QualitEye: Public and Privacy-preserving Gaze Data Quality Verification

MAYAR ELFARES, Institute of Information Security, Institute of Visualisation and Interactive Systems, University of Stuttgart, Germany
PASCAL REISERT, Institute of Information Security, University of Stuttgart, Germany
RALF KÜSTERS, Institute of Information Security, University of Stuttgart, Germany
ANDREAS BULLING, Institute of Visualisation and Interactive Systems, University of Stuttgart, Germany

Gaze-based applications are increasingly advancing with the availability of large datasets but ensuring data quality presents a substantial challenge when collecting data at scale. It further requires different parties to collaborate, therefore, privacy concerns arise. We propose QualitEye—the first method for verifying image-based gaze data quality. QualitEye employs a new semantic representation of eye images that contains the information required for verification while excluding irrelevant information for better domain adaptation. QualitEye covers a public setting where parties can freely exchange data and a privacy-preserving setting where parties cannot reveal their raw data nor derive gaze features/labels of others with adapted private set intersection protocols. We evaluate QualitEye on the MPIIFaceGaze and GazeCapture datasets and achieve a high verification performance (with a small overhead in runtime for privacy-preserving versions). Hence, QualitEye paves the way for new gaze analysis methods at the intersection of machine learning, human-computer interaction, and cryptography.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**.

Additional Key Words and Phrases: Gaze, quality, privacy, private set intersection

ACM Reference Format:
Mayar Elfares, Pascal Reisert, Ralf Küsters, and Andreas Bulling. 2018. QualitEye: Public and Privacy-preserving Gaze Data Quality Verification. *J. ACM* 37, 4, Article 111 (August 2018), 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Eye tracking has seen widespread adoption for numerous applications, such as for gaze-based human-computer interaction [32, 60, 61, 70, 81], for understanding the human visual system [1, 9], measuring user experience [28, 36] or for computational user modelling [2, 18, 27]. With eye tracking becoming pervasive [17] and increasingly integrated into personal devices [19, 48, 50], recent years have also seen a significant increase in the availability of large-scale gaze datasets [82, 87, 96, 98, 100]. Traditionally, these datasets have been collected in research contexts, but are now increasingly collected and shared by private individuals and commercial enterprises [34, 35].

A challenge amplified by these advances that has largely been neglected in the gaze community so far is verifying the quality of the acquired, collected, or shared gaze data. Although a few prior works [7, 8, 37, 45, 54, 69] focused on gaze

Authors' Contact Information: Mayar Elfares, mayar.elfares@vis.uni-stuttgart.de, Institute of Information Security, Institute of Visualisation and Interactive Systems, University of Stuttgart, Stuttgart, Germany; Pascal Reisert, Institute of Information Security, University of Stuttgart, Stuttgart, Germany, pascal.reisert@sec.uni-stuttgart.de; Ralf Küsters, Institute of Information Security, University of Stuttgart, Stuttgart, Germany, ralf.kuesters@sec.uni-stuttgart.de; Andreas Bulling, Institute of Visualisation and Interactive Systems, University of Stuttgart, Stuttgart, Germany, andreas.bulling@vis.uni-stuttgart.de.

Manuscript submitted to ACM

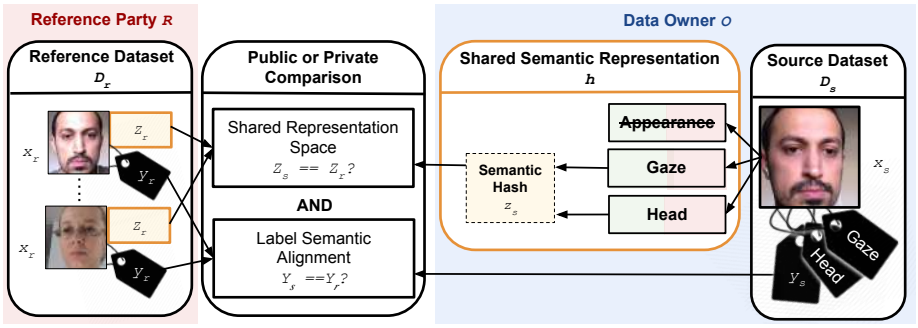

Fig. 1. To verify the image-based gaze data quality, source data owners O (blue) and the reference party R (e.g. a reliable source, red), first, disentangle the gaze direction and head pose features that correspond to the data labels, ignoring the cross-party irrelevant features (e.g. appearance) for a high domain adaptation performance instead of the raw pixel-wise data comparison. Then, features are semantically hashed for an efficient bit-wise comparison to obtain data-independent (i.e. not domain-specific), deterministic (i.e. produces the same outputs for similar semantics), and generative (i.e. learns the gaze data distribution) representations z (shown in orange). Then, they compare the hash values and corresponding labels against each other. to find the (mis)matching data samples. C.f. Figure 2 for the shared semantic representation and Figures 3 and 4 for the comparison.

data quality (e.g., evaluating the eye-tracking systems' accuracy, signal-to-noise ratio, or robustness), the verification aspect of the acquired gaze data is largely neglected, especially for image-based gaze data (e.g. features or labels). In this work, we verify the quality of the eye images and their compliance with the respective labels (e.g. gaze direction and head pose) while ignoring irrelevant cross-user features (e.g. appearance). The quality verification ensures that similar eye features correspond to similar labels, by comparing the data samples with a relatively reliable source (e.g. a publicly available dataset or a trusted party). This is particularly important as gaze data quality can be subject to inaccurate labels or inconsistent features due to technical problems with the recording setup, choice of data preprocessing methods, or calibration and systematic errors.

Formally, as shown in Figure 1, let the *source dataset* be $D_s = \{(x_s^{(i)}, y_s^{(i)})\}_{i=1}^{N_s}$, where $x_s^{(i)} \in \mathcal{X}$ denotes an image (e.g., eye or face image) and $y_s^{(i)} \in \mathcal{Y}$ denotes the associated gaze-related label (e.g., gaze direction or head pose). The source dataset corresponds to the dataset whose gaze data quality is to be verified. Let the *reference dataset* be $D_r = \{(x_r^{(j)}, y_r^{(j)})\}_{j=1}^{N_r}$, which contains image samples and the same label semantics as the source dataset (e.g., head pose and gaze direction), is assumed to have trusted or known data quality, and is used as a baseline for comparison. Therefore, we present QualitEye, a computational framework for gaze data quality verification. We first propose a new shared semantic representation function $h : \mathcal{X} \to \mathcal{Z}$, which maps raw images to a latent semantic space $\mathcal{Z}$ that preserves gaze-relevant information while discarding subject-specific or appearance-related factors. The corresponding semantic representations are $z_s^{(i)} = h(x_s^{(i)}), \quad z_r^{(j)} = h(x_r^{(j)})$. The source dataset D_s and reference dataset D_r are said to be *compatible* if the following conditions hold:

(1) **Shared Representation Space: The representations of both datasets lie in the same semantic space, i.e., $z_s^{(i)}, z_r^{(j)} \in \mathcal{Z} \quad \forall i, j$, and $\mathcal{Z}$ encodes gaze-relevant semantics consistently across datasets.**
(2) **Label Semantic Alignment:** The label spaces $\mathcal{Y}_s$ and $\mathcal{Y}_r$ are identical or there exists a mapping $\phi : \mathcal{Y}_s \to \mathcal{Y}_r$ such that $\phi(y_s^{(i)})$ is semantically comparable to $y_r^{(j)}$ (e.g., consistent gaze coordinate systems).

The need for quality verification is further amplified when raw gaze data cannot be shared due to privacy constraints, necessitating privacy-preserving processing [34, 35, 59]. Accordingly, QualitEye supports:

(1) **Public verification** for within-dataset consistency and comparison to public references.
(2) **Privacy-preserving verification** for remote comparison against private reference datasets without data disclosure. In this case, QualitEyeensures that: (i) Raw eye images and sensitive features are never directly exchanged during verification, (ii) Only essential semantic information for quality assessment is used, which is disentangled from irrelevant or privacy-sensitive attributes, (iii) Cryptographic protocols prevent inference of another party's underlying data beyond what is needed for quality comparison.

To evaluate QualitEye, we present appearance-based gaze estimation as our guiding example since it is the basic building block of gaze-based applications and has well-established publicly available datasets for evaluation. Nonetheless, QualitEye does not restrict how the dataset should be processed after the quality verification; hence, QualitEye is domain-, task-, and model-agnostic. We thereby validate our method through extensive experiments on the well-established full-face appearance-based gaze estimation datasets, MPIIFaceGaze [100] and GazeCapture [64], and achieved a gaze quality metric (Matthew Correlation Coefficient [71][1]) of 0.92 and 0.94, respectively, with a negligible overhead in runtime for the privacy-preserving setups compared to other baselines.

QualitEye, therefore, solves a fundamental concern– gaze data quality–in many eye-tracking applications including: (i) **Data collection and cleaning**, enabling validation of newly acquired gaze data against trusted sources [12], (ii) **Auto-labelling**, where labels are inferred from limited local annotations or reliable external datasets [79], (iii) **Remote learning and analytics**, e.g., federated learning [34, 35], where verifying the quality of locally held private data is critical, (iv) **Try-before-you-buy services** [83], allowing users to assess gaze data or model predictions prior to acquisition, and (v) **Predictive benchmarks and leaderboards** (e.g., Kaggle [57]), enabling private test-set verification by comparing labels without model disclosure. In summary, this work makes the following contributions[2]:

- QualitEye is the first work to investigate the problem of image-based gaze data quality verification.
- Instead of the raw pixel-wise data comparison, we propose a new generic hashed representation learning model that disentangles the gaze-specific features and ignores the cross-users' irrelevant features (e.g. appearance).
- We propose methods for public and privacy-preserving gaze data quality verification. For the latter, we extend existing protocols with semantic similarities and label matching to handle the different privacy requirements.

