Finally, all servers return their shares $[w]_i$ to all clients C_j and each C_j can locally reconstruct $w = \sum_{i=1}^{n} [w]_i$ by simply adding up the shares.
Clients provide their inputs by following the protocol of Damgard et al. [Damgård et al. 2012] (simplified above). It guarantees along with checks carried out in the MPC protocol that servers are forced to use the inputs (the shares) given to them by the clients when performing the MPC protocol. I.e., even malicious servers cannot change the client inputs. Hence, if an output is produced by the MPC computation, this output is (mathematically provably) guaranteed to correspond to the inputs provided by the clients; otherwise, no output is produced. For an efficient aggregation of the individual model updates, the MPC protocol evaluates an adaptive optimisation protocol [Reddi et al. 2020] that has been shown to adapt the model updates to the clients' heterogeneous gaze data distribution [Elfares et al. 2022] while only relying on low-computational operations on secret values for efficient privacy-preserving training (a main challenge in MPC-based protocols). Following the classical FL approach, the FL training is repeated until a certain number of rounds is reached and the final *output model* has been generated. For more details on the MPC protocol, formal security and privacy guarantees, and security proofs please refer to [Damgård et al. 2016, 2012; Keller et al. 2018].

Security guarantees of PrivatEyes. PrivatEyes provides security guarantees even if only one server (out of n servers) is honest, i.e., follows the protocol and does not collude with other servers or clients:

(1) If servers do not use the input provided to them by (honest) clients the protocol aborts.
(2) If servers deviate from the prescribed protocol the protocol aborts.
(3) If an output is produced by the protocol, then this is guaranteed to be the correct global model, i.e., the gaze estimation model that would have been obtained if all servers were honest and used the inputs provided to them by the clients.
(4) A server gains no information from the gaze data of a (honest) client beyond what is available publicly.

4.2 DualView: Gaze-specific Attack

Gaze estimation models are susceptible to a large number of vulnerabilities and attacks (c.f. Section 3). Nonetheless, such attacks have never been studied in the gaze community. In this work, we present a new gaze-specific attack, which we call DualView, to evaluate the amount of information leakage from gaze estimation models and their training process.

Threat Model. We assume an adversary $\mathcal{A}$ that can use all available knowledge to determine properties of the private training data sets used by the clients. The knowledge of $\mathcal{A}$ contains information Pub that is publicly available, e.g., the final output model or the model architecture. But it can also contain leaked knowledge Leak that the adversary gathered from the training process, e.g. by colluding with some of the servers. In particular, depending on the actual training scheme and the number of dishonest parties, the knowledge available to $\mathcal{A}$ differs. For example, in centralised

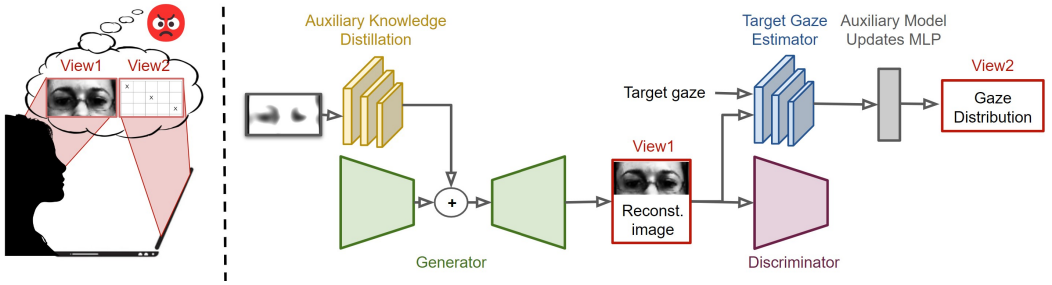

Fig. 2. DualView demonstrates the amount of gaze-specific information leakage by reconstructing the user's appearance (View1: how the user looks like) as well as inferring the corresponding gaze distribution (View2: where the user is looking) of the private dataset used to train a target gaze estimation model.

FL an adversary that colludes with the aggregating server sees all individual model updates IU and can use these updates in his attack. In contrast, in PrivatEyes , due to the use of MPC (c.f. Fig. 1), an adversary that colludes with one (or up to $n-1$) servers does not get any information on the individual updates IU and can, therefore, not use this information in the attack.

Goal of the attack. For our attack, we assume the adversary $\mathcal{A}$ wants to reconstruct private training data (given his available information from Pub and Leak) or some property of the training data (e.g. its gaze distribution). We then determine the success of $\mathcal{A}$ in Section 5.3 by measuring how much his reconstruction differs from the original training data (w.r.t. the ground truth). Naturally, if a gaze estimation training scheme like centralised FL leaks information, then this information leakage likely contains gaze-related information. Unfortunately, this is not the only information one can usually deduce from the leaked data, e.g. from a model update. For example, a model update usually also contains appearance-related information (although appearance was not the main objective of the training). Hence, in order to get a more complete picture of the overall information leakage of a gaze estimation training scheme like centralised FL or PrivatEyes, we need an attack that can successfully deduce information about a user's appearance (i.e. view1: how the user looks like) as well as the respective gaze angle distribution (i.e. view2: where the user is looking) in the respective training set. Our attack DualView (cf. Fig. 2) addresses these two objectives.

Our attack DualView. Technically, DualView consists of a generative adversarial network (GAN) that trains a generator G and a discriminator D, where the generator takes a random latent vector as input and generates face/eye images that look (superficially) authentic to human observers with realistic characteristics. The discriminator evaluates the reconstructed images by distinguishing the reconstructed images from the true data distribution. The generator takes a random input (sampled from a latent space) as a seed and is trained until it succeeds at fooling the discriminator via independent backpropagation with a training loss composed of the Wasserstein loss and a diversity loss [Yang et al. 2019] (to generate a large output distribution for our domain-agnostic reconstructed images). As a training dataset, we use a high-quality reference dataset which is publicly available, namely GazeCapture or LPW datasets, for face and eye images respectively. We choose these datasets since they come with a large number of participants, e.g. 1,474 in the case of GazeCapture, and therefore result in a generator that can reproduce realistic reconstructions of training sets.
Note that so far our attack only used public information, i.e. the publicly available data sets GazeCapture and LPW. In particular, DualView generates the GAN independently of the specific training scheme and its leakage Leak. We next want to explain how DualView uses the additional knowledge Leak gained from the gaze estimation training (e.g. by corrupting servers or clients) to reconstruct

