Assessing Privacy Risks from Feature Vector Reconstruction Attacks

Emily Wenger*, Francesca Falzon*†, Josephine Passananti*, Haitao Zheng*, Ben Y. Zhao*
* University of Chicago, † Brown University
{ewenger, josephinep, htzheng, ravenben}@uchicago.edu, {francesca_falzon}@brown.edu

Abstract—**In deep neural networks for facial recognition, feature vectors are numerical representations that capture the unique features of a given face. While it is known that a version of the original face can be recovered via "feature reconstruction," we lack an understanding of the end-to-end privacy risks produced by these attacks. In this work, we address this shortcoming by developing metrics that meaningfully capture the threat of reconstructed face images. Using end-to-end experiments and user studies, we show that reconstructed face images enable re-identification by both commercial facial recognition systems and humans, at a rate that is at worst, a factor of four times higher than randomized baselines. Our results confirm that feature vectors should be recognized as Personal Identifiable Information (PII) in order to protect user privacy.**

I. Introduction

Feature vectors allow quick identification of similar images, and both facial recognition engines and anti-facial recognition tools rely on them. Given an image x, a well-trained deep neural network (DNN) $\mathcal{F}$ maps x to a feature vector $\mathcal{F}(x) = v$ that represents x's visual and/or semantic features. Facial recognition systems use databases of feature vectors to produce identity matches via vector comparisons [1], [2], [3]. Anti-facial recognition tools like Fawkes, designed to prevent unwanted facial recognition, corrupt feature vectors to prevent identification [4]. Some tools also propose "collaborative" anti-facial recognition schemes, where individuals share their feature vectors with others for enhanced protection [5].

The privacy risks of publicly sharing face feature vectors are explicitly or implicitly assumed to be minimal. For example, several facial recognition companies state on their websites that, since face feature vectors are numerical, they will not leak sensitive personal information even if the feature vector database is hacked [6], [7]. Similarly, anti-facial recognition tools with feature vector sharing schemes do not consider whether this sharing could compromise privacy. Such assumptions are concerning, especially given the significant regulatory standards for protection of other biometric data [8], [9]. Face feature vectors also contain biometric data, so the privacy risks of using or sharing them must be better understood.

One obvious threat to face feature vector privacy comes from so-called "feature vector reconstruction" (FVR) techniques. FVR transforms the feature vector v back into the image x [10], [11], [12], [13], [14]. Prior work on inverse biometrics has concluded that reconstruction alone is a severe attack [15], [16]. However, in the context of facial recognition, FVR could lead to a more serious privacy threat – deanonymization. Since the goal of FVR is to recreate the original image x, reconstructed face images could be fed back into a facial recognition system and re-identified. In addition to the privacy loss from deanonymization, reidentified face images could be used to conduct further attacks, like creating Deepfakes or crafting a 3D model of the victim's face to fool facial recognition systems [17], [18], [19].

FVR-enabled deanonymization poses a plausible and potentially significant privacy risk, but assessing this threat requires a carefully designed metric. Prior work on face FVR measures the visual similarity between original and reconstructed images, implicitly making a strong assumption that the attacker already knows the target's identity. In practice, the target identity is unknown, and finding a target match visually would require inspecting a huge set of images. To scale the attack, an attacker needs to use a facial recognition engine to identify a small set of potential matches for visual inspection. Because facial recognition engines map images to feature vectors, the reconstructed images must have similar features to the target (to be among the potential matches) and be visually similar enough so the attacker can identify them from the matches. Thus, to measure the real-world privacy risk of FVR-enabled deanonymization we believe we need to consider both feature space and visual similarity.

To fill this gap, our work proposes two metrics. The first, *top-K matching accuracy*, assesses whether a reconstructed image would lead a facial recognition engine to produce a set of matches containing the true target identity. The second compares the *visual matching accuracy* between reconstructed and true images via a user study. Armed with these metrics, we explore the limits of FVR-enabled deanonymization attacks using state-of-the-art FVR methods, real-world facial recognition systems, and a user study. Despite the feature-space noise that current FVR methods add to reconstructed images, our findings show that FVR-enabled deanonymization poses a near-term threat and that the research community must take the necessary steps to mitigate this threat.

Our Contributions. We make the following contributions to the community's understanding of FVR-enabled deanonymization attacks.

- We develop two concrete metrics – topK matching and visual matching – that assess the real-world privacy risk of FVR-enabled deanonymization attacks (§III).
- We evaluate the deanonymization success of four state-of-

the-art FVR methods using commercial facial recognition systems and an IRB-approved user study (§V, §VI) and find that deanonymization is possible.

II. Background

Before discussing the threat model and evaluation methodology of our work, we first provide background information on feature vector reconstruction and facial recognition.

