Privacy-Preserving Image Features via Adversarial Affine Subspace Embeddings

Mihai Dusmanu[1] Johannes L. Schönberger[2] Sudipta N. Sinha[2] Marc Pollefeys[1,2]
[1] Department of Computer Science, ETH Zürich [2] Microsoft

Abstract

Many computer vision systems require users to upload image features to the cloud for processing and storage. These features can be exploited to recover sensitive information about the scene or subjects, e.g., by reconstructing the appearance of the original image. To address this privacy concern, we propose a new privacy-preserving feature representation. The core idea of our work is to drop constraints from each feature descriptor by embedding it within an affine subspace containing the original feature as well as adversarial feature samples. Feature matching on the privacy-preserving representation is enabled based on the notion of subspace-to-subspace distance. We experimentally demonstrate the effectiveness of our method and its high practical relevance for the applications of visual localization and mapping as well as face authentication. Compared to the original features, our approach makes it significantly more difficult for an adversary to recover private information.

1. Introduction

Image feature extraction and matching are two fundamental steps in many computer vision applications, such as 3D reconstruction [1, 2], image retrieval [1, 3], or face recognition [4]. Image features can be categorized into low-level [1, 5], mid-level [3, 6] or high-level [7, 8] depending on their information content and receptive field. Furthermore, features can be hand-crafted or learned using data-driven techniques. However, they are almost always represented as vectors in high-dimensional feature spaces. Multiple feature vectors are then compared using appropriate distance metrics, which forms the basis of nearest neighbor search or other retrieval and recognition techniques.

Recently, there has been rapid progress in *feature inversion* methods that reconstruct the image appearance from features extracted in the original image [9, 10, 11, 12] as shown in Figure 1. This raises serious privacy concerns, since images may contain sensitive information about the scene or subjects. Increased awareness of these privacy issues has spurred significant efforts to develop privacy-preserving machine learning systems. In recent years, researchers have proposed a large body of approaches to tackle the various aspects of the problem, including homomorphic cryptosystems [13], differential privacy [14], federated learning [15], and specific solutions for camera localization [16, 17].

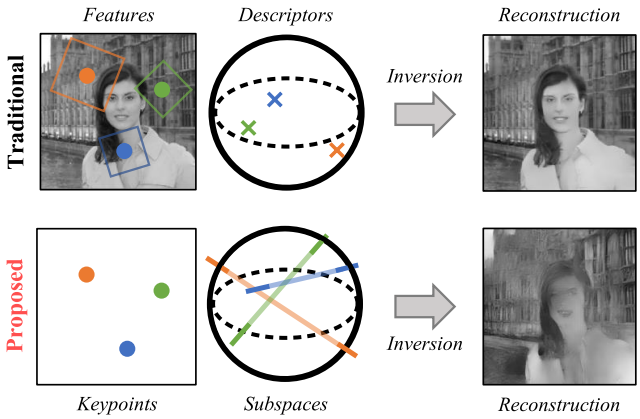

Figure 1: **Privacy-Preserving Image Features.** Inversion of traditional local image features is a privacy concern in many applications. Our proposed approach obfuscates the appearance of the original image by lifting the descriptors to affine subspaces. Distance between the privacy-preserving subspaces enables efficient matching of features. The same concept can be applied to other domains such as face features for biometric authentication. Image credit: *laylam-oran4battersea* (Layla Moran).

In this paper, we propose a new feature representation suitable for visual recognition and matching tasks that makes it significantly more difficult for an adversary to reconstruct the image contents. Our approach has only marginal computational overhead, which makes it amenable to a wide range of practical scenarios. The core idea behind our method is to represent a descriptor point in $\mathbb{R}^n$ as an affine subspace of $\mathbb{R}^n$ passing through the original point. We refer to this process as *lifting*. The chosen dimension of the subspace determines a trade-off between accuracy, runtime, and the level of privacy of the feature representation. To make inverting the representation difficult, we propose a strategy for constructing a lifted subspace containing additional adversarial feature points. We empirically demonstrate strong privacy preservation even for low-dimensional affine subspaces. Pairwise

feature comparison is a fundamental step required in many recognition tasks. In our proposed framework, such comparisons are done directly on the lifted subspaces based on either point-to-subspace or subspace-to-subspace distance.