2 Related Work

Gaze data quality. The lack of large-scale, diverse datasets remains a key limitation in eye-tracking research, hindering the study of variability across users, tasks, and settings [34, 43]. Despite advances in gaze data acquisition and processing, quality standards necessary for high-performing models are often neglected [46], primarily due to (i) the time-intensive collection and labeling of large datasets, (ii) data owners' reluctance to share private eye data, and (iii) the difficulty for single entities to gather diverse data at scale [85]. Recent work has emphasised standardised gaze data quality reporting, including metadata such as sampling rate, tracked eye(s), filter settings, recording duration, and display resolution,

[1]The Matthews Correlation Coefficient (MCC) is a metric used to assess the quality of binary classifications (i.e. match or mismatch) and it is particularly useful in cases where there is a class imbalance (e.g., when one class is much more frequent than the other), making it more reliable than metrics like accuracy in these situations (cf. Section 6).

[2]Note that, QualitEye does not assume that any single reference dataset is universally representative of all possible gaze data distributions; instead, it explicitly acknowledges and mitigates reference–source mismatch as a fundamental challenge. The framework formulates quality assessment as a problem of *relative quality verification* rather than absolute performance prediction, thereby conditioning its conclusions on the characteristics of the chosen reference data (c.f. Section 6). This design choice reflects a well-known limitation in computer vision, namely that universal ground truth distributions are generally intractable to obtain, particularly for eye and gaze imagery, where data are inherently non-independent and non-identically distributed due to variations in subject appearance, head pose, lighting conditions, capture devices, and annotation protocols [34]. Furthermore, QualitEye is explicitly designed to support the use of multiple reference datasets, enabling cross-reference validation and reducing dependence on any single dataset's coverage or bias. Consistent verification outcomes across diverse references strengthen confidence in the assessed quality, while discrepancies serve as indicators of domain mismatch or insufficient reference diversity.

to ensure accessible, reusable, and reproducible datasets [54]. While a few studies have assessed eye-tracking system accuracy and signal-to-noise ratio [7, 8, 37, 69], we propose automatic quality verification for appearance-based gaze data without manual reporting, applicable both publicly and in a privacy-preserving manner.

Unsupervised gaze representation learning. Supervised representation learning has been used to extract task-specific eye features, such as gaze estimation [96, 98] or eye contact detection [97], but it relies on carefully labelled data and does not generalise well across tasks. In contrast, self-supervised learning (SSL) enables models to learn underlying image features directly from the data without task-specific labels, capturing subtle, person-specific variations in appearance-based gaze methods [34, 42]. Early SSL approaches used joint embedding architectures (e.g., siamese networks) [68, 88, 95], but these often collapsed, producing identical embeddings. Contrastive methods address this by learning from positive and negative pairs, yet enumerating all possible pairs is intractable, leading to bias in hand-selected examples [56]. Non-contrastive approaches [23, 25, 91] theoretically require optimising latent capacity, which is often intractable. Here, we propose a VAE-based approach for gaze representation learning that is generative, non-contrastive, and leverages a "fuzzy" latent variable, with a neural network providing amortised optimisation across gaze data points.

Privacy-sensitive gaze data. Gaze data can contain personal identifiers [22, 80, 93], sensitive attributes [47, 49, 84, 92], or business-related information (e.g., devices or participant details [16]) that cannot be shared. Privacy threats in eye tracking have been largely underexplored, partly because traditional cryptographic and privacy-preserving techniques (PPTs) [6] were considered too computationally expensive. Recent advances, however, make approaches such as federated learning (FL) [34], differential privacy (DP) [15, 66, 67, 85], and secure multi-party computation (MPC) [35] practical for gaze-based tasks. QualitEye is the first work to address gaze data quality verification while explicitly incorporating privacy considerations.

Privacy-preserving techniques. Secure data comparison techniques [3, 13, 30, 52, 65, 72–74, 89] enable similarity detection between datasets while preserving privacy. Traditional methods use cryptographic tools such as homomorphic encryption (HE) [39], secure multiparty computation (MPC) [41, 94], and oblivious transfer [78]. More recent approaches employ private set intersection (PSI) protocols [11, 20, 21, 29], which reveal only the intersection (or its size), operate on a single party's input, and exchange cryptographic hashes, reducing computation and communication overhead. These techniques are used in biometric verification [13, 26, 38, 89], data linkage [3, 33, 73, 74], fraud detection [30, 52, 72, 101], and genomics [5, 62, 65], typically with tolerance for errors from formatting, typos, or biological variations. In contrast, gaze-based applications are highly sensitive to subtle variations in gaze and head pose, requiring a novel privacy-preserving comparison that focuses only on task-relevant features, excludes appearance, minimizes data transfer, and reduces the risk of information leakage.

3 QualitEye

Given our setup and formal definitions in Section 1, in this paper, we present a most common leading gaze example - appearance-based gaze estimation with eye images and their corresponding gaze direction and head pose labels[3]. We assume a horizontal data distribution where each party's dataset includes different data samples, e.g. data of different

[3]Our approach is not limited to gaze estimation and can be extended to other gaze-based applications. Nonetheless, we choose gaze estimation as our guiding example since it is the basic building block of gaze-based applications and has well-established publicly available datasets for evaluation.

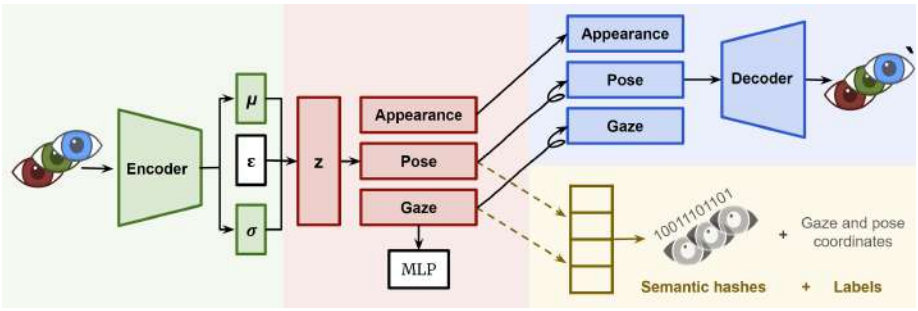

Fig. 2. To obtain the hashed semantic representations used for label-based data comparison, *1- green:* eye images are encoded into a latent vector z. *2- red:* Then, the gaze and head pose information are disentangled from the appearance via some transformations and a multi-layer perception (MLP). *3- blue:* During training, the appearance, gaze, and pose are passed as a latent vector to a decoder that reconstructs the transformed eye images. *4- yellow:* Finally, the gaze and head pose latent codes are hashed and are passed as inputs to the PSI protocol along with the corresponding labels.

participants and different numbers of samples. The data owner wishes to verify the quality of the gaze dataset with respect to the reference dataset to overcome the data collection and labelling problems mentioned above.

We define the gaze dataset quality as the number of data samples (the image features and the corresponding labels) that comply with the reference dataset, i.e. the cardinality of the intersection set. More formally, the set of mismatching data samples would be $\{(x_i^O, y_i^O) | \exists (x_i^R, y_i^R) : x_i^O = x_i^R \wedge y_i^O \neq y_i^R\}$.

Overview. In both the public and private evaluation setups, as shown in Figure 1, QualitEye operates by first mapping samples from the source and reference datasets into a shared semantic representation space using the representation function described in Section 3.1. In the public setup (QualitEye-V_0), semantic representations of the reference dataset are directly available, allowing quality verification to be performed through explicit similarity and matching operations. In the private setup, where raw data and representations cannot be shared due to privacy or proprietary constraints, QualitEye supports multiple alternative protocols (QualitEye-V_1 to V_4) that provide different trade-offs between privacy guarantees, computational requirements, and verification fidelity, allowing QualitEye to adapt to a wide range of real-world deployment scenarios while maintaining its core objective of relative data quality verification.

3.1 Semantic gaze representations

Since the direct pixel-wise comparisons of the gaze images only capture exact matches. To obtain more meaningful results, we increase the efficiency of comparisons by first encoding the images in a semantic representation. Our goal is to learn the semantic representations of the data samples that reflect the similarity of the gaze-based information (e.g. gaze direction and head pose) while ignoring other cross-party irrelevant features (e.g. appearance). In addition, this representation should be deterministic (i.e. produces the same outputs for similar semantics at different parties), generative (i.e. learns the gaze data distribution), and domain-agnostic (i.e. not dataset-specific) to be able to generalize well to different (unseen) datasets at different parties.

Variational auto-encoder (VAE). As shown in Figure 2, our VAE is composed of four main components:

- **Top-down encoder** E_ϕ that maps the input x to a latent z corresponding to a variational distribution and outputs the parameters of the distribution (e.g. mean μ and variance σ). In our case, we use a multivariate Gaussian distribution

$N(x|\mu, \sigma)$ due to the nature of RGB eye images and to estimate the average of the gaze data distribution (with likelihood $p_\theta(x|z)$) according to the *central limit theorem* for a better cross-domain adaptation. Then the prior $p_\theta(x)$ is a mixture of Gaussian distributions, and the posterior distribution $p_\theta(z|x)$ can be approximated to $q_\phi(z|x)$ (i.e. the amortized optimization).

- **Bottom-up decoder** D_θ maps the latent to the input space. Hence, our problem is to compute the conditional likelihood distribution $p_\theta(x|z)$ by the probabilistic decoder and the approximated posterior distribution $q_\phi(z|x)$ by the probabilistic encoder.
- **Feature disentanglement:** For a better domain adaptation, our model explicitly learns to disentangle the gaze direction and head pose representations as equivalent input rotations. This is achieved by training the model in a person-independent manner [68] and splitting z into three sub-vectors, (i) gaze direction, (ii) head pose, and (iii) appearance (more specifically, all other information found in the image), similar to [76], and rotating the latent sub-vectors using rotation matrices to the frontal angle and then to certain yaw and pitch angles. That is, for the same person, the model learns to transform the gaze direction and head pose of one image into the other by multiplying the sub-codes by a rotation matrix and optimizing a pixel-wise $L1$ loss function over the entire encoder-decoder and an MLP regression for the gaze sub-code.