the training data sets. Generally, the knowledge in Leak will contain some model updates, e.g. the individual model updates revealed in centralised FL. For the exact form of Leak in the different training schemes we refer to Section 5.3. Now DualView checks how accurately such a model from Leak can determine the gaze angle on generated images by G. This leads to a loss function L_{gaze}. Similarly, DualView determines a loss L_{prior} (the discriminator loss that penalises inauthentic reconstructed images) and an appearance loss $L_{\text{appearance}}$ (a cross-entropy loss). Overall he gets a total loss function[1]: $L_{\text{total}} = \alpha L_{\text{prior}} + \beta L_{\text{gaze}} + \gamma L_{\text{appearance}}$. The gaze loss is then fed to a simple multi-layer perception (MLP)-based autoencoder that regresses it to the corresponding gaze direction distribution. In addition, a kernel density estimation (KDE) function derives a probability density function to compute a Kullback-Leibler (KL) divergence in order to evaluate the predicted gaze distribution.
Now generally, Leak contains more than a single model, e.g. it contains all individual model updates IU from all rounds in centralised FL. We introduce these different model updates in Leak iteratively. Namely, DualView first optimizes w.r.t. the first round model update. The resulting reconstructed images are then (re-)introduced as auxiliary knowledge following [Zhang et al. 2020a], i.e. after passing them through an auxiliary knowledge distillation network with dilated convolution layers that insert gaps between the kernel elements (i.e. pixel skipping) to cover a larger receptive field and therewith provide a generic eye/full-face skeleton to the generator. Similarly, all available model updates in Leak (and the final output model in Pub) are used by the MLP-based autoencoder as auxiliary knowledge.
Finally, the adversary receives a reconstruction of the private training set that agrees best with his available knowledge from Pub and Leak. We remark again that the quality of this reconstruction of course depends on Leak and that a DualView instance which can use all individual model updates and all round-wise output models (i.e. in the case of centralised FL) will generally produce a more accurate reconstruction than a DualView instance which can only use the output models without individual updates (i.e. in the case of PrivatEyes).

5 EXPERIMENTS

To evaluate our method, we use several different datasets to cover the different gaze estimation setups and to prove the advantage of PrivatEyes and our privacy guarantees in real-world scenarios. The datasets cover different appearances (ethnicities, genders, glasses, and make-up), illumination conditions (indoor and outdoor), gaze distributions, head poses, modalities (images and videos with eyes or full-faces), and recording setups (remote vs. near-eye).

We mainly conducted experiments on the MPIIGaze [Zhang et al. 2015], MPIIFaceGaze [Zhang et al. 2017c], and NVGaze [Kim et al. 2019] datasets. The MPIIGaze and MPIIFaceGaze datasets contain 213,659 eye and full-face images, collected in the wild from 15 participants over the course of several months. We used the gaze estimation models originally proposed in both works and trained them using PrivatEyes. The models take eye/full-face images as input and regress them to gaze directions in normalised space. We conducted further experiments for near-eye gaze estimation on the NVGaze [Kim et al. 2019] dataset, with 32 participants recorded by a near-eye infrared camera, along with their proposed gaze estimation models. Furthermore, we used GazeCapture [Krafka et al. 2016], the gaze estimation dataset with the largest number of participants (1,474), (1) along with its corresponding iTracker CNN, to further analyse the scalability of PrivatEyes and (2) to train our DualView attack for full-face inputs. Similarly, the LPW [Tonsen et al. 2016] dataset, which contains videos of 22 participants recorded by a head-mounted eye tracker was used to train the DualView attack for eye inputs. We used the MP-SPDZ [Keller 2020] framework with Overdrive

[1] α, β, and γ are weighting constants selected according to a hyperparameter search. E.g. for attacking full-faces, $\alpha = 10$, $\beta = 6$, and $\gamma = 4$.

LowGear [Keller et al. 2018] (the state-of-the-art maliciously secure MPC protocol for a small number of servers) to implement PrivatEyes and the ML-Doctor [Liu et al. 2022a] framework to also perform standard ML attacks (c.f. Appendix D) in addition to our new attack DualView.

5.1 Baseline methods

We compared our training approach PrivatEyes (cf. Section 4.1) with three baselines and evaluated them in terms of gaze estimation error, robustness to gaze data leakage attacks, client-server communication, and computational complexity:

- **Data centre training:** In this approach, the clients' datasets are collected in one central server storage. The model is directly trained on all samples and, hence, data-sharing concerns arise.
- **Adaptive FL:** This is the state-of-the-art centralised federated learning-based gaze estimation training proposed in [Elfares et al. 2022] in which the individual model updates are treated as pseudo-random gradients for aggregation.
- **Generic MPC:** The state-of-the-art actively secure MPC approach (without FL) from [Keller et al. 2018] where the clients provide the inputs in a secret-shared form and the servers run the whole training as an MPC protocol (cf. Section 2). In particular, this approach uses only MPC and not federated training like PrivatEyes.

5.2 Gaze estimation performance

Gaze estimation performance was calculated as the mean angular error between the predicted and ground-truth gaze directions. As shown in Tab. 1, for data centre training, the gaze estimation model yielded the best performance on all datasets. This is expected given that this training approach can have direct access to original images. Of course, privacy in the data centre training relies on the trustworthiness of the central server—if the server is dishonest, *all* input data gets leaked (c.f. Section 4.1). For adaptive FL and PrivatEyes, we used the same number of rounds, epochs, and hyper-parameters as in [Elfares et al. 2022]. As expected (c.f. Section 4.1), the performance of adaptive FL and PrivatEyes is identical. In addition, we investigated how the performance loss between data centre on the one side and adaptive FL and PrivatEyes on the other side varies from client to client. For example, Fig. 8 shows a random sample of 10 clients with the respective errors. In general the gaze estimation fairness (i.e. the difference between the minimum and maximum gaze angular error across all clients) is 1.3°, 1.5°, 1.0°, and 1.5° for MPIIGaze, MPIIFaceGaze, NVGaze, and GazeCapture respectively. Finally, we were not able to run the full generic MPC benchmark for the performance, since these protocols are far too inefficient to handle realistic data sets and realistic numbers of clients (c.f. Section 5.4)[2].

Our evaluation in Tab. 1 also indicates how our approach PrivatEyes scales with the number of clients and the size of the training sets. From the four data sets in Tab. 1, MPIIGaze, MPIIFaceGaze, and NVGaze datasets represent data sets with a small number of participants (15 to 32), while GazeCapture covers a large number of participants (1,474) and therewith follows classical cross-device federated training conventions (which usually come with ≥ 100 clients [Kairouz et al. 2019]). In line with [Kairouz et al. 2019], the performance gap between data centre training/generic MPC and PrivatEyes/adaptive FL becomes smaller with a larger number of clients and larger datasets.

[2]The main reason for this lack of efficiency is that operations common in ML, e.g. comparisons, polynomial evaluations and generally floating point operations, are not naturally supported by the MPC protocols, which are optimised for arithmetic operations over finite fields. This leads to costly translations and often comes with a loss in precision and therefore a loss in the model performance. However, theoretical considerations ensure that generic MPC will perform as good as data centre training while runtime was estimated by only running the main operations and generalising them to the entire training.

	Data Centre/Generic MPC Adaptive FL		PrivatEyes
MPIIGaze	6.3°	7.6°	7.6°
MPIIFaceGaze	6.2°	7.4°	7.4°
NVGaze	0.8°	2.1°	2.1°
GazeCapture	4.0°	4.7°	4.7°

Table 1. Mean angular error for different gaze estimation datasets.