The paper is organized as follows. First, we formally present the idea of lifting and the technique for matching lifted features. Next, we analyze the performance of these features for two applications: a) image matching for visual localization and mapping as well as b) face authentication. Finally, we demonstrate that our proposed representation is resilient to potential privacy attacks. The code of our method and the evaluation protocol will be released as open-source.

2. Related Work

We first review image features used for applications considered in this paper. We then discuss existing work about privacy attacks on image features and defense mechanisms.

Feature Descriptors. In the traditional local feature extraction paradigm, after keypoint detection and shape estimation, normalized image patches are extracted from images. Feature description takes a patch as input and outputs an n-dimensional vector. Handcrafted local descriptors are based on direct pixel sampling [18] or a histogram of image gradients [1, 5]. Recent advances in deep learning have led to descriptors based on convolutional neural networks (CNNs). Such learnable descriptors are trained using triplet [19] or list-wise [20] losses and hard-negative mining techniques [21]. Local features have been successfully used for tasks such as large-scale 3D reconstruction from crowd-sourced images [2] and image retrieval [22, 23].

Face recognition methods start by face detection and alignment to obtain a canonical face image [24]. Subsequently, a well chosen low-dimensional subspace of pixel-space can provide good recognition performance [4]. More recently, CNN-based features have become the de facto choice for face descriptors. These networks are trained using different classification losses [25, 26, 27].

Feature Subspaces. Wang *et al*. [28] also use a subspace representation for feature matching. Different to their method, we consider *affine* instead of linear subspaces. Accordingly, our distance function is not based on principal angles but on the closest pair of points between the two subspaces. Further, contrary to grouping descriptors of similar patches together to improve matching performance, we add adversarial descriptors to the subspaces to improve privacy.

Feature Inversion and Compromising Privacy. Weinzaepfel *et al*. [9] proposed a method for reconstructing images from local image features using a database of patches with associated descriptors. Dosovitsky and Brox [10, 29] extended on this work by using a CNN and perceptual losses, while Pittaluga *et al*. [30] showed that it was possible to recover detailed images from sparse 3D point clouds reconstructed using structure-from-motion. Similarly, Zhmoginov and Sandler [11] and Mai *et al*. [12] proposed methods for reconstructing face images from their descriptors. Moreover, they showed that the reconstructed images could even be used by an attacker to fool an authentication system.

Privacy-Preserving Methods. *Differential privacy* [14] expands upon Dalenius [31] by formalizing the problem of querying a database without inadvertently releasing information distinguishing the individual entries in the database. An extended overview can be found in [32]. Instead of protecting information leakage from a database, our scenario is quite different in that we are interested in protecting private information in the query as well as contributing new information to a database in a privacy-preserving manner.

McMahan *et al*. [33] introduced *federated learning*, a distributed client-server framework for training a model, where training data remains with the clients, thus offering better privacy guarantees. Kairouz *et al*. [15] reviews the topic and discusses open problems. In contrast, we address a different setting, where tasks require image features computed by clients to be shared with the server. In this context, our approach makes it difficult to recover private image information from the shared features.

Existing works on local features process images encrypted using different homomorphic cryptosystems [34, 35] in the cloud. Jiang *et al*. [36] proposed an alternative by additively splitting the image into two ciphertext matrices using a private prime modulus. These methods guarantee that the original images remain private, but they do not prevent information leakage by inverting the obtained local features. One could also use ℓ_2 distance computation on encrypted feature vectors [37, 38, 39]. However, recent works regarding homomorphic representation search [38, 39] remain computationally expensive, while our method only comes with marginal overhead. Furthermore, these cryptosystems provide security through encryption, where a breach of the secret keys is a privacy risk. In contrast, our system does not have the same single point of failure and provides parameters to trade off accuracy, runtime, and privacy.