  The aim of disentangling gaze direction and head pose from the rest of the image is to (i) only compare the information that is relevant to the ground-truth labels and (ii) to minimize the amount of information exchanged in the cross-party setup (i.e. data minimization) for better privacy, runtime, and communication.
- **Hashing:** Once the latent vector is disentangled, the gaze direction and head pose sub-codes are hashed with a locality-sensitive hash function. Locality-sensitive hashing (LSH) is a fuzzy hashing technique that maps similar inputs to the same hash value with a certain probability. Such hashes (i) reduce the dimensionality of the semantic representations, which likelyimproves efficiency (e.g. computational runtime and communication), (ii) allow efficient comparisons of the bit-wise representations (i.e. comparing the hashed values in the hamming space instead of the latent space), (iii) are data-independent (i.e. not domain-specific to generate the same hash for different inputs at different parties), and most importantly, (iv) produce identical hashes for images with similar features (i.e. to tolerate small systematic errors in data acquisition commonly found in eye tracking data).

VAEs are specifically interesting for gaze-based data because of (i) the inherent mutual information in appearance-based eye data (higher mutual information yields better disentanglement [24]), (ii) the independence between the latent variables (e.g. gaze direction and head pose) encourages interpretability yielding better semantics [24, 44], and (iii) they are more generalisable[4] [4].

4 Comparing Hashes in the Public Setting (QualitEye-V_0)

As the raw pixel-wise eye image comparison cannot be used for quality verification given the cross-user variations in appearance, the hashed disentangled gaze direction and head pose representation can be used instead. Hence, a gaze data owner O can verify the within-dataset consistency by checking that similar representations of gaze direction or head pose have similar respective labels. Additionally, in case of systematic or calibration errors (e.g. resulting from changes in user position during data collection), O can compare the collected dataset against the publicly available datasets (as the reference) to either check for possible errors (i.e. correcting the non-compliant data samples) or for auto-labelling the dataset.

[4]The generalisation capability makes VAEs more robust against adversarial attacks [4]. This increases privacy but remains out of the scope of this paper.

However, in other scenarios, the datasets might not be available at one party, therefore, O might need to verify the dataset quality against a different dataset owned by another reference party R. A (public) solution would be for one party (e.g. R) to send their data as hashed disentangled representations along with the labels to the other party (e.g. O) for comparison.

5 Comparing Hashes in the Private Setting (QualitEye-V_1 to V_4)

The (public) solution mentioned above does not guarantee privacy as O can perform a dictionary attack (i.e. tries all possible hashes of the input space) and recover the plain representations, especially when the dictionary has a computationally reasonable size, e.g. in the case of gaze estimation. Note that, even if the plain representations do not include the appearance and the raw images cannot be reconstructed [35] (via our data minimization step of disentanglement), O can still deduce information beyond the (mis)matching samples such as the number of samples in R's dataset, the semantic meaning of all other samples, the plaintext labels, the kind of error (if any)... etc. Therefore, a better solution is to use a cryptographic solution with formal provable guarantees – private set intersection (PSI), where both parties interactively compute the intersection (i.e. one party cannot compute the intersection without the other party's help, e.g. mitigating dictionary attacks). However, since PSI usually come with a computational overhead or a drop in utility, we, therefore, present several versions of QualitEye with different tradeoffs between efficiency (i.e. runtime and communication), privacy, and utility.

We assume a semi-honest (a.k.a. honest-but-curious) security model, i.e. parties will not deviate from the defined protocol; however, they might try to learn possible information from the legitimately received messages. Note that, in this work, an adversary refers to the main parties R and O.

Private set intersection (PSI) Preliminaries. PSI is a secure multiparty computation (MPC) protocol that allows two (or more) parties to compare elements in their sets by computing the intersection without revealing any information beyond this intersection. Recently, PSI constructs [11, 20, 21, 29] included different adversarial models, efficiency tradeoffs, and security guarantees. In this paper, we focus on the semi-honest (a.k.a. honest-but-curious) adversarial model [40]. This assumes that different data owners have a mutual interest in working together and improving the quality of their own datasets. We therefore assume that they stick to the rules and follow a previously agreed protocol. However, they nevertheless are happy to gain any information leaked by the protocol. Such data owners are called honest-but-curious. QualitEye uses different cryptographic primitives to ensure that a curious data owner gains only minimal knowledge about other parties' data, e.g. oblivious transfer. PSI constructions can include different cryptographic primitives. QualitEye relies on the following primitives:

- **Key agreement protocols:** In cryptography, key agreement protocols allow two or more parties to agree on a cryptographic key. Among these protocols, the Diffie–Hellman protocol [31] is one of the earliest practical protocols. More specifically, we base our privacy-preserving constructions of QualitEye$_{v1,v12,v3}$ on a Diffie-Hellman-based PSI protocol where, as shown in Figure 3, the two parties O and R hash all their data samples x_i^j [5], and raise the hashed values to their private keys. Then, these values are exchanged and compared by O. A match means that both x_i^O and

[5]Parties get a cryptographic hash of the semantic hash values using their private keys K_O and K_R. Samples are hashed to a primitive root modulo p where p is a large prime number that parties agree on, i.e. $H(x_i^j)^{K_j} \mod p$. In modular arithmetic, a number g is a primitive root modulo n if every number a coprime to n is congruent to a power of g modulo n. That is, g is a primitive root modulo n if for every integer a coprime to n, there exists some integer k for which $g^k \equiv a(modn)$. In other words, every invertible number is of the form g^k for some integer k.

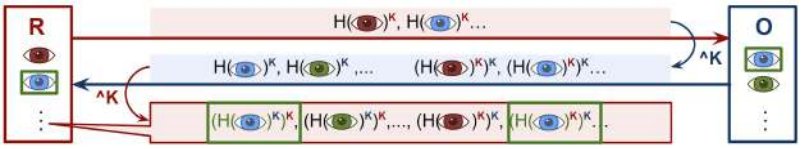

Fig. 3. In **QualitEye-V**$_1$, both parties O and R exchange their elements as hashed values raised to their private keys. Then, they raise the other party's received values again to their private keys. R then start the comparison to find the matching inputs. Note that, we further send the eye labels (e.g. gaze direction and head pose) as an additional encrypted payload to each element. In **QualitEye-V**$_2$, O shuffles the second message to only reveal the cardinality of the intersection. In **QualitEye-V**$_3$, R includes the first message, e.g. as a package, before the start of the protocol. In all versions, only the private keys are secret, and all other values are sent in the clear. Security is still guaranteed due to the hardness of the 'discrete logarithm problem', i.e. it is hard to infer the private keys [31].

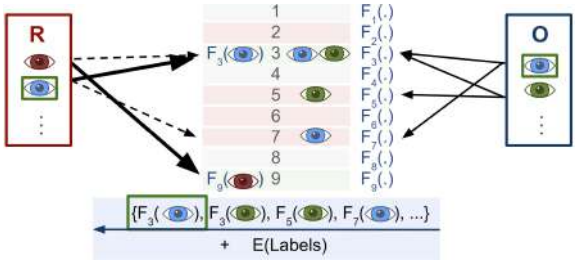

Fig. 4. In **QualitEye-V**$_4$, each party computes two values per element (i.e. the PRF output of the hashed semantic representations). R only selects one value per element (via cuckoo hashing). O send his values (along with the encrypted blinded labels) to R. R sends back all received information along with her encrypted blinded labels to O, who makes the comparison and finds the (mis)matching elements.

x_i^R are similar. A mismatch reveals no information about the inputs[6]. This way, dictionary attacks are not possible since R is committed to certain values in the first message, and the intersection requires the interaction between both parties to get both private keys (i.e. O will not remain available for R to try all different dictionary entries). Note that an eavesdropper can intercept the communication but does not have access to the private keys; therefore, the eavesdropper cannot infer the intersection [51].

- **Oblivious transfer (extension):** Oblivious transfer (OT) [78] and Oblivious transfer extension (OTe) [10, 53] are cryptographic protocols where one party sends one of many pieces of information to a receiver, but remains oblivious as to what exact piece has been sent.
- **Oblivious pseudorandom functions (OPRF):** OPRF is a cryptographic protocol where two parties jointly compute a pseudorandom function (PRF), i.e. a function which emulates a random oracle. Similar to OT, the sender does not learn any information about the other party's input, i.e. the sender is oblivious to what exact values has been sent. More specifically, we base our PSI construction of QualitEye$_{v4}$ on an OPRF-based construction [63]. As shown in Figure 4, both parties hash their inputs to two hash values. R only selects one of the two hash values (using cuckoo hashing [75]). O sends all his inputs (as two PRF outputs per input corresponding to the two hash values) to R. R compares these outputs to his to compute the intersection. This way, R only learns the matching elements while all other elements in O's dataset look random to him, and O learns nothing about R's inputs.

[6]As proved by Diffie-Hellman [31], raising $H(x_i)$ to an exponent (i.e. the secret key) makes it indistinguishable from random (i.e. the protocol hides the inputs when there is no match), even if the exponent is used elsewhere (i.e. it is safe to reuse $H(x_i)^{K_j}$ as in QualitEye$_{v3}$).