Feature Vector Reconstruction. FVR methods use a reconstructor $\mathcal{R}$ to transform a feature vector $\mathcal{F}(x) = v$ into a reconstructed image $\tilde{x}$. More formally, let W, H, C, and M be integers and let $\mathcal{F} : \mathbb{R}^{W \times H \times C} \to \mathbb{R}^{M}$ be a function defined by a feature extractor that takes as input an image x and outputs a face feature vector v; v is an M-dimensional vector which represents x in the feature space of $\mathcal{F}$. Given a feature vector v we want to generate an image $\tilde{x}$, such that $\mathcal{F}(\tilde{x}) = v$. We thus define the problem of feature vector reconstruction as follows:

Definition 1: Given black-box access to a model $\mathcal{F}$ and a feature vector $v = \mathcal{F}(x)$ of an unknown image x, the goal of *feature vector reconstruction* (FVR) is to recover an image $\tilde{x}^* = \arg\min_{\tilde{x}} \mathcal{L}(\mathcal{F}(\tilde{x}), v)$, where $\mathcal{L}$ is some loss function.

The feature vector reconstructor $\mathcal{R}_{\mathcal{F}}$ is designed to invert feature vectors produced by $\mathcal{F}$, so $\mathcal{R}_{\mathcal{F}}(v) = \tilde{x}$. In practice, $\mathcal{R}_{\mathcal{F}}$ is typically either a trained model (parametric method) or an optimization procedure (nonparametric method). *Parametric reconstruction methods* rely on a reconstruction model $\mathcal{R}$ trained specifically to invert vectors produced by a model $\mathcal{F}$ [13], [10], [20]. *Nonparametric reconstruction methods* use an iterative optimization process $\mathcal{R}$ to reconstruct images from $\mathcal{F}$'s feature vectors [14], [21].

Facial Recognition. Since we evaluate FVR in the context of facial recognition, we now briefly describe how modern facial recognition systems work. Facial recognition works in three phases: *enrollment*, *query*, and *matching*, shown in Fig. 1. In the enrollment phase, labeled face images are transformed into feature vectors and stored in a reference database. In the query phase, a user runs an unlabeled image against the reference database, and the system returns the top-K matches for the image, based on the similarity of their feature vectors. Finally, in the matching phase, the user chooses the best match from the top-K set [22], [3]. This general setup is commonly used in commercial facial recognition systems [23].

Relationship between FVR and Facial Recognition. The three phases of facial recognition correspond to stages of FVR-enabled deanonymization attacks. Attackers must first obtain feature vectors to reconstruct. These could be obtained from leaked facial recognition reference databases (created in the *enrollment phase*) [24], [25] or from proposed vector-sharing schemes [4], [5]. Once the attacker reconstructs these feature vectors, their reconstructions must be good enough to be identified in both the *query* and *matching* phases for the images to be deanonymized.

III. Threat Model and Evaluation Metrics

With this background in mind, we now formalize the threat model and evaluation metrics for FVR-enabled deanonymization attacks on face images. We assume the attacker attempts deanonymization via a facial recognition engine, as manual reidentification of reconstructed images is time-consuming and does not scale.

A. Attack Stages and Assumptions

We consider an adversary $\mathcal{A}$, whose FVR-enabled deanonymization attack operates in three stages, illustrated in Figure 2. We assume $\mathcal{A}$ has a feature vector v and blackbox access to the DNN $\mathcal{F}$ that created it (*initial stage*). To conduct the attack, $\mathcal{A}$ inverts v to recover $\tilde{x}$, an approximation of real image x of person P (*stage 1*); deanonymizes $\tilde{x}$ to recover P's identity (*stage 2*); and potentially conducts additional attacks using this knowledge (*stage 3*). Below, we describe the stages in more detail.

Attack stage 1: reconstruction. $\mathcal{A}$ first chooses a reconstructor $\mathcal{R}_{\mathcal{F}}$ and uses it to invert $\mathcal{F}(x) = v$, producing $\tilde{x} = \mathcal{R}_{\mathcal{F}}(v)$. We assume $\mathcal{A}$ has a feature vector $v = \mathcal{F}(x)$ generated from a face image x from unknown person P whose identity they wish to recover. Furthermore, we assume $\mathcal{A}$ has black-box access to the feature extractor $\mathcal{F}$ used to create v and trains $\mathcal{R}_{\mathcal{F}}$ using dataset $\mathcal{X}_{\mathcal{R}}, \mathcal{Y}_{\mathcal{R}}$, $x \notin \mathcal{X}_{\mathcal{R}}$.

Attack stage 2: deanonymization (primary attack). Next, $\mathcal{A}$ uses a *secondary facial recognition system* $\mathcal{F}'$ and/or *visual identification* to deanonymize $\tilde{x}$. In most cases, these two methods are used together, but in limited cases $\mathcal{A}$ could use visual identification alone – if, for example, P is a public figure who is easily recognized. In the more common scenario that both $\mathcal{F}'$ and visual identification are used, we assume $\mathcal{F}'$ is a commercial FR system that provides the top-K potential identity matches for an input image $\tilde{x}$ (see Figure 1). $\mathcal{A}$ then chooses the best visual match (if any) from the K possibilities.