Speciale *et al*. [16, 17] proposed solutions tailored to image-based localization, where geometric information is concealed by lifting 2D or 3D points to randomly oriented lines passing through the original locations. Recent work extends on the same idea to solve the full structure-from-motion problem [40, 41]. We draw inspiration from their approach, but instead lift feature descriptors to higher dimensional affine subspaces to conceal appearance information.

3. Method

In this paper, we will represent features from a particular domain as vectors in $\mathbb{R}^n$, where n is the dimensionality of the original feature space. We denote $\mathrm{span}(v_1, \ldots, v_m) =$

$\{\sum_{i=1}^{m} \lambda_i v_i | \lambda_i \in \mathbb{R}\}$ the linear span of a set of vectors $v_i \in \mathbb{R}^n$. An m-dimensional affine subspace $\mathcal{A}$ will be represented as the vector sum of a translation vector a_0 and a linear subspace $\text{span}(a_1, \ldots, a_m)$, giving $\mathcal{A} = a_0 + \text{span}(a_1, \ldots, a_m)$. The core idea of our method is to lift the original feature vector or descriptor $d \in \mathbb{R}^n$ to an m-dimensional affine subspace $\mathcal{D} \subset \mathbb{R}^n$ satisfying $d \in \mathcal{D}$. We denote the lifted affine subspace representation as private features. There are two major requirements that we must address. Firstly, we need to define a distance function that can be used to reliably and efficiently compare two features in this new representation. Secondly, we must construct the affine subspace in a way that effectively conceals the original feature vector d and makes it difficult for an attacker to carry out a successful privacy attack aiming to recover the vector d given the private representation $\mathcal{D}$.

3.1. Distance Functions

Most applications require feature descriptor comparison, which is accomplished using appropriate pairwise distance measures. In our analysis, we restrict ourselves to the Euclidean distance (denoted $\|\cdot\|$) as it is most commonly used in practice. To compute the distance between private features, we either use the point-to-subspace or subspace-to-subspace distance. Note that both distances are upper bound by the original point-to-point distance.

Point-to-Subspace Distance. To compute the distance between a private descriptor d represented as an affine subspace $\mathcal{D}$ and a regular descriptor e, one can use the point-to-subspace distance defined as:

$$\text{dist}(\mathcal{D}, e) = \min_{x \in \mathcal{D}} \|e - x\| = \|e - p_{\perp}^{\mathcal{D}}(e)\| \ , \qquad (1)$$

where $p_{\perp}^{\mathcal{D}}(e)$ denotes the orthogonal projection of e onto $\mathcal{D}$.

Subspace-to-Subspace Distance. To compute the distance between two private descriptors d, e represented as affine subspaces $\mathcal{D}, \mathcal{E}$ of dimensions $m_{\mathcal{D}}, m_{\mathcal{E}}$, one can use the subspace-to-subspace distance defined as:

$$\text{dist}(\mathcal{D}, \mathcal{E}) = \min_{x \in \mathcal{D}, y \in \mathcal{E}} \|y - x\| \ . \qquad (2)$$