Adaptation of existing PSI protocols. PSI protocols typically operate on input messages to find matches between two (or more) sets. In our case, three additional challenges arise:

(1) The comparison of the raw eye images only reflects the exact similarity. We solve this issue by comparing the hashed semantic representations of the inputs in each dataset (generated in subsection 3.1).
(2) In addition to finding matching elements in the parties' sets, we further need to check the compliance of the labels. In the following, we solve this problem by extending the protocols with additional payloads.
(3) The accuracy of some labels (e.g. gaze direction) in the reference gaze-based data can include some error, currently >3 degrees for gaze direction [90]. Therefore, we allow some error tolerance for the semantic representations and their respective payloads.

QualitEye and the Different Privacy Tradeoffs. QualitEye adapts privacy-preserving gaze data verification according to different efficiency tradeoffs. The different versions cover: (i) Dataset sizes: hundreds vs thousands of samples, (ii) difference in datasets size: symmetric (i.e. when both parties have the same amount of data) vs asymmetric (i.e. when one party has a relatively larger dataset) dataset distributions, (iii) resulting information: the intersection set vs its cardinality, (iv) receiver of this information: one vs both parties.

- **QualitEye-V$_1$** can be used when both parties have relatively small datasets (i.e. tens to hundreds of samples) and would like to know the exact (mis)matching samples. More specifically, QualitEye-V$_1$ is based on the Diffie-Hellman key exchange protocol, as shown in Figure 3.
- **QualitEye-V$_2$** can be used when both parties have relatively small datasets (i.e. tens to hundreds of samples) but are only interested in knowing the cardinality of the intersection (i.e. the size of the (mis)matching samples and not the exact samples). Similarly, QualitEye-V$_1$ can be extended to only reveal the cardinality of the intersection (i.e. the intersection size and not the exact samples). In QualitEye-V$_1$, R sends $M2b$ (Figure 3) in the same order sent by O so that O can find the corresponding elements in her set. In QualitEye-V$_2$, R shuffles those elements before sending them to O. Hence, O can only count the number of matching elements but cannot map them to the raw elements (i.e. elements appear uniform).
- **QualitEye-V$_3$** can be used when one party has a relatively small dataset while the reference party owns a larger dataset (i.e. thousands to billions of samples). In this case, $M1$ in QualitEye-V$_1$ and QualitEye-V$_2$ (Figure 3) can be sent in advance (e.g. offline as a built-in package in an eye-tracking software or as a publicly-published data) as it does not depend on the other party's input. Note that this does not break privacy and can be shared with multiple parties (or protocol instances) [7].
- **QualitEye-V$_4$** can be used when one or both parties own large datasets where the previous versions are highly inefficient (c.f. 6). As shown in Figure 4, QualitEye-V$_4$ is based on a OPRF-based PSI protocol [63]. QualitEye-V$_4$ can be an alternative to QualitEye-V$_3$ when changes to the data of the reference party are frequently made.

For all versions, the final information can be revealed to (i) one party, depending on which party starts the protocol or (ii) both parties with an additional communication step containing the resulting information (i.e. the intersection set or its cardinality). We further propose the following adaptations: In addition to the hashed semantic representations of the gaze direction and head pose, both parties input the corresponding labels encrypted as Elgamal ciphertext (an asymmetric key encryption based on the Diffie-Hellman key exchange) under their private keys. For instance, in Figure 4, O sends the PRF outputs along with the corresponding encrypted (under O's key) labels. If R finds a matching

[7] Kales et al. [58] further proposed an efficient encoding mechanism that could be used with $M1$ to enhance efficiency.

PRF representation in his dataset, he encrypts his labels and sends both encrypted labels to O. If there is no match, R re-encrypts O's label. Then, O decrypts both labels to find the mismatching inputs. Note that O cannot distinguish between the cases where a matching representation does not exist in R's dataset and a matching representation exists with a matching label; he only learns the mis-matched labels (i.e. the non-compliant samples). On the other hand, O only learns the cardinality of the matching set. Additionally, parties do not learn the labels in the clear as labels are further blinded following [12].

Furthermore, to accommodate the (unavoidable) systematic errors in the labels (e.g. 3 degrees for gaze direction) in the hashing step, similar to the gaze representations, we adjust the probability that different latent codes are mapped to the same hash values and drop the least significant bits in the labels accordingly. The parties agree on the exact values that can be adjusted according to the data collection setups (controlled vs in-the-wild, remote vs near-eye cameras... etc) and the corresponding established error values (e.g. eye-tracker drift or calibration errors).

6 Experiments

6.1 Datasets

To evaluate QualitEye, we use different appearance-based gaze estimation datasets covering different conditions: appearances (genders, ethnicities, glasses, and make-up), illumination (indoor and outdoor), and gaze direction and head pose distributions. Mainly, experiments were conducted on the full-face MPIIFaceGaze [99] and GazeCapture [64] datasets. The MPIIFaceGaze [99] dataset contains ~200 thousand full-face images collected in the wild from 15 participants. The GazeCapture [64] dataset contains ~2.5 million frames of ~1.5 thousand participants.

6.2 Results and Implementation Details

We train our VAE with an enhanced ResNet [90] backbone on the training set (80% of GazeCapture) in a person-specific fashion since the inter-subject anatomical differences are known to affect the performance of gaze-based tasks [64, 99]. We then test our model on the remaining 20% of GazeCapture and the full MPIIFaceGaze dataset.

6.2.1 ***Training the VAE.*** We use the well-established loss function [76] adapted to our disentanglement criteria: $L_{\text{full}} = \lambda_{recon}L_{\text{recon}} + \lambda_{EC}L_{\text{EC}} + \lambda_{gaze}L_{\text{gaze}} + \lambda_{KL}L_{\text{KL}}$ where L_{recon} is the reconstruction loss that guides the encoding-decoding process pixel-wise, L_{EC} is the embedding consistency loss that ensures the embedding of the same appearance into the same features even with different (disentangled) gaze direction and head pose, L_{gaze} is the gaze direction loss between the estimated gaze of the MLP and the true gaze direction, and L_{KL} is the VAE Kullback-Leibler divergence loss that regularizes the model by approximating the prior distribution via the encoded distribution and penalising deviations from the model. For the coefficients, we use $\lambda_{recon} = 2$, $\lambda_{EC} = 1$, $\lambda_{gaze} = 0.1$, and $\lambda_{KL} = 1$ with a batch size of 128 and a learning rate of $5 \cdot 10^{-7}$. This yielded a reconstruction loss of 0.3859, a gaze angular error of 5.0543°, and a KL-divergence loss of 12.9531. The EC_{gaze} is 5.8974 and the EC_{pose} is 10.4678.

6.2.2 ***Feature Disentanglement.*** As shown in Figure 2, to disentangle the appearance, gaze direction, and head pose features, the encoder processes the input image into three distinct latent codes of size 64, 2, and 16, respectively. Qualitatively, as shown in Figure 5, we intentionally neglect the appearance code for (i) data minimisation, i.e. sharing less information to enhance privacy and (ii) for transferability across different subjects. The head pose code captures rotation and orientation (i.e. pitch and yaw angles) and excludes gaze direction to ensure that the gaze features remain consistent even if head orientation changes. This is achieved by a transformation (i.e. a rotation matrix), proposed

Fig. 5. A qualitative example of the correct predictions (green) and incorrect predictions (red) of the gaze direction verification. Each pair represents different aspects of appearance, e.g. different subjects, head poses, lighting conditions, glasses, makeup, genders, and race. Therefore, QualitEye can successfully disentangle the appearance code, removing its effect from the overall method.

Fig. 6. A qualitative example of the output of the decoder after the disentanglement and rotation of gaze direction (top, with a fixed head pose and rotating gaze direction) and head pose (bottom, with a fixed gaze direction and rotating head pose).

Table 1. Within-dataset performance for gaze verification averaged over different participant-based data splits.

	Dataset	TP	TN	FP	FN	**MCC**
Gaze Direction	MPIIFaceGaze	0.9628	0.9967	0.0033	0.0372	0.9220
	GazeCapture	0.9808	0.9966	0.0034	0.0192	0.9402
Head Pose	MPIIFaceGaze	0.8381	0.9062	0.0938	0.1619	0.746
	GazeCapture	0.9564	0.9289	0.0711	0.0436	0.8856

by Park et al. [76], that maps each gaze representation relative to a canonical (or frontal) head position, leading to successfully disentangling the gaze direction and head pose as shown in Figure 6.

Quantitatively, correctly matching samples between source and reference datasets provides a direct quantitative indicator of successful feature disentanglement. We report the quality metric as a Matthews correlation coefficient (MCC) [71], to account for true positives (TP, the correct compliant matching samples), true negatives (TN, the correct non-compliant mismatching samples), false positives, and false negatives (FP and FN, the incorrect predictions). MCC is particularly interesting as it can be used for classes of different sizes [14], i.e. both symmetric and asymmetric scenarios. A coefficient of +1 indicates a perfect match, 0 represents a random prediction, and −1 is a total disagreement between the predicted match and the true match.

We run experiments on the within- and cross-participant for the same domain, e.g. MPIIFaceGaze in Figure 7. We further report the average MCC with varying participant-based data splits to handle the unbalanced data distribution across data sources in Table 1, and cross-datasets in Table 2 quality verification. Note that, we use the available dataset's labels as our ground-truth which is subject to error since creating labels is a complex task in the first place given the eye-head interplay, eye registration error, occlusions, appearance biases... etc. Hence, we adapt our hashing step to compensate for such (unavoidable) errors in the data where the dimensionality of the target projected space is reduced to 80 bits with a collision probability of 0.05[8].

Hence, the disentanglement is not achieved implicitly (e.g., via adversarial losses), but rather by explicitly enforcing how specific latent variables must behave under known physical transformations (i.e., geometric rotations).

[8]Values are calculated according to the current SOTA remote gaze estimation models [90]. For instance, the normal gaze range is [-45,45] with a few outliers and an error of 3 and 5 degrees for gaze direction and head pose, respectively.