Attack stage 3: integrity attack (secondary attack). Though our work measures the threat of deanonymization, $\mathcal{A}$ may conduct additional attacks after $\tilde{x}$ has been deanonymized. For example, $\mathcal{A}$ could use $\tilde{x}$ to generate Deepfakes imitating P or construct a 3D representation of P's face to fool biometric face recognition systems [19]. To conduct each of these, the attacker needs only internet access and a reasonably powerful computer. Code for both integrity attacks is available freely online [26], [27].

B. Evaluation Metrics

An efficient adversary would attempt deanonymization via a secondary facial recognition engine $\mathcal{F}' \neq \mathcal{F}$ that produces a set of potential identity matches for $\tilde{x}$. This attack technique naturally yields two evaluation metrics, one based on $\mathcal{F}'$'s **top-K matching accuracy**, and one based on the **visual similarity**, for measuring the deanonymization risk posed by different FVR methods. Both metrics are derived from key operational components of modern facial recognition systems (see Figure 1). Below, we describe the metrics in detail and note their limited use in prior work.

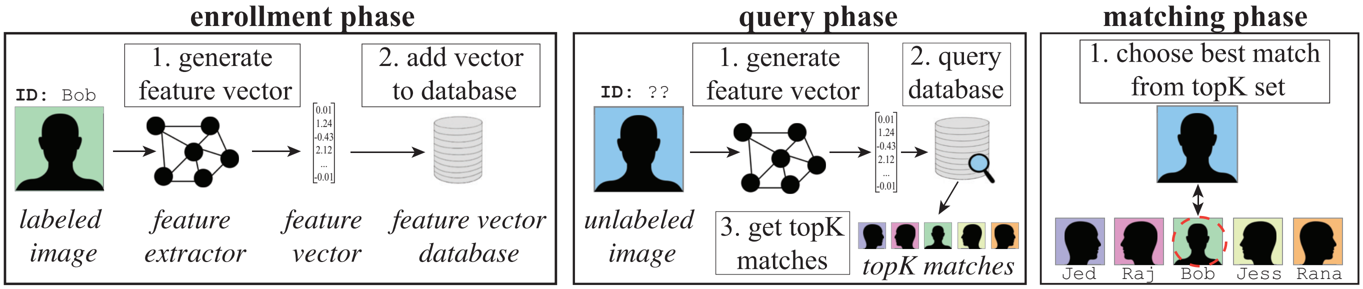

Fig. 1: *Overview of three stages of facial recognition: enrollment, query, and matching.*

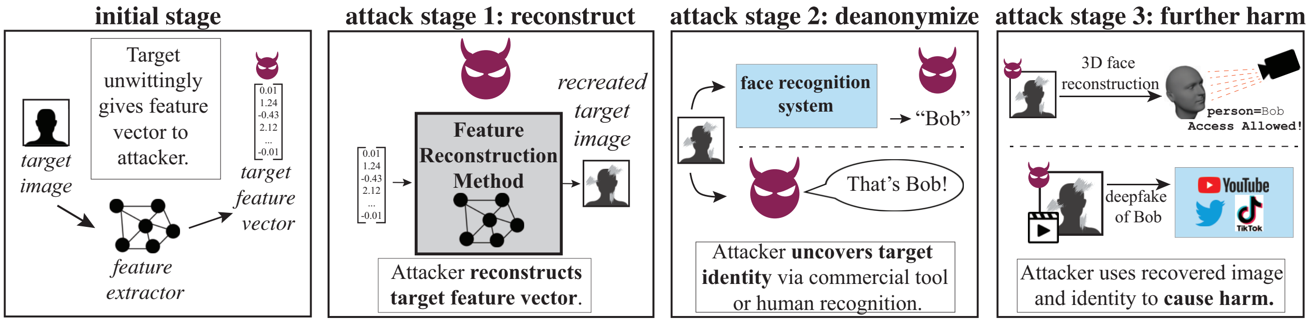
initial stage
Target unwittingly gives feature vector to attacker.
target image
target feature vector
feature extractor
0.01
1.24
-0.43
2.12
...
-0.01
attack stage 1: reconstruct
Feature Reconstruction Method
recreated target image
Attacker reconstructs target feature vector.
attack stage 2: deanonymize
face recognition system
"Bob"
That's Bob!
Attacker uncovers target identity via commercial tool or human recognition.
attack stage 3: further harm
3D face reconstruction
person=Bob
Access Allowed!
deepfake of Bob
YouTube
TikTok
Attacker uses recovered image and identity to cause harm.