Let us denote a closest pair of points in the two subspaces as $x^* \in \mathcal{D}$ and $y^* \in \mathcal{E}$, respectively. Then, we have:

$$x^* = d_0 + \sum_{i=1}^{m_{\mathcal{D}}} \alpha_i d_i \ , y^* = e_0 + \sum_{i=1}^{m_{\mathcal{E}}} \beta_i e_i \ , \qquad (3)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{m_{\mathcal{D}}}, \boldsymbol{\beta} \in \mathbb{R}^{m_{\mathcal{E}}}$. In the following derivation, we assume that both subspaces have the same dimension $m = m_{\mathcal{D}} = m_{\mathcal{E}}$ for simplicity. A sufficient and necessary condition for $\text{dist}(\mathcal{D}, \mathcal{E}) = \|y^* - x^*\|$ is that the line $y^* - x^*$ is orthogonal to both $\mathcal{D}$ and $\mathcal{E}$:

$$\begin{cases} (y^* - x^*)^T d_i = 0 \\ (y^* - x^*)^T e_i = 0 \end{cases} \ , \qquad (4)$$

which can be rewritten as:

$$\begin{cases} (e_0 - d_0)^T d_i = \sum_{j=1}^{m} \alpha_j d_i^T d_j + \sum_{j=1}^{m} \beta_j (-d_i^T e_j) \\ (e_0 - d_0)^T e_i = \sum_{j=1}^{m} \alpha_j e_i^T d_j + \sum_{j=1}^{m} \beta_j (-e_i^T e_j) \end{cases} . \qquad (5)$$

This system can be formulated in a more compact form:

$$\begin{bmatrix} DD^T & -DE^T \\ ED^T & -EE^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} D \\ E \end{bmatrix} (e_0 - d_0) \ , \qquad (6)$$

where $D = \begin{bmatrix} d_1 \ldots d_m \end{bmatrix}^T, E = \begin{bmatrix} e_1 \ldots e_m \end{bmatrix}^T \in M_{m \times n}(\mathbb{R})$.

If the bases of the subspaces are orthonormal ($DD^T = EE^T = I$), the system further simplifies to:

$$\begin{bmatrix} I & -DE^T \\ ED^T & -I \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} D \\ E \end{bmatrix} (e_0 - d_0) \ . \qquad (7)$$

Thus, finding the subspace-to-subspace distance requires solving a linear system with $2m$ unknowns and equations. Let $M = -DE^T$. Under the assumption that the matrix $N = I - MM^T$ is invertible, the block-matrix inversion formula can be used to rewrite Eq. 7 as follows:

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} N^{-1} & N^{-1}M \\ -M^T N^{-1} & -M^T N^{-1} M - I \end{bmatrix} \begin{bmatrix} D \\ E \end{bmatrix} (e_0 - d_0) \ . \qquad (8)$$

The solutions to $\boldsymbol{\alpha}, \boldsymbol{\beta}$ can be substituted into Eq. 3 to obtain the subspace-to-subspace distance as $\|x^* - y^*\|$. Note that the problem can also be formulated using the dual representation of a subspace as the intersection of $n - m$ hyperplanes. We provide a derivation of the dual formulation and further discussion in the supplementary material.

3.2. Affine Subspace Embedding

Each subspace embedding is defined by a translation vector d_0 and a basis $\{d_1, \ldots, d_m\}$. The choice of these and the distribution of the original descriptors has significant impact on the effectiveness of our approach and the required dimensionality m of the subspace to achieve sufficient privacy preservation. For example, it is common practice to ℓ_2-normalize descriptors [1, 21, 26, 27]. In such cases, lifting descriptors to affine lines ($m = 1$) is not secure. This is because a line intersects the unit hyper-sphere in at most 2 points. It can be easy to detect which of the two intersections is statistically plausible and thereby exactly recover the original point. However, any value of $m > 1$ generally produces infinite intersection points and thus provides much more ambiguity which is desirable for privacy preservation. We now describe different lifting strategies, which we later compare in our experimental evaluation.

Random Basis. One could sample random direction vectors for the linear subspace, *i.e.*, $d_i \sim \mathcal{U}([-1, 1])^n$ referred to as *random lifting*. In our experiments, we found this approach to be vulnerable to relatively simple privacy attacks.

The original descriptor can be approximated by the nearest entry from a database of real-world descriptors according to the point-to-subspace distance. This is because random subspaces generally intersect the descriptor manifold once.