Table 2. Cross-domain performance for gaze quality verification on a TITAN X 12G GPU

Dataset	TP	TN	FP	FN	**MCC**	Runtime
MPIIFaceGaze/ GazeCapture	0.9740	0.9966	0.0360	0.0260	0.9331	152*s*

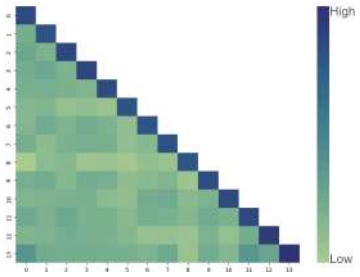

Fig. 7. Cross-participant performance as normalised MCC on the MPIIFaceGaze dataset

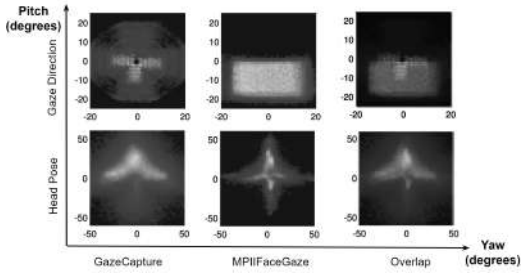

Fig. 8. The pitch and yaw distribution of gaze direction and head pose on the GazeCapture (reference R) dataset, the MPIIFaceGaze (owner O) dataset, and the corresponding distribution overlap. QualitEye captures the overlapping samples and misses the samples in O that do not have a matching sample in R.

6.2.3 ***Datasets' Discrepencies***. Although the GazeCapture and MPIIFaceGaze datasets are widely used in gaze research, they differ significantly in terms of data collection methodology, device settings, environmental conditions, and participant demographics. As shown in Table 1 and Figure 7, our method was able to successfully disentangle gaze direction and head poses within the same dataset, regardless of the appearance differences such as the different participants (1400 vs 15), demographics (diverse range of age vs university students), and background and lightning conditions (outdoor vs indoor) in GazeCapture and MPIIFaceGaze, respectively. The performance slightly degrades for head pose since the GazeCapture dataset was self-collected by participants using mobile devices (iOS phones and tablets), leading to a high variation in head pose due to uncontrolled conditions. It was also annotated with gaze points that are less precise due to mobile device limitations. Meanwhile, the MPIIFaceGaze dataset was recorded using laptops with a front-facing camera in indoor setups. This led to limited head pose variation and precise annotations due to controlled lab settings. Hence, the GazeCapture dataset is larger, more variable, and more general, and the same domain was used to train the VAE, hence the better performance.

Nonetheless, QualitEye mainly relies on a reference dataset for comparison. Therefore, as shown in Table 2 and Figure 8, QualitEye also succeeds at comparing different domains (i.e. datasets) but mainly fails in cases where a sample in one dataset does not have a matching sample in the other dataset when comparing ground-truth labels (i.e. outliers). In practice, carefully selecting a reference dataset that meets all requirements is recommended.

6.2.4 ***Privacy-preserving gaze data quality verification***. The private cross-party gaze data verification setups maintain the same performance as the local public one in terms of data quality metrics. We use a computational security parameter $k = 128$ and a statistical security parameter of $\sigma = 40$ following [63].[9] However, privacy comes with a computational overhead in runtime and communication. Since the magnitude of this overhead depends on multiple

[9] $k = 128$, $\sigma = 40$ is a standard security parameter choice. Other values are possible. Generally, lower values might reduce the runtime of the cryptographic parts but can be more vulnerable to attacks. The security parameter choice does not affect accuracy.

factors, e.g., dataset size (c.f. Section 5), our proposed versions account for the practical runtime while maintaining the desired privacy guarantees, as shown in Table 3.

Table 3. LAN online runtime in ms as overhead with respect to the public verification runtime (QualitEye_{v0}) on a TITAN X 12G GPU for different dataset sizes: 2^8 (small-hundreds), 2^{12} (small-thousands), 2^{18} (~ MPIIFaceGaze), 2^{21} (~ GazeCapture), and 2^{24} (larger datasets). The table shows the symmetric scenarios, however different asymmetric splits are also possible (c.f. 2).

Protocol/Size	2^8	2^{12}	2^{18}	2^{21}	2^{24}
QualitEye_{v0}	128	2,048	131,072	1,048,576	8,388,608
QualitEye_{v1}	+ ~ 100	+ ~ 1,600	+ ~ 102,400	+ ~ 819,200	+ ~ 6,553,600
QualitEye_{v2}	+ ~ 101	+ ~ 1,601	+ ~ 102,402	+ ~ 819,202	+ ~ 6,553,602
QualitEye_{v3}	+ ~ 51	+ ~ 802	+ ~ 51,202	+ ~ 409,602	+ ~ 3,276,802
QualitEye_{v4}	+ ~ 201	+ ~ 223	+ ~ 1,284	+ ~ 13,894	+ ~ 58,600

6.2.5 ***Robustness vs. Privacy***. We assess robustness by comparing hashes of the disentangled features (Table 1) to the original images (without disentanglement). For the latter, we also modify the appearance, while maintaining the gaze and pose vectors, with different collision and evasion attacks (c.f. Appendix A): Collisions correspond to incorrectly matching samples with different gaze or head pose, while evasion attacks correspond to failing to match samples with the same gaze/pose due to minor appearance changes. Our results show that hashing raw images is highly vulnerable to evasion (small appearance changes cause large hash differences - MCC = -0.87 and -0.9 for MPIIGaze and GazeCapture), whereas disentangled perceptual hashes strike the desired balance: They resist evasion via irrelevant appearance changes while maintaining low collision rates across genuinely different gaze configurations.

We also evaluate privacy. Perceptual hashes are irreversible, as they compress the input into a low-dimensional code that preserves only the features relevant for similarity comparison while discarding all other details. By comparing images of the same participant (i.e., same appearance), we find that the similarity metric remains too low to reliably map them to the same person (MCC = 0.01 and -0.07 for MPIIGaze and GazeCapture, respectively), indicating that appearance information is effectively removed. Beyond appearance, no other attributes can be reliably inferred. Although the datasets utilised in this study do not provide participant demographic information, precluding any further analysis based on age, gender, ethnicity, or other subject-specific factors, the irreversibility of the mapping from an input x to a disentangled latent representation $z = E(x)$ and subsequently to a perceptual hash $s = h(z)$ can be formalised using information-theoretic arguments. The perceptual hash $h(\cdot)$ compresses z into a fixed-length code $s \in \{0,1\}^d$. Because the latent space $\mathcal{Z}$ has much higher intrinsic dimensionality than the hash length d, the composition $F = h \circ E : \mathcal{X} \to \{0,1\}^d$ is many-to-one, i.e., multiple distinct inputs $x_1 \neq x_2$ may yield the same hash s. From an information-theoretic perspective, the entropy of the input satisfies $H(X) \gg H(S) \leq d$, implying that the mutual information $I(X;S) \leq H(S) \ll H(X)$, and therefore no deterministic or probabilistic function can reconstruct x from s. Together, robustness to irrelevant variations and irreversibility of sensitive details make perceptual hashes essential for QualitEye.

7 Discussion

Our results show that the gaze data quality verification problem can be solved efficiently under different public and privacy-preserving setups while achieving a good tradeoff between performance, runtime, and communication.

Our experiments demonstrate that the proposed VAE effectively disentangles appearance, gaze direction, and head pose

into separate latent codes. Gaze and head pose representations remain consistent across variations in appearance (MCC gaze: 0.92–0.94, head pose: 0.75–0.89). This success is achieved by explicitly enforcing known physical transformations, i.e. geometric rotations, rather than relying on implicit adversarial objectives. QualitEye also performs robustly across datasets with different characteristics (MCC = 0.93), demonstrating that the method can handle discrepancies in environment, devices, and annotation quality. In addition, the privacy-preserving implementation maintains the same verification performance as the public setup while introducing manageable runtime overhead. The hashes are irreversible, preventing the inference of other participant attributes beyond gaze and head pose. We further confirm our theoretical hypotheses in section 3: QualitEye$_{v1}$ can be used when both parties have relatively small datasets (i.e. tens to hundreds of samples) and would like to know the exact (mis)matching samples. QualitEye$_{v2}$ can be used when both parties have relatively small datasets (i.e. tens to hundreds of samples) and are only interested in knowing the cardinality of the intersection with an additional shuffling step. QualitEye$_{v3}$ can be used when one party has a relatively small dataset while the reference party owns a larger dataset (i.e. thousands to billions of samples) by offloading the larger dataset computation to an offline phase, i.e. asymmetric scenarios. QualitEye$_{v4}$ can be used when one or both parties own large datasets as the runtime with OPRF-based approach decreases significantly with respect to the DH-based approaches when the dataset size increases.

Limitations and future work. We focus on gaze angles and head poses, as they are the most commonly available labels in gaze-based datasets. While this problem has not been previously studied, we hypothesise that gaze data quality verification can be further improved by disentangling additional factors, such as illumination conditions [55]. We restrict our scope to image-based gaze tasks, noting that other modalities (e.g., scanpaths or videos) require specialised models and privacy assumptions due to temporal dependencies. Although we present a two-party computation (2PC) protocol, it can be extended to multi-party computation (MPC), where incorporating additional data sources may further reduce verification errors. Finally, QualitEye assumes a semi-honest threat model[10]; stronger security guarantees under malicious adversaries are possible, eliminating collision and evasion attacks, albeit reduced efficiency [77].