Fig. 2: *The main phases of FVR-enabled attacks.*

Metric 1: Top-K Matching Accuracy. The first phase of a facial recognition system is *query matching*, in which the system produces a set of K possible identity matches for a queried image. For $\tilde{x}$ to be identified, the system must return the P's true identity in the top-K set for $\tilde{x}$. Therefore, our first metric tests the frequency with which images of P appear in the top-K sets produced for $\tilde{x}$ by secondary facial recognition models, $\mathcal{F}' \neq \mathcal{F}$. We assume that P's true identity is enrolled in $\mathcal{F}'$. Prior FVR evaluations almost exclusively measure matching accuracy between x and $\tilde{x}$ using the *inverted model* $\mathcal{F}$, rather than a secondary model $\mathcal{F}'$. Top-K match accuracy using $\mathcal{F}$ provides a minimal baseline for feature vector reconstruction performance.

Metric 2: Visual Similarity. If images from correct class P appear in the top-K matches returned by $\mathcal{F}'$, $\mathcal{A}$ must identify them from the match set in the *matching* phase of a facial recognition system. Thus, visual similarity between $\tilde{x}$ and P's true appearance is the second key metric for assessing FVR-enabled deanonymization risk. In this work, we employ a user study to measure this, which allows us to empirically assess whether a human could deanonymize reconstructed images from a top-K set. No prior FVR evaluation has used a user study to evaluate reconstruction quality. Prior work primarily relies on SSIM [28] to measure visual similarity. However, SSIM compares structural similarity and fails to capture broader similarity between faces, especially if image backgrounds vary [29].

IV. Methodology

Before presenting our evaluations using these metrics, we describe the FVR methods, facial recognition models, and datasets used in our analysis and provide a brief overview of the evaluation process.

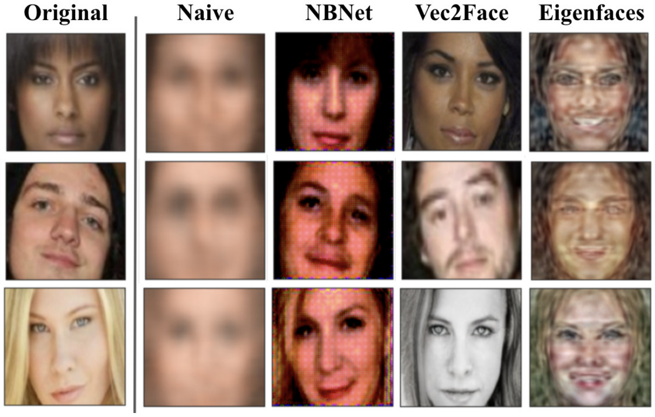

Fig. 3: *Examples of images reconstructed by each method.*

FVR Methods. We measure the deanonymization success of four state-of-the-art FVR methods. These methods are chosen because they encompass all currently identified methods of FVR (i.e. parametric and non-parametric), outperform older work, and have public or partially public implementations. Below, we provide a brief overview of each method. Examples of reconstructions from each method are in Fig. 3.

`NBNet` [13] trains a model $\mathcal{R}_{\mathcal{F}}$ using a deconvolutional neural network architecture similar to DenseNet [30]. We use the publicly available codebase [31] and train $\mathcal{R}_{\mathcal{F}}$ using the *NBNet-B* architecture for 80 epochs total with a batch size of 64. We use pixel mean-absolute-error loss for 60 epochs and add perceptual loss for the last 20 epochs.

`Vec2Face` [10] is generative adversarial net (GAN)-based and trains $\mathcal{R}_{\mathcal{F}}$ using a joint loss function balancing visual and feature space similarity between x and $\tilde{x}$. The GAN architecture is based on PoGAN [32], so we use the official PoGAN codebase to implement Vec2Face (for which no code was provided). Due to unanswered author correspondence, we omit ℓ_{biject} from the loss.[1] We train using the recommended batch size scaling for PoGAN and 15,000,000 training images.

`Naive` [20] is a convolutional neural net (CNN) built from stacked, inverted ResNet blocks [3]. We re-create the architecture based on the authors' detailed instructions and train each $\mathcal{R}_{\mathcal{F}}$ for 5 epochs with a batch size of 128.

`Eigenfaces` [21] is a *nonparametric* FVR method that works by iteratively adding Gaussian blobs to $\tilde{x}$ which minimizes the feature space distance between the $\mathcal{F}(\tilde{x})$ and v. At each step, $\tilde{x}$ is normalized using pre-computed eigenface representations to encourage visual similarity between x and $\tilde{x}$. We use author-provided code and run each $\mathcal{R}_{\mathcal{F}}$ procedure for 2000 iterations with batch size 64.

Facial Recognition Systems. We use several facial recognition (FR) models in our analysis. One set of local FR systems is used to train parametric FVR methods and generate feature vectors for testing ($\mathcal{F}$). The second set of commercial FR systems are used to evaluate Metric 1, the top-K matching accuracy of FVR methods in secondary models ($\mathcal{F}'$).

$\mathcal{F}$: Models for FVR Training/Testing. We use two local FR models to train and test FVR methods. Both are trained on the MS1M [33] dataset using the ArcFace [22] loss function. One uses a ResNet50 backbone [3], and the other uses an EfficientNet backbone [34]. We refer to these extractors as Res50 and Efficient respectively.