5.3 DualView performance

In this section, we want to compare different training schemes, i.e. data centre training, adaptive FL, PrivatEyes and generic MPC, w.r.t. their privacy leakage. For the two federated approaches, we consider a training with $t = 10$ rounds. To simplify the evaluation we further assume that $C = \{1, \ldots, N\}$, i.e. that all clients are chosen in each round.

We assume that at least one and at most $n - 1$ servers (if $n \geq 2$) are corrupted. Hence, the data centre training is completely insecure and the one corrupted central server leaks all private training data. For adaptive FL, the corrupted server leaks all individual model updates and all output models in all rounds, i.e. $\mathsf{Leak}_{adaptive\ FL} = \{\mathsf{IU}_{j,k}, \mathsf{OM}_k : 1 \leq k \leq n\}$. Furthermore, we assume that at least one client is corrupted. Recall from Section 4.1 that in PrivatEyes each client receives the round-wise output models and hence a corrupted client leaks these output models to the adversary. However, the individual models (of an honest party) are only send in shared form in PrivatEyes and therefore no corrupted server or client can deduce information about them. Thus, in PrivatEyes, we have $\mathsf{Leak}_{PrivatEyes} = \{\mathsf{OM}_k : 1 \leq k \leq t\}$. Finally, in generic MPC, no information is leaked apart from public information, e.g. the final output model OM_t.

We then run three instances of DualView: (i) against adaptive FL, where DualView uses $\mathsf{Leak} = \mathsf{Leak}_{adaptive\ FL}$; (ii) against PrivatEyes , where $\mathsf{Leak} = \mathsf{Leak}_{PrivatEyes}$ contains only all roundwise output model updates; (iii) against generic MPC, where DualView receive no non-public information, i.e. $\mathsf{Leak}_{\mathrm{MPC}} = \{\}$ and DualView stops once he used all publicly available information for the reconstruction. In each of the three cases DualView outputs a reconstruction (i.e. View1 and View2) of the training set of each client at each round. We then evaluate the reconstructions in terms of:

- **Appearance similarity:** A user study (N = 60 respondents) was conducted to see if users could correctly map the reconstructed images to the corresponding participant.[3] Additionally, the participants were asked to rate the similarity of a reconstruction with a qualitative visual score. See Appendix A for more details.
- **Pixel-wise similarity:** A Peak Signal-to-Noise Ratio (PSNR) was used to quantify the fluctuation between the original (private) image and the corresponding reconstructed image. Higher PSNR scores indicate higher similarities.
- **Gaze direction similarity:** A mean angular error (MAE) was calculated to capture the gaze direction similarity between the ground truth and the reconstructed images.
- **Gaze distribution similarity:** A KL-divergence is used to calculate the statistical distance between the ground truth and the inferred gaze probability distributions of all images used to train a target model. A value of 0 indicates identical quantities of information (i.e. identical gaze directions).

Our results for the MPIIFaceGaze dataset are included in Tab. 2 and Fig. 3. Some additional results are contained in Appendix B for the remaining datasets. Note that the same performance behaviour can be seen across all datasets (c.f. Appendix B). As expected our evaluation shows that more

[3]A re-identification network could serve as a metric for appearance similarity. However, we opted for the user study to avoid bias.

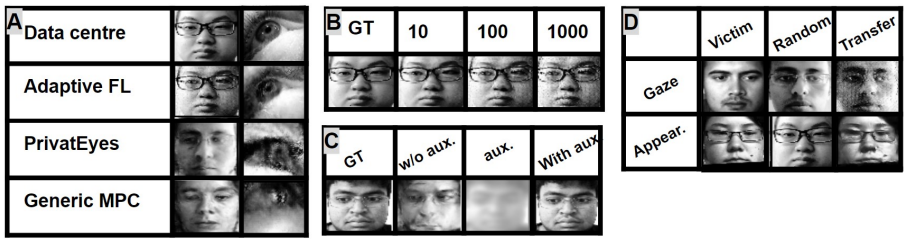

Fig. 3. Sample reconstruction results. (A) Reconstructed face/eye images of the same participant for the different baselines (Data centre is the ground truth due to the direct access to the raw data). (B) Effect of the number of training samples per round. (C) Using previous reconstructions as auxiliary knowledge against adaptive FL training affects the reconstruction. (D) Effect of the attack loss function, e.g. transferring the gaze of a victim to a random face (1st row) and transferring the appearance of a victim to a random gaze (2nd row) in adaptive FL.

leaked information naturally leads to better reconstructions of the training set, i.e. the data leakage increases from generic MPC and PrivatEyes to adaptive FL and data centre training. Surprisingly, PrivatEyes reaches an on-par privacy level as generic MPC which is (at least concerning privacy) as good as one might hope for, since for generic MPC an attack can only use publicly available information, namely the final output model. This means that the knowledge of several intermediate output models from different rounds does not really improve the ability of the adversary. In other words, the aggregated intermediate output models contain (already in our evaluation setup with only 15 clients) only a small amount of individual training data.

Moreover, we observe that a larger number of rounds, and hence more leakage, results in more accurate appearance reconstructions (re-identification, visual score, and PSNR) given that the user's appearance stays the same over all training sets of a client. However, the later updates only improve the reconstructions slightly (if at all). The reason is that the gaze estimation model converges, i.e. towards the last round(s), less information could be leaked from the target models as the effect of the training data on the model updates becomes small, and the multiple aggregation steps in FL dissipate individual information (c.f. Fig. 9). Figure 3-C shows the strong effect of the earlier updates introduced as auxiliary knowledge compared to a reconstruction (w/o aux.) that only relies on the last round update.

For the gaze-related properties (MAE, KL), DualView tries to reconstruct the gaze angle distribution used during training. However, in contrast to appearance, the gaze angles in the k-th round training set $D_{j,k}$ of a client C_j do not (necessarily) depend on the previous training sets. In particular, the previously learned models (if introduced as aux.) can only slightly improve the final gaze reconstruction of $D_{j,t}$ as they contain little information on $D_{j,t}$ due to the described convergence issue. Nonetheless, the auxiliary knowledge provides information about which parameters already converged. For instance, in a late round, some training samples will no longer update the gaze estimator gradients (i.e. have a negligible effect on the training) while the previous model updates (if introduced as aux.) provide information about which parameters already converged. DualView uses this knowledge by allowing weights corresponding to already converged parameters to nevertheless contribute to the reconstruction (c.f. Appendix C).

With respect to scaling Fig. 3-B shows the effect of the number of training samples per round on privacy. In general, a smaller number of training samples per round (i.e. overfitting) leaks more information with a slower convergence (i.e. more training rounds). We also investigate the effect of our attack loss by replacing the total loss function from Section 4.2 by L_{gaze} or $L_{\text{appearance}}$. For example, Fig. 3-D shows how the adversary can reconstruct (transfer) the gaze of some target client

	Re-identified	Visual score	PSNR	MAE	KL
Data centre	15/15	100%	max.	0	0
Adaptive FL	15/15	83%	19.8	8.1	0.201
PrivatEyes	0/15	7%	5.1	10.6	0.269
Generic MPC	0/15	3%	5.0	10.8	0.273

Table 2. DualView performance on the MPIIFaceGaze dataset.