8 Conclusion

We presented QualitEye– the first work to investigate the problem of gaze data quality verification. We introduced a new generic hashed representation learning model that disentangles the gaze direction and head pose features for a high-domain adaptation performance, ignoring the cross-user irrelevant features (e.g. appearance) to allow for a label-specific comparison. Furthermore, we extended existing privacy-preserving interactive protocols with semantic similarities and labels matching to handle the different privacy and trust requirements. Our results show that QualitEye is efficient under different public and privacy-preserving setups in terms of performance, runtime, and communication.

Acknowledgments

M. Elfares was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 548713845. P. Reisert and R. Küsters were supported by the German Federal Ministry of Research, Technology and Space as part of the QCyber project under grant agreement 16KIS2590K, and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grants 411720488 and 548713845.

[10]The semi-honest model aligns with our intended deployment in collaborative research and data-sharing settings where parties (e.g., research groups or dataset providers) are expected to follow the protocol but may attempt to infer additional information from exchanged messages, without manipulation.

Privacy and Ethics Statement

QualitEye prioritizes privacy, enabling data sharing and cross-domain evaluation without raw eye images, thus supporting reproducibility and collaboration. It benefits applications in human-computer interaction, accessibility, and healthcare, where high-quality gaze data is needed but privacy is critical. Deployments should comply with data protection regulations (e.g., GDPR), ensure informed consent, and maintain transparency. All datasets used are public, and QualitEye's design mitigates potential misuse.

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)* 3, 3 (2019), 1–27. doi:10.1145/3351227

[2] Ahmed Abdou, Ekta Sood, Philipp Müller, and Andreas Bulling. 2022. Gaze-enhanced Crossmodal Embeddings for Emotion Recognition. In *ACM ETRA (ETRA, Vol. 6)*. ACM, 1–18. doi:10.1145/3530879

[3] Allon Adir, Ehud Aharoni, Nir Drucker, Eyal Kushnir, Ramy Masalha, Michael Mirkin, and Omri Soceanu. 2022. Privacy-preserving record linkage using local sensitive hash and private set intersection. In *International Conference on Applied Cryptography and Network Security*. Springer, 398–424.

[4] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2016. Deep Variational Information Bottleneck. *CoRR* abs/1612.00410 (2016). arXiv:1612.00410 http://arxiv.org/abs/1612.00410

[5] Nour Almadhoun, Erman Ayday, and Özgür Ulusoy. 2020. Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics* 36, 6 (2020), 1696–1703.

[6] David W. Archer, Borja de Balle Pigem, Dan Bogdanov, Mark Craddock, Adria Gascon, Ronald Jansen, Matjaž Jug, Kim Laine, Robert McLellan, Olga Ohrimenko, Mariana Raykova, Andrew Trask, and Simon Wardley. 2023. UN Handbook on Privacy-Preserving Computation Techniques. arXiv:2301.06167 [cs.CY]

[7] Samantha Aziz and Oleg Komogortsev. 2022. An assessment of the eye tracking signal quality captured in the HoloLens 2. In *2022 Symposium on eye tracking research and applications*. 1–6.

[8] Samantha Aziz, Dillon J Lohr, Lee Friedman, and Oleg Komogortsev. 2024. Evaluation of Eye Tracking Signal Quality for Virtual Reality Applications: A Case Study in the Meta Quest Pro. *arXiv preprint arXiv:2403.07210* (2024).

[9] Dale J Barr. 2008. Analyzing 'visual world'eyetracking data using multilevel logistic regression. *Journal of memory and language* 59, 4 (2008), 457–474.

[10] Donald Beaver. 1996. Correlated pseudorandomness and the complexity of private computations. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. 479–488.

[11] Bellare, Namprempre, Pointcheval, and Semanko. 2003. The one-more-RSA-inversion problems and the security of Chaum's blind signature scheme. *Journal of Cryptology* 16 (2003), 185–215.

[12] E. Blass and F. Kerschbaum. 2023. Private Collaborative Data Cleaning via Non-Equi PSI. In *IEEE S&P*. 1419–1434.

[13] Remco Bloemen, Bryan Gillespie, Daniel Kales, Philipp Sippl, and Roman Walch. 2024. Large-scale MPC: Scaling private iris code uniqueness checks to millions of users. *arXiv preprint arXiv:2405.04463* (2024).

[14] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one* 12, 6 (2017), e0177678.

[15] 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.

[16] 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).

[17] Andreas Bulling and Hans Gellersen. 2010. Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12. doi:10.1109/MPRV.2010.86

[18] Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2011), 741–753. doi:10.1109/TPAMI.2010.86

[19] 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)*. doi:10.1145/3313831.3376449

[20] Jan Camenisch, Markulf Kohlweiss, Alfredo Rial, and Caroline Sheedy. 2009. Blind and anonymous identity-based encryption and authorised private searches on public key encrypted data. In *Public Key Cryptography–PKC 2009: 12th International Conference on Practice and Theory in Public Key Cryptography, Irvine, CA, USA, March 18-20, 2009. Proceedings 12*. Springer, 196–214.

[21] Jan Camenisch and Gregory M Zaverucha. 2009. Private intersection of certified sets. In *Financial Cryptography and Data Security: 13th International Conference, FC 2009, Accra Beach, Barbados, February 23-26, 2009. Revised Selected Papers 13*. Springer, 108–127.

[22] 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.

[23] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. 2019. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2959–2968.

[24] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems* 29 (2016).

[25] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15750–15758.

[26] Jose Contreras and Hardik Gajera. 2022. DeV-IP: A k-out-n Decentralized and verifiable BFV for Inner Product evaluation. *Cryptology ePrint Archive* (2022).

[27] Antoine Coutrot, Janet H Hsiao, and Antoni B Chan. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods* 50, 1 (2018), 362–379.

[28] Benjamin Cowley, Marco Filetti, et al. 2016. The psychophysiology primer: a guide to methods and a broad review with a focus on human–computer interaction. *Found. Trends Hum.-Comput. Interact.* 9, 3-4 (2016), 151–308.

[29] Emiliano De Cristofaro and Gene Tsudik. 2012. Experimenting with fast private set intersection. In *International Conference on Trust and Trustworthy Computing*. Springer, 55–73.

[30] Pierpaolo Della Monica, Ivan Visconti, Andrea Vitaletti, and Marco Zecchini. 2024. Trust Nobody: Privacy-Preserving Proofs for Edited Photos with Your Laptop. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 14–14.

[31] Whitfield Diffie. 1976. New direction in cryptography. *IEEE Trans. Inform. Theory* 22 (1976), 472–492.

[32] Heiko Drewes. 2010. *Eye gaze tracking for human computer interaction.* Ph. D. Dissertation. LMU Munich. doi:10.5282/edoc.11591

[33] Thai Duong, Duong Hieu Phan, and Ni Trieu. 2020. Catalic: Delegated PSI cardinality with applications to contact tracing. In *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 870–899.

[34] Mayar Elfares, Zhiming Hu, Pascal Reisert, Andreas Bulling, and Ralf Küsters. 2022. Federated Learning for Appearance-based Gaze Estimation in the Wild. *NeurIPS-GMML* (2022). doi:10.48550/arXiv.2211.07330

[35] Mayar Elfares, Pascal Reisert, Wenwu Tang, Zhiming Hu, Ralf Küsters, and Andreas Bulling. 2024. PrivatEyes: Appearance-based Gaze Estimation Using Federated Secure Multi-Party Computation. *ACM ETRA* (2024).

[36] Yasmine Elfares, Gül Çalikli, and Mohamed Khamis. 2025. GazeCopilot: Evaluating Novel Gaze-Informed Prompting for AI-Supported Code Comprehension and Readability. arXiv:2511.08177 [cs.HC] https://arxiv.org/abs/2511.08177

[37] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Proceedings of the 2017 Chi conference on human factors in computing systems*. 1118–1130.

[38] Jesús García-Rodríguez, Stephan Krenn, and Daniel Slamanig. 2024. To pass or not to pass: Privacy-preserving physical access control. *Computers & Security* 136 (2024), 103566.

[39] Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 169–178.

[40] Oded Goldreich. 2004. *Foundations of Cryptography, Volume 2.* Cambridge university press Cambridge.

[41] Oded Goldreich, Silvio Micali, and Avi Wigderson. 2019. How to play any mental game, or a completeness theorem for protocols with honest majority. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*. 307–328.

[42] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. 2021. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988* (2021).

[43] Céline Gressel, Rebekah Overdorf, Inken Hagenstedt, Murat Karaboga, Helmut Lurtz, Michael Raschke, and Andreas Bulling. 2023. Privacy-Aware Eye Tracking: Challenges and Future Directions. *IEEE Pervasive Computing* 22, 1 (2023), 95–102. doi:10.1109/MPRV.2022.3228660

[44] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

[45] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye tracker data quality: What it is and how to measure it. In *Proceedings of the symposium on eye tracking research and applications*. 45–52.

[46] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye tracker data quality: what it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 45–52. doi:10.1145/2168556.2168563

[47] 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. doi:10.3389/fnhum.2018.00105

[48] Michael Xuelin Huang, Tiffany CK Kwok, Grace Ngai, Stephen CF Chan, and Hong Va Leong. 2016. Building a personalized, auto-calibrating eye tracker from user interactions. In *ACM CHI*. 5169–5179.

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

[50] 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.

[51] Bernardo A Huberman, Matt Franklin, and Tad Hogg. 1999. Enhancing privacy and trust in electronic communities. In *Proceedings of the 1st ACM conference on Electronic commerce*. 78–86.

[52] Apple Inc. 2021. CSAM Detection - Technical Summary. https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf.

[53] Yuval Ishai, Joe Kilian, Kobbi Nissim, and Erez Petrank. 2003. Extending oblivious transfers efficiently. In *Annual International Cryptology Conference*. Springer, 145–161.