Adversarial Basis. To address this issue, one can ensure that the subspace passes through multiple regions of the descriptor manifold. We propose to use a database of real-world descriptors $W = \{w_1, \dots, w_s\}$ as an approximation of the manifold and sample the basis vectors as $d_i = w_i - d$, where $w_i \sim \mathcal{U}(W)$. We call this approach *adversarial lifting*, as it intentionally introduces plausible samples in the subspace to conceal the original descriptor. Moreover, a defender can choose adversarial samples to hide specific private information, *e.g.*, to hide the gender of a person, one can pick a feature vector from another gender, as shown in our experimental evaluation. Adversarial sampling improves privacy but reduces descriptor matching accuracy, because the chance of accidental subspace intersections increases. To balance the accuracy and privacy trade-offs, we propose combining the adversarial and random lifting strategies, which we call *hybrid lifting*. In hybrid lifting, a subset of the basis vectors are selected randomly while the rest are chosen using adversarial sampling. There are different ways to implement the adversarial and hybrid strategies depending on the task at hand. We describe a few such variants in the context of local features and face descriptors in Section 4.

Translation Vector. The origin can be set to any point in the subspace, except for the vector d itself, since it is precisely what we must conceal. Thus, we sample a random point and project it to the subspace, as follows:

$$d_0 = p_{\perp}^{d+\mathrm{span}(d_1,\dots,d_m)}(e) \text{ where } e \sim \mathcal{U}([-1,1])^n \ . \quad (9)$$

Information Leakage. It is important to carefully construct the subspace to avoid accidental leakage of information. For instance, in the adversarial formulation described above, all basis vectors (d_i for $i > 0$) point "away from" the initial descriptor d. An attacker could target parts of the descriptor manifold where these directions are feasible. More precisely, one could look for real-world descriptors $\tilde{d}$ such that $\tilde{d} + \lambda d_i$ also intersects the descriptor manifold. To mitigate this, given an initial subspace $\mathcal{D}$, we generate a random basis as:

$$d_i = p_{\perp}^{\mathcal{D}}(e_i) - d_0 \text{ where } e_i \sim \mathcal{U}([-1,1])^n, \forall i \ . \quad (10)$$

4. Experimental Evaluation

In this section, we evaluate our method on two applications. First, we experiment with local features on the task of image matching for visual localization and mapping. Second, we apply our method to global image features for face verification. We report results in these two settings and assess the trade-offs between the degree of privacy preservation achieved, the accuracy of the target task and the computational complexity. As we cannot provide any theoretical guarantees on privacy preservation, we implement plausible privacy attacks and empirically demonstrate that our approach is robust against them.

Dist.	Time (ms)	Subspace dimension 2	4	8
s-to-s	GPU	2.02 ± 0.14	6.02 ± 0.14	N/A
	CPU	107.87 ± 0.95	195.50 ± 2.02	540.98 ± 25.18
p-to-s	GPU	2.02 ± 0.14	2.10 ± 0.30	4.17 ± 0.38
	CPU	25.25 ± 1.14	37.71 ± 0.55	63.24 ± 1.08

Table 1: **Runtime.** We report the average runtime over 100 runs of the distance matrix computation for an image pair, when varying the lifting dimension. Each image has 1000 128-dimensional floating point features. We consider both the subspace-to-subspace (s-to-s) and the point-to-subspace (p-to-s) distance. For the former, we implemented specialized CUDA solvers for lifting dimensions 2 and 4. Hardware: NVIDIA RTX 2080Ti, Intel Core i9-9900K.