$\mathcal{F}'$: Models for Attack Evaluation. We measure FVR-enabled deanonymization success using both local FR models (i.e. Res50 and Efficient, described above) and real-world FR systems, Microsoft Azure and Amazon Rekognition. For all systems, we enroll a large set of labeled images to form our "matching set". When unlabeled images are submitted for identification, the systems return any potential matches found, along with a confidence score for each.

Datasets. We use five well-known facial recognition datasets in our evaluation: one to train models Res50 and Efficient, three to train the parametric FVR methods $\mathcal{R}$ (i.e. `NBNet`, `Vec2Face`, and `Naive`), and one to test reconstruction performance. The datasets, and their uses, are in Table I.

Use	Dataset	# Labels	# Images	Citation
$\mathcal{F}$ training	MS1M	85742	1,776,809	[33]
$\mathcal{R}$ training and testing	FaceScrub	530	57,838	[35]
	VGGFace2	8,631	1,047,297	[36]
	WebFace	10,575	475,137	[37]
$\mathcal{R}$ testing	LFW	5749	13233	[38]

TABLE I: *Datasets used for training and testing models $\mathcal{F}$ and reconstructors $\mathcal{R}$.*

Evaluation Overview. To perform our evaluation, we use the local FR systems Res50 and Efficient as $\mathcal{F}$ to train 18 FVR models $\mathcal{R}_{\mathcal{F}}$[2]. We then reconstruct images using the trained FVR models (i.e. parametric methods) as well as the non-parametric `Eigenfaces` method and measure deanonymization success on the reconstructions using Metrics 1 and 2 (§III-B). Our evaluation is structured as follows:

- §V reports results on Metric 1, top-K matching accuracy, using both local and commercial facial recognition systems.
- §VI reports results on Metric 2, visual similarity, via a survey-based user study.
- §VII synthesizes results and discusses the real-world threat of face image FVR, while proposing future work.

We report the results of reconstruction methods on Res50 feature vectors generated from FaceScrub images. Parametric FVR methods are trained using the WebFace dataset. Reconstruction results on Efficient feature vectors and other datasets are comparable and omitted for brevity.

V. Evaluating Metric 1: Top-K Matching

After $\mathcal{A}$ uses FVR to recover a face image $\tilde{x}$, their first step towards deanonymization is to run $\tilde{x}$ through a secondary facial recognition system $\mathcal{F}'$. $\mathcal{F}'$ returns the top-K matches for a test image, the K images with the highest similarity scores to $\tilde{x}$. Assuming P's true identity is enrolled in this system[3], real images of P should appear in the top-K options if the FVR attack is successful.

Measuring Metric 1. We report the top-K matching accuracy for a FVR method using two measurements: the *true positive rate* (*tpr*: the proportion of $\tilde{x}$ for which $\mathcal{F}'$ returned $>= 1$ real image of P in top-K) and the *false positive rate* (fpr: proportion of $\tilde{x}$ for which $\mathcal{F}'$ returned only incorrect matches in the top-K). We first compute these on our local Res50 and Efficient models, which allow more fine-grained measurements, before running experiments on Azure and Rekognition.

A. Evaluation on Local Models

To measure tpr and fpr, we run experiments varying K (match set size) and N (number of enrolled identities). This allows us to simulate real-world facial recognition systems with different parameters, providing deeper insight.

Results. The nonparametric method (`Eigenfaces`) vastly outperforms parametric methods on both tpr and fpr in top-K matching, regardless of K or N size. Increasing K and decreasing N slightly improves top-K tpr for all methods, but `Eigenfaces` still dominates. As Figures 4 and 5 show, `Eigenfaces` has a tpr of at least 80% across all K and N values, while the next best method, `Vec2Face` has a tpr

[1] Since Vec2Faces assigns a comparatively small weight (0.01) to the ℓ_{biject} loss term, this omission does not significantly affect performance

[2] We use 3 reconstruction methods (`NBNet`, `Vec2Face`, `Naive`) and 3 training datasets (WebFace, VGGFace2, FaceScrub) to invert each $\mathcal{F}$, resulting in 9 $\mathcal{R}_{\mathcal{F}}$ for a single $\mathcal{F}$ and 18 total $\mathcal{R}_{\mathcal{F}}$ across both $\mathcal{F}$. Recall that Eigenfaces does not require training a $\mathcal{R}$.