[54] Deborah N. Jakobi, Daniel G. Krakowczyk, and Lena A. Jäger. 2024. Reporting Eye-Tracking Data Quality: Towards a New Standard. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications* (<conf-loc>, <city>Glasgow</city>, <country>United Kingdom</country>, </conf-loc>) *(ETRA '24)*. Association for Computing Machinery, New York, NY, USA, Article 47, 3 pages. doi:10.1145/3649902.3655658

[55] Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. 2023. NeRFFaceLighting: Implicit and Disentangled Face Lighting Representation Leveraging Generative Prior in Neural Radiance Fields. *ACM Transactions on Graphics* 42, 3 (2023), 1–18.

[56] Swati Jindal and Roberto Manduchi. 2023. Contrastive representation learning for gaze estimation. In *Annual Conference on Neural Information Processing Systems*. PMLR, 37–49.

[57] Kaggle. 2024. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com.

[58] Daniel Kales, Christian Rechberger, Thomas Schneider, Matthias Senker, and Christian Weinert. 2019. Mobile private contact discovery at scale. In *28th USENIX Security Symposium (USENIX Security 19)*. 1447–1464.

[59] Thivya Kandappu, Archan Misra, Shih-Fen Cheng, Randy Tandriansyah, and Hoong Chuin Lau. 2018. Obfuscation at-source: Privacy in context-aware mobile crowd-sourcing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–24.

[60] Mohamed Khamis, Daniel Buschek, Tobias Thieron, Florian Alt, and Andreas Bulling. 2017. EyePACT: Eye-Based Parallax Correction on Touch-Enabled Interactive Displays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 4 (2017), 1–18. doi:10.1145/3161168

[61] Mohamed Khamis, Ludwig Trotter, Ville Mäkelä, Emanuel von Zezschwitz, Jens Le, Andreas Bulling, and Florian Alt. 2018. CueAuth: Comparing Touch, Mid-Air Gestures, and Gaze for Cue-based Authentication on Situated Displays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 7 (2018), 1–22. doi:10.1145/3287052

[62] Miran Kim and Kristin Lauter. 2015. Private genome analysis through homomorphic encryption. In *BMC medical informatics and decision making*, Vol. 15. Springer, 1–12.

[63] Vladimir Kolesnikov, Ranjit Kumaresan, Mike Rosulek, and Ni Trieu. 2016. Efficient batched oblivious PRF with applications to private set intersection. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 818–829.

[64] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *IEEE ICPR*. 2176–2184.

[65] Huining Li, Xiaoye Qian, Ruokai Ma, Chenhan Xu, Zhengxiong Li, Dongmei Li, Feng Lin, Ming-Chun Huang, and Wenyao Xu. 2023. TherapyPal: Towards a Privacy-Preserving Companion Diagnostic Tool based on Digital Symptomatic Phenotyping. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[66] 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.

[67] Ao Liu, Lirong Xia, Andrew Duchowski, Reynold Bailey, Kenneth Holmqvist, and Eakta Jain. 2019. Differential privacy for eye-tracking data. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–10.

[68] Gang Liu, Yuechen Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. 2018. A differential approach for gaze estimation with calibration.. In *BMVC*, Vol. 2. 6.

[69] Dillon J Lohr, Lee Friedman, and Oleg V Komogortsev. 2019. Evaluating the data quality of eye tracking signals from a virtual reality system: Case study using SMI's eye-tracking HTC vive. *arXiv preprint arXiv:1912.02083* (2019).

[70] Päivi Majaranta and Andreas Bulling. 2014. Eye tracking and eye-based human–computer interaction. *Advances in physiological computing* (2014), 39–65. doi:10.1007/978-1-4471-6392-3_3

[71] Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, 2 (1975), 442–451.

[72] Meta. 2019. Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer. https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

[73] Microsoft. 2015. PhotoDNA. https://www.microsoft.com/en-us/photodna.

[74] Dimitris Mouris, Daniel Masny, Ni Trieu, Shubho Sengupta, Prasad Buddhavarapu, and Benjamin Case. 2024. Delegated Private Matching for Compute. *Proceedings on Privacy Enhancing Technologies* (2024).

[75] Rasmus Pagh and Flemming Friche Rodler. 2001. Cuckoo hashing. In *European Symposium on Algorithms*. Springer, 121–133.

[76] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. 2019. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9368–9377.

[77] Benny Pinkas, Mike Rosulek, Ni Trieu, and Avishay Yanai. 2020. PSI from PaXoS: fast, malicious private set intersection. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 739–767.

[78] Michael O Rabin. 2005. How to exchange secrets with oblivious transfer. *Cryptology ePrint Archive* (2005).
[79] Christoph Sager, Christian Janiesch, and Patrick Zschech. 2021. A survey of image labelling for computer vision applications. *Journal of Business Analytics* 4, 2 (2021), 91–110.
[80] 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.
[81] Ryo Shimata, Yoshihiro Mitani, and Tsumoru Ochiai. 2015. A study of pupil detection and tracking by image processing techniques for a human eye-computer interaction system. In *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, 1–4.
[82] 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.
[83] Qiyang Song, Jiahao Cao, Kun Sun, Qi Li, and Ke Xu. 2021. Try before you buy: Privacy-preserving data evaluation on cloud-based machine learning data marketplace. In *Annual Computer Security Applications Conference*. 260–272.
[84] Julian Steil and Andreas Bulling. 2015. Discovery of everyday human activities from long-term visual behaviour using topic models. In *ACM UbiComp (UbiComp)*. ACM, 75–85. doi:10.1145/2750858.2807520
[85] Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling. 2019. Privacy-aware eye tracking using differential privacy. In *ACM ETRA*. 1–9. doi:10.1145/3314111.3319915
[86] Lukas Struppek, Dominik Hintersdorf, Daniel Neider, and Kristian Kersting. 2022. Learning to break deep perceptual hashing: The use case neuralhash. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 58–69.
[87] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *IEEE ICPR*. 1821–1828.
[88] Yunjia Sun, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2021. Cross-encoder for unsupervised gaze representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3702–3711.
[89] Erkam Uzun, Simon P Chung, Vladimir Kolesnikov, Alexandra Boldyreva, and Wenke Lee. 2021. Fuzzy labeled private set intersection with applications to private {Real-Time} biometric search. In *30th USENIX Security Symposium (USENIX Security 21)*. 911–928.
[90] Yunhan Wang, Xiangwei Shi, Shalini De Mello, Hyung Jin Chang, and Xucong Zhang. 2023. Investigation of Architectures and Receptive Fields for Appearance-based Gaze Estimation. arXiv:2308.09593 [cs.CV]
[91] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. 2020. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6509–6518.
[92] Victoria Yaneva, Le An Ha, Sukru Eraslan, Yeliz Yesilada, and Ruslan Mitkov. 2018. Detecting autism based on eye-tracking data from web searching tasks. In *Proceedings of the 15th International Web for All Conference*. 1–10.
[93] Edwin Yang, Qiuye He, and Song Fang. 2022. WINK: Wireless Inference of Numerical Keystrokes via Zero-Training Spatiotemporal Analysis. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS' 2022)*. ACM, 3033–3047.
[94] Andrew C Yao. 1982. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*. IEEE, 160–164.
[95] 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.
[96] 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 *European Conference on Computer Vision*. Springer, 365–381.
[97] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017. 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.
[98] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based Gaze Estimation in the Wild. In *IEEE CVPR*. 4511–4520. doi:10.1109/CVPR.2015.7299081
[99] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's written all over your face: Full-face appearance-based gaze estimation. In *IEEE CVPR Workshops*. 51–60.
[100] 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* 41, 1 (2019), 162–175. doi:10.1109/TPAMI.2017.2778103
[101] Ke Zhong and Sebastian Angel. 2024. Oryx: Private detection of cycles in federated graphs. *Cryptology ePrint Archive* (2024).

A Collision and Evasion Attacks

In the context of hashing-based and similarity-driven verification, adversarial behaviour is commonly characterised by *collision* and *evasion* attacks, which target different failure modes of the representation. We build our attacks on the work of Struppek et al [86]. Struppek et al. analyse the robustness of perceptual image hashing by explicitly formulating two complementary attack models: collision attacks and evasion attacks, both realised as optimisation problems over the image space.

A.1 Collision Attacks

A collision attack aims to identify two distinct inputs $x_1 \neq x_2$ that produce the same hash value, i.e., $h(x_1) = h(x_2)$. The objective is to falsely link or impersonate samples that are semantically different. In perceptual or similarity hashing, collisions are inherently possible due to the low-dimensional and many-to-one nature of the hash space. However, a secure and task-appropriate hash function ensures that collisions occur only for inputs that are semantically equivalent with respect to the target attributes (e.g., identical gaze direction and head pose), while remaining unlikely for dissimilar inputs.

Technically, an adversary starts with a target hash (or target image) and iteratively modifies another image so that its perceptual features align with the same low-dimensional representation used by the hash function. Because perceptual hashes rely on compressed, robust features, attackers manipulate precisely those components while allowing high-frequency or imperceptible changes elsewhere. This manipulation is done through a gradient-based collision attack. Since the final binary hash is non-differentiable, the attack operates on the continuous internal embedding produced by the hashing model before thresholding. An optimization objective is defined that (i) encourages the embedding of the modified image to align with the target embedding and (ii) penalizes perceptual distortion using similarity metrics. Using gradient-based optimization methods (e.g., projected gradient descent or Adam), the image pixels are iteratively adjusted in the direction that reduces hash distance, with updates constrained to valid pixel ranges. The attack succeeds because small changes in the continuous embedding—especially near threshold boundaries—can flip multiple hash bits, allowing two semantically different images to produce identical or near-identical perceptual hashes. The result is two semantically different images that are deemed similar by the hashing system, undermining content authentication or deduplication.