	Data Centre	Adaptive FL	PrivatEyes	Generic MPC
Runtime	~ 357.0 s	~ 357.2 s	~ 411.9 s	~ 10 months
Communication	~ 909 MB	~ 1 MB/round	~ 7 MB/round	~ 3840 MB

Table 3. Training runtime and communication for the baselines vs PrivatEyes on MPIIGaze on a NVIDIA V100 GPU. Generic MPC and PrivatEyes both use three servers; adaptive FL and data centre both use one server and 10 training rounds.

(victim) to some randomly chosen face, and vice versa. Although the advantage of PrivatEyes over adaptive FL is larger when more private information is leaked (i.e. with our proposed total loss), such effects could be of independent interest (e.g. for image forensics such as Deepfakes [Westerlund 2019] or gaze data synthesis [Qin et al. 2023]).

In summary, since PrivatEyes does not leak any information on individual updates, attacks can only use the aggregated models at each round. This represents a key advantage compared to i) data centre training which requires access to the input images and ii) adaptive FL which allows access to both the individual model updates (IU) as well as the output model (OM) of each round. Results of other well-established data leakage attacks further confirm our findings (c.f. Appendix D). Finally, PrivatEyes reaches an on-par privacy level as generic MPC, which comes with optimal privacy guarantees (under our assumption that one server is honest). However, generic MPC remains inefficient in real-world applications, e.g. gaze estimation (c.f. Section 5.4 below).

5.4 Efficiency

While the previous results suggest that generic MPC achieves slightly better performance and privacy than PrivatEyes, this does not imply that generic MPC can be successfully used in practice. The main obstruction to employing generic MPC is that it can currently not be implemented efficiently. Fortunately, our combination of FL and MPC does not suffer the same disadvantage. The reason is that the MPC computation in generic MPC is much more complex than in our FL-based scheme PrivatEyes. In PrivatEyes, the MPC protocol is dominated by additions of the (shares of) the individual updates. These additions can be computed locally (i.e. without communication among the servers) and therewith efficiently by each server. In contrast, a generic MPC evaluation must additionally handle all the non-linear training operations, which in PrivatEyes are done locally by the clients. This includes e.g. approximations of activation functions by high-degree polynomials or derivatives. By design of the MPC protocols, these non-linear operations lead to a significant communication overhead among the servers. For instance, the first convolution layer on the MPIIGaze CNN requires $\sim 900k$ and $\sim 4{,}000k$ communicated 128-bit values per iteration for the forward and backward pass, respectively (with a total of $\sim 30{,}000K$ for the entire CNN). As a result, the estimated training time for three parties on a single computer with no network delay is ~ 10 months and 3840 MB communication for training and ~ 4 hrs for inference. In real-world setups, e.g. over the Internet, network delays and bandwidth restrictions further slow down the training (significantly). Overall, the generic MPC approach is currently not ready for real-world applications in gaze estimation given its runtime, as shown in Tab. 3. While PrivatEyes is much more efficient it nevertheless comes with

a small efficiency loss compared to adaptive FL, since the clients need to send one share of their individual update to each server (plus some additional communication needed to get an actively secure scheme as described in Section 4.1). For instance Table 3 shows that PrivatEyes for 3 servers comes with an approx. factor 7 communication overhead compared to adaptive FL, which is well in accordance with theory (it results from sending 3 shares plus authentication for active security). The communication scales (just as generic MPC) linearly in the number of servers. The runtime on the other hand is more dominated by the local training of the clients and not affected strongly by the communication overhead. Please note also, that shares can be sent in parallel to all servers (and back) and hence the input and output phase in adaptive FL and PrivatEyes have similar runtime (although more data is sent in PrivatEyes). We generally think that the small absolute increase in runtime and communication is justified given the significantly better privacy guarantees of PrivatEyes over adaptive FL. Finally, data centre training needs significantly more communication since all the raw training data has to be sent to the central server.

In summary, we have seen that our combination of FL and MPC in PrivatEyes has significant advantages compared to all other currently available gaze estimation training approaches (on distributed data sets). It comes with (i) good gaze estimation performance (which is only slightly worse than data centre training and equal to centralized FL approaches), (ii) reasonable computational and communication costs, (iii) strong privacy guarantees that are almost optimal and significantly better than centralised FL approaches and data centre training.

6 DISCUSSION

In the following, we discuss the potential of PrivatEyesalong three axes: the privacy guarantees for the gaze data, the fairness of the gaze estimator, and the feasibility of our trust assumption.

Privacy guarantees for the gaze data. FL structurally prevents access to the client's raw data through data minimisation (i.e. the *individual model updates)*. In addition, FL facilitates data-stewardship to ensure that clients control and approve of how their data will be used and have governance over their data, hence, transparency and consent principles are applied [Bonawitz et al. 2022]. Unfortunately, our evaluation in Section 5 shows that FL is still vulnerable to attacks which can deduce very accurate reconstructions of the training data from the individual model updates. PrivatEyes does not leak the individual model updates and therefore reaches a far better privacy level. In particular, PrivatEyes provides provable privacy guarantees as long as at least one server remains honest (see Section 4 for details). Other approaches with similar privacy guarantees like generic MPC, are currently too inefficient to be applied in real-world gaze applications.

Fairness of the gaze estimator. In collaborative approaches, e.g. FL or PrivatEyes, concerns about model fairness are magnified due to the heterogeneous data distribution across clients. Nonetheless, our experiments (Section 5.2) show that (i) PrivatEyes is able to adapt to the different individual model updates and yields good performance for each client (Fig. 8), (ii) increases the generalisation capability of gaze estimators, and (iii) maintains a high scalability performance. PrivatEyes, therefore, incentivises collaborative training (a main goal of our paper) while preserving data privacy.

Trust and dishonest majority. Trust plays a crucial role in shaping clients' willingness to share information, such as model updates. The evolving capabilities of AI and adversarial attacks, as discussed in Section 3, have given rise to the 'trust crisis' [Yu et al. 2017]. Addressing this issue, [Steil et al. 2019] conducted a comprehensive survey on users' attitudes towards sharing their eye data. The study revealed that clients are more inclined to share their data if the co-owner, e.g. a server in PrivatEyes, belongs to entities perceived as trustworthy, such as governmental, healthcare, education, or research institutes. Conversely, there is a lack of trust in international and profit-oriented

organizations. Interestingly, clients are more open to sharing their data in aggregated forms, e.g. output models in PrivatEyes, while being reluctant to share raw data. Hence, PrivatEyes, specifically guided by the dishonest majority assumption, emerges as a solution for (i) ensuring the trustworthy processing of eye data for any service provider (i.e. server) if only one server (i.e., the sole honest server) is affiliated with trusted entities and (ii) aligning with clients' trust dynamics.