4.1. Runtime

Previous approaches to privacy-preserving descriptors take advantage of homomorphic encryption. While these methods guarantee an exact distance computation, they are severely limited in terms of practical applicability, especially in real-time scenarios. A recent work about encrypted representation search [39] reports that computing the distances between a single 128-dimensional query vector and a database with 1000 entries takes around 1 second (*c.f.* Figure 3 [39]). Thus, obtaining the full distance matrix for an image pair with 1000 features each would take around 16 minutes. In comparison, our method only induces minimal computational overhead, as shown in Table 1. For completeness, the runtime for computing the point-to-point distance matrix in the same setting is 1.01 ± 0.10ms on GPU and 1.05 ± 0.46ms on CPU, respectively.

4.2. Local Feature Descriptors

In order to demonstrate the robustness and generalizability of our approach, we perform experiments using the arguably most popular hand-crafted local feature (SIFT [1]) as well as a recent state-of-the-art learned descriptor (HardNet [21]). Both descriptors are by default ℓ_2-normalized. We evaluate the private descriptors on the tasks of image matching, structure-from-motion and visual localization.

Subspace Selection. The adversarial lifting database is obtained by clustering 10 million local features from 60,000 images of the Places365 dataset [43] into $s = 256{,}000$ clusters using spherical k-means [44]. In the context of 3D computer vision tasks, it is usually desirable to have many thousands of features per image [45]. Let us consider the

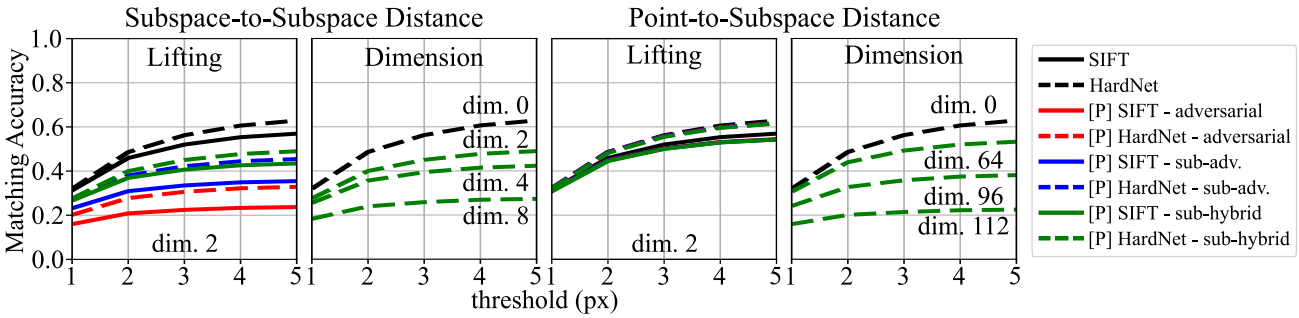

Figure 2: **Matching evaluation.** We plot the mean matching accuracy at different thresholds on the HPatches sequences [42]. Methods using our private representation are prefixed by [P]. We report results with different lifting methods and dimensions. HardNet outperforms SIFT on this benchmark and the ordering is respected after lifting as well.

case of lifting to descriptor planes ($m = 2$) using uniform random sampling from the database of 256,000 centroids. Given an image pair with 8,000 descriptors per image, for each feature in the second image, there is a $1/16$ chance of sampling a feature already selected in the first one. Such a collision causes subspace intersections and thus leads to wrong feature matches. This is further exacerbated by typical match filtering strategies (*e.g.*, mutual check, ratio-test [1]).

To reduce the number of wrong matches, we randomly split the database into S pairwise disjoint sub-databases $W_1, \dots, W_S$ satisfying $W = \cup_{i=1}^{S} W_i, \text{card}(W_i) = s/S$. For an image I, we then first randomly select a sub-database $\mathcal{W} \sim \mathcal{U}(\{W_1, \dots, W_S\})$. Next, the basis vectors are generated using only elements of $\mathcal{W}$, *i.e.*, $v_i = w_i - d$, where $w_i \sim \mathcal{U}(\mathcal{W})$. If two images select different sub-databases in this *sub-adversarial lifting* strategy, the probability of random collision is 0. For images using the same sub-database, the number of collisions is very high. Overall, with this strategy, instead of degrading the matching performance for all image pairs, we achieve good matching performance in $15/16$ cases for $S = 16$. In addition, we also evaluate a *sub-hybrid lifting* strategy, where half of the basis vectors are random and the other half uses a sub-database.