[3] A not-unreasonable assumption given the scale of modern facial recognition systems, c.f. Clearview.ai [39]

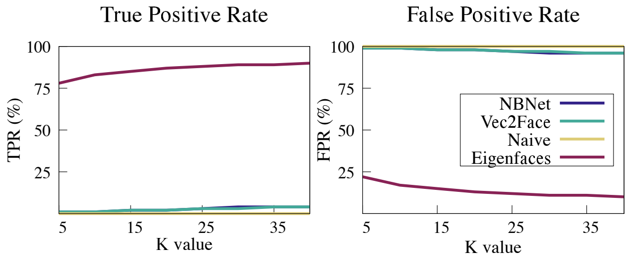

Fig. 4: *tpr and fpr rates for local $\mathcal{F}$ facial recognition models when $N = 1000$ and K varies from 5 to 40.*

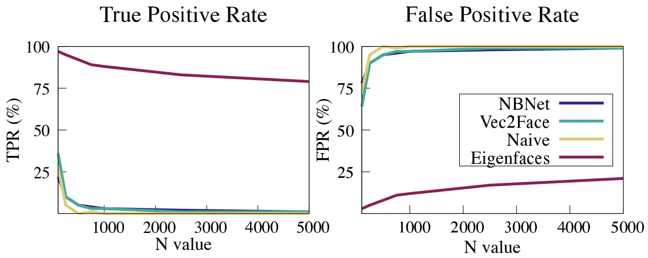

Fig. 5: *tpr and fpr rates for local $\mathcal{F}$ facial recognition models when $K = 25$ and N varies from 50 to 5000.*

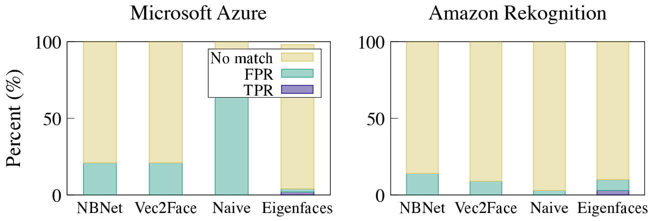

Fig. 6: *True positive rates, false positive rates, and "no match" rates for Azure and Rekognition.*

of 40% but only when the number of enrolled identities is $N < 50$. In a commercial system, N would be much larger, so this result overinflates `Vec2Face`'s performance.

Furthermore, `Vec2Face`, `NBNet`, and `Naive` have a high fpr ($> 90\%$) across all values of N and K. This means that deanonymization of images reconstructed with these methods is difficult.

B. Evaluation on Commercial Models

For our tests on Rekognition and Azure, we first enroll $N = 5000$ identities in each system by uploading labeled face images to their databases. 500 identities are drawn from the FaceScrub dataset, while 4500 are drawn from VGGFace2. The identities associated with all reconstructions $\tilde{x}$ we test are also enrolled in the system. To evaluate success, we query the reconstructions $\tilde{x}$ against the enrolled identity database. We set $K = 5$ (max), but since both systems only return a match when the similarity score exceeds an unspecified threshold, some $\tilde{x}$ receive no matches. We ensure the small set of overlapping classes between FaceScrub and VGGFace2 does not affect our results.

Results. As before, `Eigenfaces` outperforms all other FVR methods on both Azure and Rekognition (see Figure 6). However, for both systems, even `Eigenfaces` has $tpr < 5\%$. The other methods fare slightly better on Azure, but all have $tpr \leq 1\%$. Both Azure and Rekognition have high fpr for all methods (Fig. 6) – up to 85%. Interestingly, many inversions from the same method are matched to the same false positive class. Table II shows the percent of non-unique false positives (i.e. same class) for each system/method pair. Rekognition produces a higher percentage of non-unique false positives than Azure on average.

System	NBNet	Vec2Face	Naive	Eigenfaces
Rekognition	67%	54%	86%	77%
Azure	24%	9%	99%	23%

TABLE II: *Percent of false positives which are from the same class for each method and system.*

C. Discussion of Results

Three out of four FVR methods evaluated have nontrival top-K matching success, indicating a strong potential for full deanonymization. In addition to this key result, we observe the following notable behaviors.

Nonparametric method performs best. Images reconstructed using `Eigenfaces` are matched at higher rates than images reconstructed by other methods. This may be due to generalization errors in the parametric reconstruction models, a well-known problem for machine learning models [40]. As nonparametric methods do not rely on a trained model, they do not suffer from this problem.

Noise from reconstruction process affects performance. The many non-unique false positives observed in the commercial systems likely indicates that the reconstruction models add nontrivial, similar noise to all $\tilde{x}$. This noise causes the consistent false positive classifications observed. Even `Eigenfaces` has a 77% (23%) same-class false positive rate in Rekognition (Azure). This presents an obstacle for FVR-enabled deanonymization attacks. If the noise from the reconstruction process is too strong, deanonymization will become more difficult. However, future improvents to FVR techniques may mitigate this problem.

VI. Evaluation of Metric 2: Visual Similarity

If real images of P appear in the top-K set for $\tilde{x}$ (i.e. $tpr > 0$), $\mathcal{A}$ must be able to identify them to complete the deanonymization process (c.f. Fig 1). Consequently, visual similarity between $\tilde{x}$ and P is a key component of successful deanonymization attacks. In this section, we evaluate human ability to match $\tilde{x}$ to the true target identity in top-K matching sets for all four FVR methods using an IRB-approved user study (IRB information omitted for anonymous submission).