Limitations and future work. One of the major challenges of collaborative gaze estimation research is to transfer conventional gaze estimation methods to decentralised datasets. For example, the selection of hyper-parameters and their optimisation (i.e. AutoML [Kohavi and John 1995]) currently does not have an efficient distributed counterpart (especially for secure setups). Furthermore, PrivatEyes, similar to other FL paradigms (e.g [Elfares et al. 2022]), considers a supervised gaze estimation task where data annotation is assumed to be available at each client. Applying FL (or even PrivatEyes) to semi-supervised or unsupervised learning (i.e. [Jindal and Manduchi 2022; Yu and Odobez 2020]) also remains an open problem.

7 CONCLUSION

We presented PrivatEyes—the first privacy-preserving training approach for appearance-based gaze estimation methods that combines FL with MPC. Our evaluation shows that PrivatEyes reaches the same gaze estimation performance as the currently best (non-secure) distributed training scheme [Elfares et al. 2022], is domain-agnostic (i.e. can be used with any gaze estimation model and dataset), and can scale to a large number of clients, with a rather moderate efficiency overhead compared to centralised FL. Finally, PrivatEyes provides strong security guarantees (cf. Section 4). It then reaches almost optimal privacy guarantees, if only one out of n servers is honest, and is therewith significantly better than centralised FL and data centre training (cf. Section 5). Overall, PrivatEyes currently provides the most practical and private gaze estimation training on distributed datasets. It hits a "sweet spot" in terms of privacy, model accuracy, and performance.

ACKNOWLEDGMENTS

M. Elfares was funded by the Ministry of Science, Research and the Arts Baden-Württemberg in the Artificial Intelligence Software Academy (AISA). Z. Hu was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016. P. Reisert and R. Küsters were supported by the German Federal Ministry of Education and Research under Grant Agreement No. 16KIS1441 and from the French National Research Agency under Grant Agreement No. ANR-20-CYAL-0006 (CRYPTECS project). A. Bulling was funded by the European Research Council (ERC; grant agreement 801708).

REFERENCES

Mihai Bâce, Sander Staal, and Andreas Bulling. 2020. Quantification of Users' Visual Attention During Everyday Mobile Device Interactions. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 1–14. https://doi.org/10.1145/3313831.3376449

Shumeet Baluja and Dean Pomerleau. 1993. Non-intrusive gaze tracking using artificial neural networks. *Advances in Neural Information Processing Systems* 6 (1993).

Daniel Bernau, Philip-William Grassal, Jonas Robl, and Florian Kerschbaum. 2019. Assessing differentially private deep learning with membership inference. *arXiv preprint arXiv:1912.11328* (2019).

Pradipta Biswas et al. 2021. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3143–3152.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.

Kallista Bonawitz, Peter Kairouz, Brendan Mcmahan, and Daniel Ramage. 2022. Federated learning and privacy. *Commun. ACM* 65, 4 (2022), 90–97.

Efe Bozkir, Onur Günlü, Wolfgang Fuhl, Rafael F Schaefer, and Enkelejda Kasneci. 2021. Differential privacy for eye tracking with temporal correlations. *Plos one* 16, 8 (2021), e0255979.

Efe Bozkir, Süleyman Özdel, Mengdi Wang, Brendan David-John, Hong Gao, Kevin Butler, Eakta Jain, and Enkelejda Kasneci. 2023. Eye-tracked Virtual Reality: A Comprehensive Survey on Methods and Privacy Challenges. *arXiv preprint arXiv:2305.14080* (2023).

Efe Bozkir, Ali Burak Ünal, Mete Akgün, Enkelejda Kasneci, and Nico Pfeifer. 2020. Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework. In *ACM symposium on eye tracking research and applications*. 1–5.

Andreas Bulling and Hans Gellersen. 2010. Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12.

Andreas Bulling and Daniel Roggen. 2011. Recognition of Visual Memory Recall Processes Using Eye Movement Analysis. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 455–464. https://doi.org/10.1145/2030112.2030172

Andreas Bulling, Christian Weichel, and Hans Gellersen. 2013. EyeContext: Recognition of high-level contextual cues from human visual behaviour. In *Proceedings of the sigchi conference on human factors in computing systems*. 305–308.

Lukas Burkhalter, Hidde Lycklama, Alexander Viand, Nicolas Küchler, and Anwar Hithnawi. 2021. Rofl: Attestable robustness for secure federated learning. *arXiv preprint arXiv:2107.03311* (2021).

Mihai Bâce, Sander Staal, and Andreas Bulling. 2020. Quantification of Users' Visual Attention During Everyday Mobile Device Interactions. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. https://doi.org/10.1145/3313831.3376449

Virginio Cantoni, Chiara Galdi, Michele Nappi, Marco Porta, and Daniel Riccio. 2015. GANT: Gaze analysis technique for human identification. *Pattern Recognition* 48, 4 (2015), 1027–1038.

Jixu Chen and Qiang Ji. 2008. 3D gaze estimation with a single camera without IR illumination. In *2008 19th International Conference on Pattern Recognition*. IEEE, 1–4.

Jinsoo Choi, Byungtae Ahn, Jaesik Parl, and In So Kweon. 2013. Appearance-based gaze estimation using kinect. In *2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence*. IEEE, 260–261.

Amrita Roy Chowdhury, Chuan Guo, Somesh Jha, and Laurens van der Maaten. 2021. EIFFeL: Ensuring Integrity for Federated Learning. *arXiv preprint arXiv:2112.12727* (2021).

Ivan Damgård, Kasper Damgård, Kurt Nielsen, Peter Sebastian Nordholt, and Tomas Toft. 2016. Confidential benchmarking based on multiparty computation. In *International Conference on Financial Cryptography and Data Security*. Springer, 169–187.

Ivan Damgård, Valerio Pastro, Nigel Smart, and Sarah Zakarias. 2012. Multiparty computation from somewhat homomorphic encryption. In *Annual Cryptology Conference*. Springer, 643–662.

Ye Dong, Xiaojun Chen, Kaiyun Li, Dakui Wang, and Shuai Zeng. 2021. FLOD: Oblivious defender for private Byzantine-robust federated learning with dishonest-majority. In *European Symposium on Research in Computer Security*. Springer, 497–518.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

Mayar Elfares, Zhiming Hu, Pascal Reisert, Andreas Bulling, and Ralf Küsters. 2022. Federated Learning for Appearance-based Gaze Estimation in the Wild. In *Proc. NeurIPS Workshop on Gaze Meets ML (GMML)*. 1–11.

Myrthe Faber, Robert Bixler, and Sidney K D'Mello. 2018. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods* 50, 1 (2018), 134–150.

Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Helen Möllering, Thien Duc Nguyen, Phillip Rieger, Ahmad-Reza Sadeghi, Thomas Schneider, Hossein Yalame, et al. 2021. SAFELearn: Secure aggregation for private federated learning. In *2021 IEEE Security and Privacy Workshops (SPW)*. IEEE, 56–62.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM CCS*. 1322–1333.

Filippo Galli, Sayan Biswas, Kangsoo Jung, Tommaso Cucinotta, and Catuscia Palamidessi. 2022. Group privacy for personalized federated learning. https://doi.org/10.48550/ARXIV.2206.03396

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems* 33 (2020), 16937–16947.

Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International conference on machine learning*. PMLR, 201–210.

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* 32, 3 (2009), 478–500.

Zecheng He, Tianwei Zhang, and Ruby B Lee. 2019. Model inversion attacks against collaborative inference. In *Proceedings of the 35th Annual Computer Security Applications Conference*. 148–162.

Craig Hennessey, Borna Noureddin, and Peter Lawrence. 2006. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the 2006 symposium on Eye tracking research & applications*. 87–94.

Keita Higuchi, Soichiro Matsuda, Rie Kamikubo, Takuya Enomoto, Yusuke Sugano, Junichi Yamamoto, and Yoichi Sato. 2018. Visualizing gaze direction to support video coding of social attention for children with autism spectrum disorder. In *23rd International Conference on Intelligent User Interfaces*. 571–582.

Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. 2017. Deep models under the GAN: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 603–618.

Philip S Holzman, Leonard R Proctor, Deborah L Levy, Nicholas J Yasillo, Herbert Y Meltzer, and Stephen W Hurt. 1974. Eye-tracking dysfunctions in schizophrenic patients and their relatives. *Archives of general psychiatry* 31, 2 (1974), 143–151.

Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. 2018. Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience* (2018), 105.

Michael Xuelin Huang, Tiffany CK Kwok, Grace Ngai, Stephen CF Chan, and Hong Va Leong. 2016a. Building a personalized, auto-calibrating eye tracker from user interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5169–5179.

Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2016b. Stressclick: Sensing stress from gaze-click patterns. In *Proceedings of the 24th ACM international conference on Multimedia*. 1395–1404.

Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2017. Screenglint: Practical, in-situ gaze estimation on smartphones. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2546–2557.

J Thomas Hutton, JA Nagel, and Ruth B Loewenson. 1984. Eye tracking dysfunction in Alzheimer-type dementia. *Neurology* 34, 1 (1984), 99–99.

Swati Jindal and Roberto Manduchi. 2022. Contrastive Representation Learning for Gaze Estimation.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, K. A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G.L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2019. Advances and Open Problems in Federated Learning. (2019). https://arxiv.org/abs/1912.04977

Marcel Keller. 2020. MP-SPDZ: A versatile framework for multi-party computation. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 1575–1590.

Marcel Keller, Valerio Pastro, and Dragos Rotaru. 2018. Overdrive: making SPDZ great again. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 158–189.

Daniel Kifer and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 193–204.

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 2019 CHI conference on human factors in computing systems*. 1–12.

Ron Kohavi and George H John. 1995. Automatic parameter selection by minimizing estimated error. In *Machine Learning Proceedings 1995*. Elsevier, 304–312.

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*. 2176–2184.

Jing Li, Zejin Chen, Yihao Zhong, Hak-Keung Lam, Junxia Han, Gaoxiang Ouyang, Xiaoli Li, and Honghai Liu. 2022. Appearance-Based Gaze Estimation for ASD Diagnosis. *IEEE Transactions on Cybernetics* (2022).

Jingjie Li, Amrita Roy Chowdhury, Kassem Fawaz, and Younghyun Kim. 2021. {Kalεido}:{Real-Time} Privacy Control for {Eye-Tracking} Systems. In *30th USENIX Security Symposium*. 1793–1810.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2020. Membership inference attacks and defenses in supervised learning via generalization gap. *ArXiv* abs/2002.12062 (2020).

Ke Liang, Youssef Chahir, Michele Molina, Charles Tijus, and François Jouen. 2013. Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*. 17–23.

Ao Liu, Lirong Xia, Andrew Duchowski, Reynold Bailey, Kenneth Holmqvist, and Eakta Jain. 2019. Differential privacy for eye-tracking data. In *ACM Symposium on Eye Tracking Research & Applications*. 1–10.

Jian Liu, Mika Juuti, Yao Lu, and Nadarajah Asokan. 2017. Oblivious neural network predictions via minionn transformations. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 619–631.

Pengrui Liu, Xiangrui Xu, and Wei Wang. 2022b. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. *Cybersecurity* 5, 1 (2022), 1–19.

Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. 2022a. {ML-Doctor}: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *31st USENIX Security Symposium (USENIX Security 22)*. 4525–4542.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 691–706.

Payman Mohassel and Peter Rindal. 2018. ABY3: A mixed protocol framework for machine learning. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 35–52.

Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 19–38.

Carlos Hitoshi Morimoto, Arnon Amir, and Myron Flickner. 2002. Detecting eye position and gaze from a single camera and 2 light sources. In *2002 International Conference on Pattern Recognition*, Vol. 4. IEEE, 314–317.

Thien Duc Nguyen, Phillip Rieger, Roberta De Viti, Huili Chen, Björn B Brandenburg, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, et al. 2022. {FLAME}: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*. 1415–1432.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth sensitivity and sampling in private data analysis. In *ACM Symposium on Theory of Computing*. 75–84.

Lucas Paletta, Martin Pszeida, Amir Dini, Silvia Russegger, Sandra Schuessler, Anna Jos, Eva Schuster, Josef Steiner, and Maria Fellner. 2020. MIRA–A Gaze-based Serious Game for Continuous Estimation of Alzheimer's Mental State. In *ACM Symposium on Eye Tracking Research and Applications*. 1–3.

Jiawei Qin, Takuru Shimoyama, Xucong Zhang, and Yusuke Sugano. 2023. Domain-Adaptive Full-Face Gaze Estimation via Novel-View-Synthesis and Feature Disentanglement. arXiv:2305.16140 [cs.CV]

Mayank Rathee, Conghao Shen, Sameer Wagh, and Raluca Ada Popa. 2023. Elsa: Secure aggregation for federated learning with malicious actors. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1961–1979.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295* (2020).

Bita Darvish Rouhani, M Sadegh Riazi, and Farinaz Koushanfar. 2018. Deepsecure: Scalable provably-secure deep learning. In *Proceedings of the 55th annual design automation conference*. 1–6.

Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. 2020. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In *29th USENIX security symposium (USENIX Security 20)*. 1291–1308.

Negar Sammaknejad, Hamidreza Pouretemad, Changiz Eslahchi, Alireza Salahirad, and Ashkan Alinejad. 2017. Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in Cognitive Psychology* 13, 3 (2017), 232.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *IEEE S&P*. 3–18.

Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 271–280.

Julian Steil and Andreas Bulling. 2015. Discovery of everyday human activities from long-term visual behaviour using topic models. In *Proceedings of the 2015 acm international joint conference on pervasive and ubiquitous computing*. 75–85.

Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling. 2019. Privacy-aware eye tracking using differential privacy. In *ACM Symposium on Eye Tracking Research & Applications*. 1–9.

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1821–1828.

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 1, 3 (2017), 1–21.

Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2016. Labelled Pupils in the Wild: A Dataset for Studying Pupil Detection in Unconstrained Environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (Charleston, South Carolina) *(ETRA '16)*. Association for Computing Machinery, New York, NY, USA, 139–142. https://doi.org/10.1145/2857491.2857520

Roberto Valenti, Nicu Sebe, and Theo Gevers. 2011. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing* 21, 2 (2011), 802–815.

Roel Vertegaal et al. 2003. Attentive user interfaces. *Commun. ACM* 46, 3 (2003), 30–33.

Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016a. A 3D Morphable Eye Region Model for Gaze Estimation. In *Proc. European Conference on Computer Vision (ECCV)*. 297–313. https://doi.org/10.1007/978-3-319-46448-0_18

Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016b. Learning an appearance-based gaze estimator from one million synthesised images. In *Proc. ACM International Symposium on Eye Tracking Research and Applications (ETRA)*. 131–138. https://doi.org/10.1145/2857491.2857492

Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F Naughton. 2016. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*. IEEE, 355–370.

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. 2019. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024* (2019).

Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3025171.3025219

Yu Yu and Jean-Marc Odobez. 2020. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7314–7324.

Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. 2018. Training Person-Specific Gaze Estimators from Interactions with Multiple Devices. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 1–12. https://doi.org/10.1145/3173574.3174198

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020b. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*. Springer, 365–381.

Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017b. Everyday eye contact detection using unsupervised gaze target discovery. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*. 193–203.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based Gaze Estimation in the Wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4511–4520. https://doi.org/10.1109/CVPR.2015.7299081

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017c. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 51–60.

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)* 41, 1 (2019), 162–175. https://doi.org/10.1109/TPAMI.2017.2778103

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020a. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 253–261.

Zhifei Zhang, Yang Song, and Hairong Qi. 2017a. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5810–5818.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610* (2020).

APPENDIX

A QUALITATIVE EVALUATION

To get the qualitative assessment of the DualView reconstruction results, a user study was conducted (N = 60 - each image was shown to 3 participants) on a randomly-picked subset of 120 images covering all cases discussed in Section 5. Participants were asked to assess each reconstructed image in terms of similarity and re-identification. An example of the survey is shown in Fig. 4 and Fig. 5.

Fig. 4. For similar ground truth (left) and reconstructed images (right) from the 3 different datasets, participants were asked the following questions:
A.1. Do both images represent the same person? (Yes/No)
A.2. On a scale from 1 to 10, how would you rate the overall similarity of both images? (Very similar to very dissimilar)

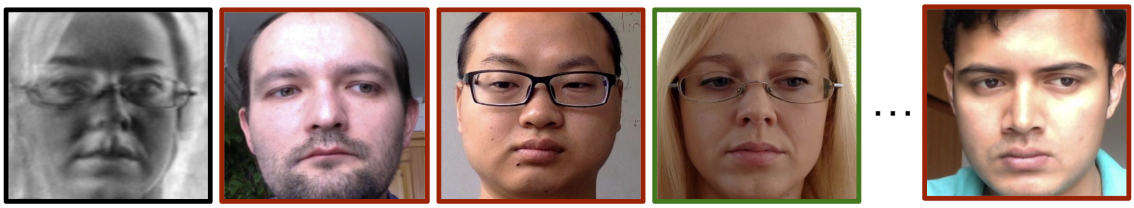

Fig. 5. For similar reconstructed images (left), participants were asked to map the full-face/eye image to the corresponding ground-truth clients.

B ADDITIONAL RESULTS

In Tab. 4 and Tab. 5, we show the DualView performance on the remaining datasets, namely MPIIGaze and NVGaze. The mean angular error (MAE) has been calculated with gaze estimators that already have MAE of 6.3, 6.2, and 0.8 for MPIIGaze, MPIIFaceGaze, and NVGaze, respectively.

	Re-identified	Visual score	PSNR	MAE	KL
Data centre	15/15	100%	max.	0	0
Adaptive FL	15/15	80%	18.3	7.8	0.198
PrivatEyes	0/15	4%	4.8	10.1	0.255
Generic MPC	0/15	5%	4.8	9.9	0.257

Table 4. DualView performance on the MPIIGaze dataset.

	Re-identified	Visual score	PSNR	MAE	KL
Data centre	32/32	100%	max.	0	0
Adaptive FL	32/32	92%	12.4	2.0	0.167
PrivatEyes	0/32	13%	4.1	6.7	0.181
Generic MPC	0/32	10%	4.0	6.6	0.180

Table 5. DualView performance on the NVGaze dataset.

In Fig. 6 and Fig. 7, we show additional examples for reconstructed face and eye images, respectively, covering different participants (i.e. appearances) and gaze directions. As previously proven by our evaluation metrics (c.f. Section 5), PrivatEyes and generic MPC leak less information in comparison to adaptive FL. The resulting faces/eyes remain nonetheless realistic due to the discriminator loss in DualView.

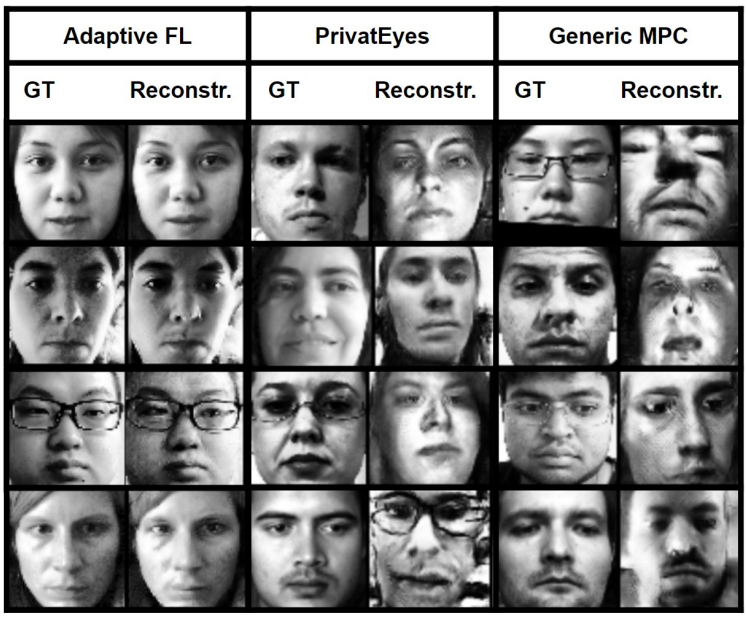

Fig. 6. Additional examples of faces reconstructed by DualView from the MPIIFaceGaze dataset.