Image Matching. We compare raw descriptors with their private counterparts on the image sequences from the HPatches dataset [42]. This dataset consists of 116 scenes with 6 images each: 57 of them exhibit illumination changes, while the other 59 show significant viewpoint changes. For each scene, we match the first image against the other 5 yielding 580 image pairs in total. For evaluation, we follow protocol introduced by Dusmanu *et al*. [46] which reports the mean matching accuracy of a mutual nearest neighbors matcher for different values of the threshold up to which a match is considered correct.

Figure 2 shows results for both distances with different lifting methods and dimensions. Random lifting is not plotted as it performs identical with the raw descriptors. As mentioned above, adversarial lifting performs poorly for local features due to subspace collisions. This is, in part, addressed by the use of sub-databases and further improved by sub-hybrid lifting. The point-to-subspace distance only preserves the privacy of one image and is useful for cloud- and client-based visual localization systems, equivalent to Speciale *et al*. [16, 17]. This approach is able to achieve good matching performance even for very high lifting dimensions.

Structure-from-Motion. Next, we integrate the best performing private representation from above (sub-hybrid lifting) into an end-to-end 3D reconstruction pipeline [47] and evaluate it on the crowd-sourced 3D reconstruction benchmark of Schönberger *et al*. [45]. For each image, we retrieve the top 50 most similar images using NetVLAD [48] and only match against these. Next, we run geometric verification (with a minimum inlier ratio of 0.1) followed by sparse reconstruction using COLMAP [47, 49] and finally report the reconstruction statistics in Table 2. For this evaluation, we preserve the privacy of all input images. As already observed in our image matching evaluation, the private features come with accuracy trade-offs. As we increase the dimensionality of the subspace, the reconstruction completeness degrades accordingly. Despite the fewer number of registered images, the 3D models remain relatively accurate and clearly distinguishable. The generally lower track length for private features is caused by missing matches leading to longer feature tracks being split into multiple smaller ones.

Visual Localization. We also consider the case of localizing to an already built map on the challenging Aachen Day-Night long-term visual localization dataset [50]. This is equivalent to the scenario tackled by Speciale *et al*. [17], where the goal is to protect the privacy of users of an image-based localization service, such as Google Visual Positioning System [51] or Microsoft Azure Spatial Anchors [52]. We first triangulate the database model from the given camera poses and intrinsics using DoG keypoints with raw SIFT and HardNet descriptors, respectively. For each query image (824 day-time and 98 night-time), we retrieve the top 50 database images using NetVLAD [48]. We preserve the

Dataset	Method	Reg. images	Sparse points	Track length	Reproj. error
Madrid Metropolis 1344 images	SIFT	400	28,862	7.01	0.72
	[P] SIFT - dim. 2	302	17,232	6.37	0.59
	[P] SIFT - dim. 4	227	11,461	5.54	0.56
	HardNet	459	42,180	7.25	0.89
	[P] HardNet - dim. 2	367	28,367	6.49	0.68
	[P] HardNet - dim. 4	268	15,562	6.32	0.58
Gendarmen-markt 1463 images	SIFT	896	74,348	6.37	0.84
	[P] SIFT - dim. 2	783	64,554	5.44	0.71
	[P] SIFT - dim. 4	458	33,291	5.23	0.60
	HardNet	999	112,245	6.68	0.96
	[P] HardNet - dim. 2	864	89,865	5.98	0.80
	[P] HardNet - dim. 4	751	63,862	5.50	0.69
Tower of London 1576 images	SIFT	635	64,490	7.78	0.70
	[P] SIFT - dim. 2	525	55,439	6.58	0.61
	[P] SIFT - dim. 4	439	37,819	6.10	0.56
	HardNet	749	89,818	7.85	0.81
	[P] HardNet - dim. 2	557	69,161	7.19	0.68
	[P] HardNet - dim. 4	498	49,570	6.69	0.61