Study Procedure. The study asks participants to match reconstructions to real images and then rank their confidence in the match. In each question, participants are given the top K=5

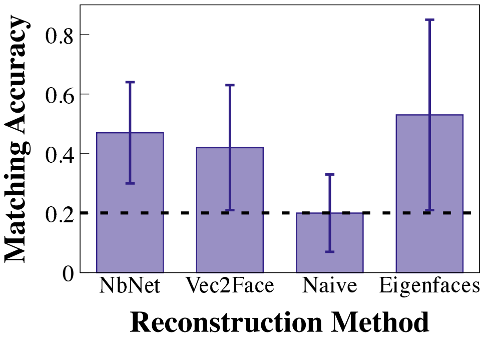

Fig. 7: *Proportion of survey responses with ≥ 1 match per method. The black dashed line represents the probability of a random correct guess.*

matches for $\tilde{x}$ generated using one of the four FVR methods. These are the same top-K matches produced by local models $\mathcal{F}'$ used in the prior section. Participants indicate which (if any) of the top-K images match the identity of the person in $\tilde{x}$ and how confident they are in their choice(s). Confidence options are given on a 5-point Likert scale, ranging from "not confident at all" to "very confident." Participants perform top-K matching on 5 images from each FVR method (20 images total). Random attention checks are included in the survey.

Study Participants. We recruit 200 study participants via Prolific (https://www.prolific.co/). Of our participants, 57% identified as female (39% male; 4% undisclosed). The participants are all 18+ years old spanning multiple age groups: 18-29 (58%), 30-39 (25%), 40-49 (14%), 50-59 (2%), 60+(1%). The survey was designed to take 10 minutes on average, and participants received $2 as compensation.

Study Results. Overall, users have the highest proportion of successful matches (and thus deanonymizations) on `Eigenfaces` reconstructions (Fig. 7). Since $K = 5$, baseline random guessing results in 20% matching accuracy, but participants successfully identify 50% of matches on average. The upper quartile for the matching accuracy of Eigenfaces is 85% – *four times greater than baseline random guessing*. This clearly demonstrates that manual re-identification of the reconstructed images poses a very real threat – a threat that can only increase as more robust FVR methods are developed.

Users are less confident in their responses to `Eigenfaces` questions (Table III) than they are for other successful methods. Respondents have the most confidence in their answers with `NBNet` and `Vec2Face`. More than 90% of respondents were confident to some degree with these two methods, and despite the poorer performance of the naive method, more than 70% of respondents still felt confident about their selection. We conjecture that users feel more confident about their choices for `NBNet` and `Vec2Face` due to the high visual quality of the images produced by these methods (see Figure 3).

VII. Discussion

Key Takeaways. Our results demonstrate that FVR-enabled deanonymization attacks pose a real-world threat. Images reconstructed by the FVR methods we test produce successful top-K matches in secondary facial recognition engines $\mathcal{F}'$, and humans can successfully identify the target identity from these top-K sets.

Beyond this main finding, there are two other takeaways. First, nonparametric reconstruction methods are more flexible and, in our experience, suffer less generalization error than parametric methods. While visual image quality may suffer in the nonparametric setting (e.g. compare `Eigenfaces` to `Vec2Face` in Fig. 3), future improvements may mitigate this problem. Second, evaluations of future FVR methods should focus more on feature space similarity and less on visual similarity to provide more meaningful performance metrics. Since top-K matching precedes visual matching in a practical attack leveraging facial recognition engines, current visual-focused metrics fall short in measuring the real-world threat.

Future Work. The success of our FVR-enabled attacks highlights a few key avenues for future work.

Better defenses against FVR methods. The non-negligible probability of deanonymization shown in our work nullifies the claims that feature vectors can be considered "secure" in any meaningful, cryptographic sense. As such, defenses against these attacks should be developed. State-of-the-art defenses against FVR specifically (not evaluated in this work) assume the attacker uses a parametric reconstruction method [20], [41], [42]. However, we found that the nonparametric FVR method had the highest deanonymization success rate, so future FVR defense work ought to broaden its scope.

Legislation. Work along two lines is needed. First, the implications of face feature vectors as *both* personal *and* biometric data need to be understood. For example, the EU's General Data Protection Regulation (GDPR) defines *biometric data* as "Personal data resulting from specific technical processing ... which allow or confirm [the] unique identification" (Article 4(14)) [8]. Personal and biometric data is subject to more stringent regulations, and – if feature vectors were to be classified as such – practitioners would need to take appropriate measures to mitigate privacy risks associated with plaintext feature vectors. As a second line of work, new legislation governing data privacy laws should include face feature vectors. There is increasing advocacy for the co-design of legislation and computer science [43], [44] and new regulations should account for privacy risks associated with face feature vectors and facial recognition software.

Ethics. Our user study was approved by our local IRB and was designed to maximally protect participant privacy. Furthermore, we conducted our FVR attacks using public face datasets designed for academic research use.