A.2 Evasion Attacks

An evasion attack seeks to modify an input x into x' such that $h(x') \neq h(x)$, despite the modification being visually imperceptible or semantically irrelevant. The goal is to prevent correct matching by exploiting the excessive sensitivity of the hashing function to nuisance factors such as lighting, texture, or minor appearance variations. Evasion attacks, therefore, directly assess the robustness of the representation to irrelevant perturbations.

In practice, the attacker starts from an original image and iteratively perturbs it to increase the Hamming distance between the original hash and the modified hash beyond the detection threshold. Because the final hash is binary and non-differentiable, the attack operates on the continuous internal embedding produced by the model before binarization. By computing gradients of a loss that encourages divergence in embedding space, the attacker identifies which pixels most strongly influence specific hash bits. The optimization is constrained by a perceptual similarity metric—such as MSE, SSIM, or LPIPS—to ensure that the modified image remains visually close to the original. At each iteration, the algorithm adjusts pixel values in directions that maximally shift the embedding while keeping distortion within a predefined bound. Since perceptual hashes typically rely on thresholding continuous features (e.g., sign of embedding components), small shifts near decision boundaries can flip multiple bits simultaneously. The attack therefore focuses on regions or frequency components that disproportionately affect these embedding dimensions.

A.3 Key Distinctions

The key distinction is that collision attacks cause *false matches* between different samples, whereas evasion attacks cause *missed matches* between equivalent samples.

In this work, we (i) fine-tune and optimise the models to our domain and, in addition to [86], (ii) incorporate mean squared error (MSE) and Learned Perceptual Image Patch Similarity (LPIPS). While SSIM captures structural and luminance-based similarities aligned with human perception, MSE provides a pixel-wise measure of absolute intensity differences, and LPIPS evaluates perceptual similarity in a deep feature space learned by neural networks. Using these complementary metrics allows us to assess robustness across low-level, structural, and semantic perturbations, yielding a more comprehensive characterisation of image modifications and their impact on perceptual hashing.

Our qualitative results are shown in Figures 9 and 10. When passing the resulted images to our pipeline, as discussed in the main text, perceptual hashes computed from full images without disentanglement are highly susceptible to attacks, whereas hashes derived from disentangled features exhibit strong robustness and are largely resistant to both collision and evasion attempts.

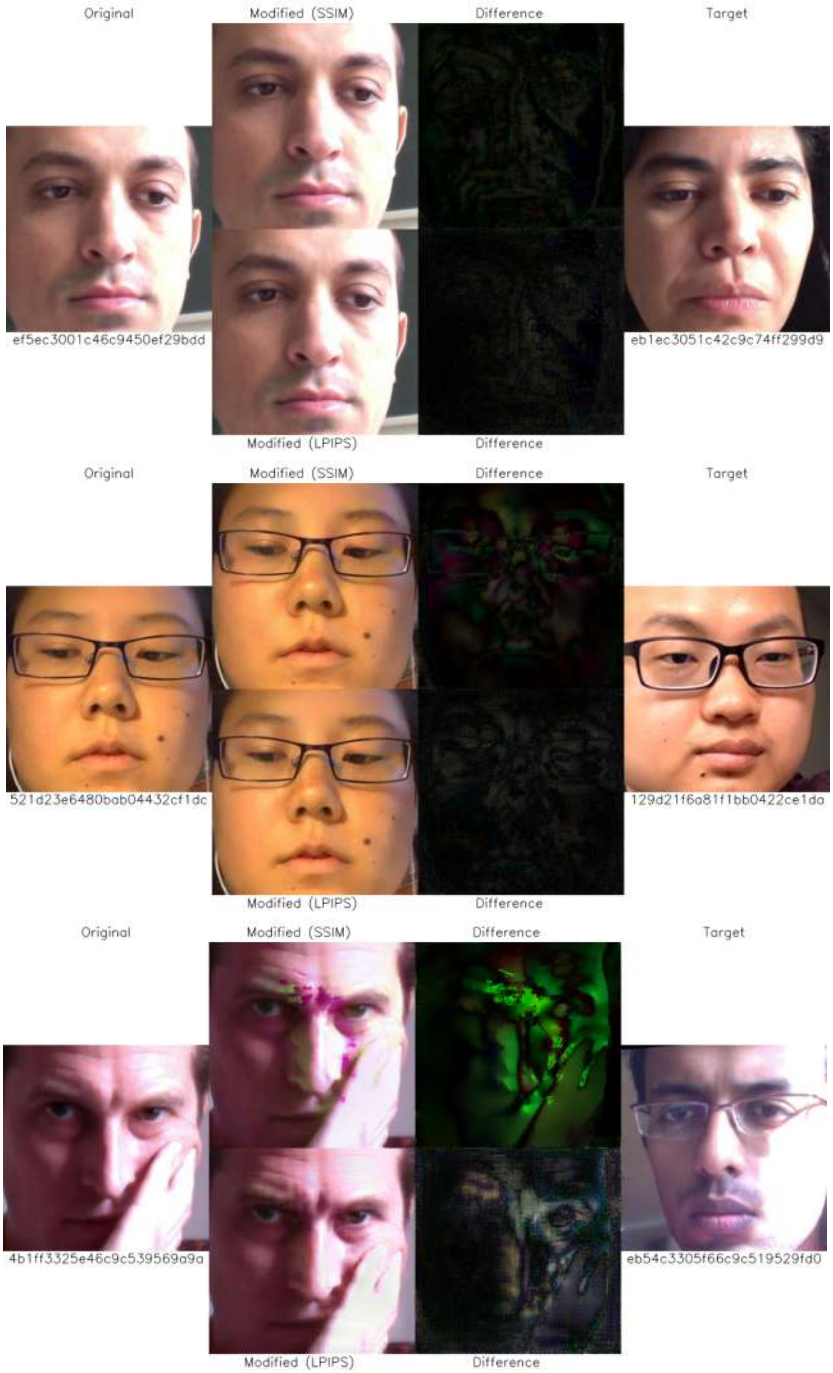

Fig. 9. Collision attack results. Each row presents the original image (left), the adversarially modified image (middle), and the target image (right) for a different participant. The corresponding difference plots for the modified images are also displayed; these plots are generated by subtracting the original and modified images using the absolute difference and scaling the result to enhance the visibility of changes. Below the original image, the starting hash value is displayed in hexadecimal, while below the target image, the hash value of the modified image is shown.

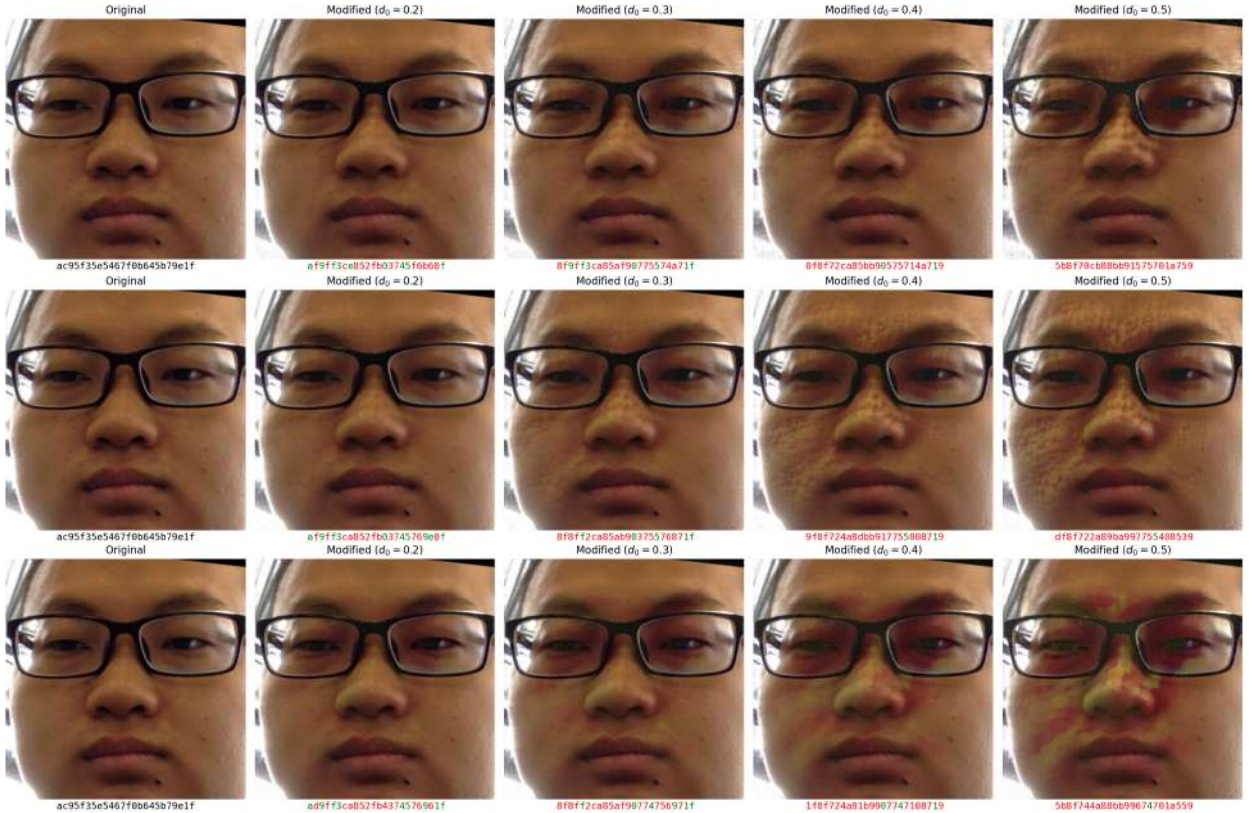

Fig. 10. Evasion modification with *LPIPS* (top), *MSE* (middle), and *SSIM* (bottom) with thresholds (0.2 to 0.5) with evasion attacks. Hash values are displayed below each image in hexadecimal format, with modified hashes colour-coded: red for changes and green for no change.