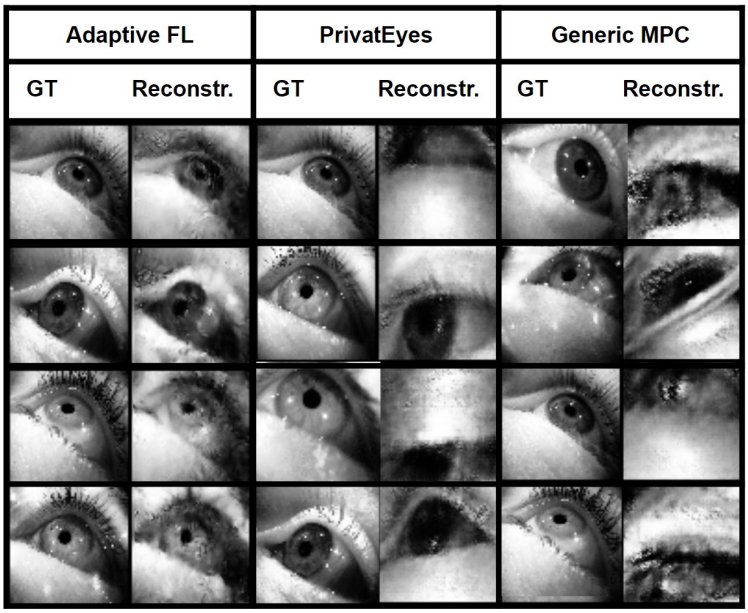

Fig. 7. Additional examples of eyes reconstructed by DualView from the NVGaze dataset.

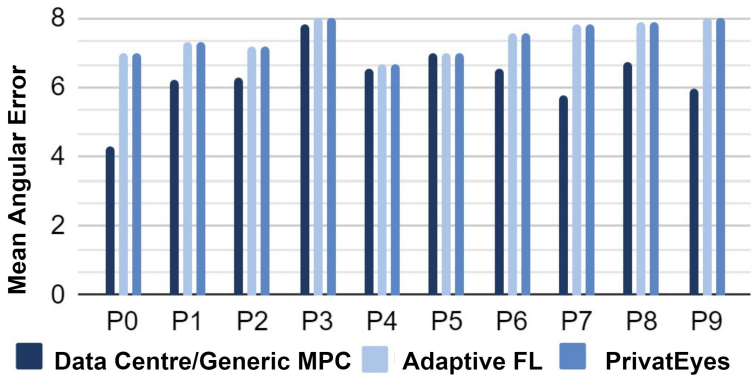

Fig. 8. Mean gaze estimation error for 10 randomly-selected participants from MPIIFaceGaze.

C CONVERGENCE EFFECT ON PRIVACY

In Fig. 9, the convergence effect can be seen from the left figure (i.e. the first training round) to the middle figures (i.e. a later training round) where some data samples no longer contribute to updating the trained model (gradients). Without aux. knowledge (middle), the prediction takes into account all training samples that previously contributed to the training of the entire model regardless of the current data of the target model. Hence, adding the previous model updates as aux. knowledge (right) enhances the reconstruction (i.e. KL-divergence).

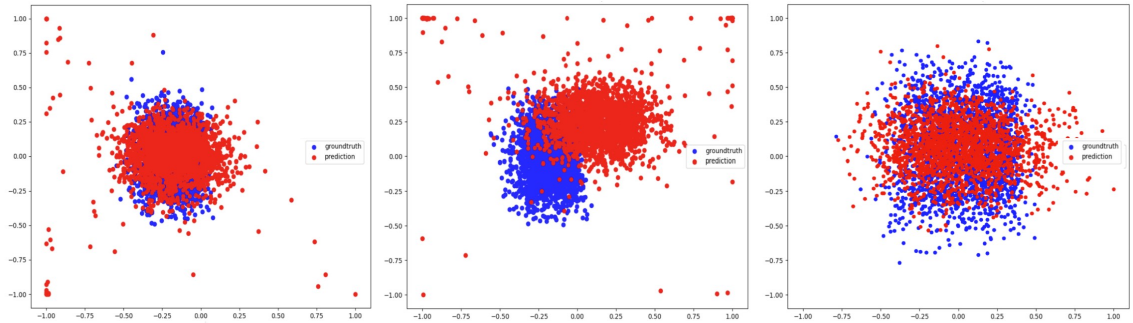

Fig. 9. Example of the ground-truth vs predicted gaze data distribution of the normalised gaze directions with KL-divergence = 0.198 (left), KL-divergence = 0.260 (middle), and KL-divergence = 0.202 (right).

D ROBUSTNESS TO OTHER DATA LEAKAGE ATTACKS

We further investigated the potential information leakage by other existing generic attacks using the ML-doctor [Liu et al. 2022a] framework which benchmarks state-of-the-art attacks on ML models.

Membership inference attacks. Membership inference attacks [Shokri et al. 2017] aim to determine if a particular data sample was used to train the model. For instance, in privacy-sensitive gaze applications (e.g. mental disorders detection [Higuchi et al. 2018; Li et al. 2022; Paletta et al. 2020]), this membership knowledge could imply the inclusion of an individual in a certain group (e.g. mental disorder). To perform the attack the adversary constructs a shadow dataset to train a shadow model on the same task (gaze estimation). In addition, the adversary trains a binary classifier (member or non-member) by querying the shadow model on training and out-of-training shadow datasets. Finally, target images are fed to the target model (the gaze estimator) and posteriors are compared to the binary classifier. For all datasets, the attack correctly identified more than 90%(first round) to 74% (last round) of the clients from individual updates (adaptive FL). For output models (PrivatEyes), for the MPIIFaceGaze dataset, membership inference was less successful as the number of rounds increased, i.e. the average accuracy dropped from 40% (first round) to 33% (last round) correctly determined samples. Moreover, for the MPIIGaze dataset, the attack's accuracy dropped from 33% (first round) to 13% (last round). Similarly, for the NVGaze dataset, the attack's accuracy dropped from 37% (first round) to 12% (last round). Therefore, PrivatEyes reduces the gaze information leakage for membership inference attacks significantly due to the inaccessibility of individual updates and the information obfuscation during the aggregation step.

Attribute inference attacks. Attribute inference attacks [Melis et al. 2019] exploit additional knowledge gained by ML models, which is not needed for the training task, to infer private attributes about the data. This attack is similar to membership inference attacks but substitutes the binary classifier with the attack model trained on the to-be-inferred attributes[4]. We opted for the simple attack of predicting the gender of the participants whose data was used to train the gaze estimation model. Consequently, the attack model is pre-trained on the UTKFace dataset [Zhang et al. 2017a]. For all datasets, the attack correctly predicted more than 95% (first round) to 65% (last round) of clients' genders from individual updates. From output models, the attack correctly predicted the gender of 5/15 (first round) to 2/15 (last round) participants for MPIIFaceGaze, 3/15 to 0/15 for MPIIGaze, and 10/32 to 1/32 for NVGaze.

[4]In DualView, the adversary reconstructs the input image, the gaze distributions, and the corresponding loss values which could later be fed to attribute classifiers.