	NBNet	**Vec2Face**	**Naive**	**Eigenfaces**
Confident	0.92	0.94	0.77	0.85
Not Confident	0.08	0.06	0.23	0.15

TABLE III: *Average confidence in survey results from respondents. Though respondents ranked their confidence on a Likert scale, we categorize responses as "confident" (e.g. a response indicating some confidence) or "not confident" for easier presentation.*

References

[1] "Azure face recognition." [Online]. Available: \url{https://azure.microsoft.com/en-us/services/cognitive-services/face/}

[2] "Amazon rekognition." [Online]. Available: \url{https://aws.amazon.com/rekognition}

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016, pp. 770–778.

[4] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *Proc. of USENIX*, August 2020.

[5] I. Evtimov, P. Sturmfels, and T. Kohno, "Foggysight: A scheme for facial lookup privacy," in *Proc. of PETS*, 2021.

[6] Facefirst, "Facefirst's privacy commitment," 2021.

[7] Anyvision, "Ethical and responsible ai at anyvision," 2019.

[8] E. Commission, "General data protection regulation." [Online]. Available: https://gdpr-info.eu/

[9] I. G. Assembly, "(740 ilcs 14/) biometric information privacy act." 2019.

[10] C. N. Duong, T.-D. Truong, K. Luu, K. G. Quach, H. Bui, and K. Roy, "Vec2face: Unveil human faces from their blackbox features in face recognition," in *Proc. of CVPR*, 2020.

[11] A. Zhmoginov and M. Sandler, "Inverting face embeddings with convolutional neural networks," *CoRR*, 2016.

[12] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. of CVPR*, 2015.

[13] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, "On the reconstruction of face images from deep face templates," *IEEE TPMI*, 2019.

[14] A. Razzhigaev, K. Kireev, E. Kaziakhmedov, N. Tursynbek, and A. Petiushko, "Black-box face recovery from identity features," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds., 2020.

[15] M. Gomez-Barrero and J. Galbally, "Reversing the irreversible: A survey on inverse biometrics," *Computers & Security*, 2020.

[16] K. Cao and A. K. Jain, "Learning fingerprint reconstruction: From minutiae to image," *IEEE Transactions on information forensics and security*, 2014.

[17] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Proc. of NeurIPS*, 2019.

[18] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proc. of CVPR*, 2019.

[19] J. Rice-Jones, "Cops, hackers and your mom can unlock your phone with a 3d printed model of your head," 2018.

[20] H. Guo, B. Dolhansky, E. Hsin, P. Dinh, C. C. Ferrer, and S. Wang, "Deep poisoning: Towards robust image data sharing against visual disclosure," in *Proc. of WACV*, 2021.

[21] A. Razzhigaev, K. Kireev, I. Udovichenko, and A. Petiushko, "Darker than black-box: Face reconstruction from similarity queries," *arXiv preprint arXiv:2106.14290*, 2021.

[22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. of CVPR*, 2019, pp. 4690–4699.

[23] "Facial recognition technology: Privacy and accuracy issues related to commericial uses," 2020.

[24] Biostar, "Report: Data breach in biometric security platform affecting millions of users," 2021.

[25] X. Shen, "Facial recognition data leaks are rampant in china as covid-19 pushes wider use of the technology," 2020.

[26] A. Siarohin *et al.*, "First order motion model for image animation," 2021.

[27] Y. Deng and S. Xu, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set —— pytorch implementation," 2021.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, 2004.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of CVPR*, 2018.

[30] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[31] "Code for nbnet reconstruction," 2021. [Online]. Available: https://github.com/csgcmai/NBNet

[32] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. of ICLR*, 2018.

[33] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," 2016.

[34] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of ICML*, 2019.

[35] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. of ICIP*, 2014.

[36] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018.

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014.

[38] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep., 2007.

[39] K. Hill, "The secretative company that may end privacy as we know it," 2020.

[40] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli, "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks," in *Proc. of USENIX*, 2019.

[41] A. Li, J. Guo, H. Yang, and Y. Chen, "Deepobfuscator: Adversarial training framework for privacy-preserving image classification," *CoRR*, vol. abs/1909.04126, 2019.

[42] P. Vepakomma, A. Singh, O. Gupta, and R. Raskar, "Nopeek: Information leakage reduction to share activations in distributed deep learning," in *Proc. of ICDMW*, 2020.

[43] K. Nissim, "Privacy: From database reconstruction to legal theorems," ser. PODS'21. New York, NY, USA: Association for Computing Machinery, 2021, p. 33–41.

[44] K. Nissim, A. Bembenek, A. Wood, M. Bun, M. Gaboardi, U. Gasser, D. O'Brien, T. Steinke, and S. Vadhan, "Bridging the gap between computer science and legal approaches to privacy," vol. 31, Harvard Journal of Law & Technology. Harvard Journal of Law & Technology, 2016 2018, pp. 687–780.