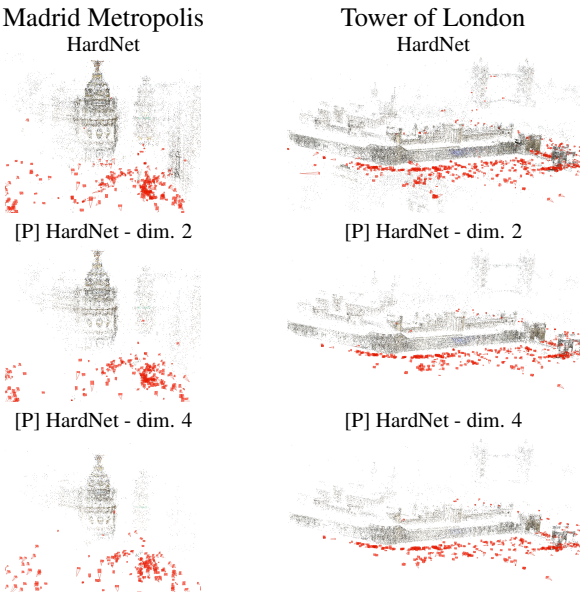

Table 2: **Local Feature Evaluation Benchmark.** We report reconstruction statistics such as the number of registered images and sparse points and the average track length and reprojection error on internet photo collections of landmarks [45]. Methods prefixed by [P] use sub-hybrid lifting for all features of input images. On the right side, we visualize the final sparse models.

Query	Method	Thresholds 0.25m, 2°	0.5m, 5°	5.0m, 10°
Day (824)	SIFT	82.9%	89.6%	92.2%
	[P] SIFT - dim. 2	79.5%	87.0%	91.1%
	[P] SIFT - dim. 4	79.6%	86.5%	91.1%
	[P] SIFT - dim. 16	76.7%	84.0%	87.4%
	HardNet	86.3%	92.5%	95.6%
	[P] HardNet - dim. 2	84.3%	89.8%	94.3%
	[P] HardNet - dim. 4	83.5%	90.2%	93.6%
	[P] HardNet - dim. 16	82.0%	88.3%	92.2%
Night (98)	SIFT	41.8%	48.0%	55.1%
	[P] SIFT - dim. 2	32.7%	36.7%	42.9%
	[P] SIFT - dim. 4	32.7%	38.8%	43.9%
	[P] SIFT - dim. 16	25.5%	31.6%	34.7%
	HardNet	60.2%	67.3%	73.5%
	[P] HardNet - dim. 2	49.0%	53.1%	58.2%
	[P] HardNet - dim. 4	40.8%	44.9%	49.0%
	[P] HardNet - dim. 16	32.7%	37.8%	43.9%

Table 3: **Aachen Day-Night Localization Challenge.** We report the percentage of localized query images for both day and night scenarios under different camera pose accuracy threshold on the Aachen Day-Night dataset [50]. For the private methods (prefixed by [P]), we use sub-hybrid lifting for query images and point-to-subspace distance for matching.

privacy of all query images with sub-hybrid lifting and use point-to-subspace distance for matching. Finally, we use the COLMAP [47] image registrator with fixed intrinsics to obtain poses that are submitted to the long-term visual localization benchmark [53].

Following the standard evaluation protocol, we report the percentage of localized query images for different real-world thresholds in Table 3. On the day queries, we are able to achieve competitive performance even when lifting to 16 dimensional subspaces. As previously, the accuracy gradually decreases when increasing the lifting dimension. Furthermore, even on the extremely hard night-to-day matching queries where pose estimation has very low inlier ratios, we are still able to localize a reasonable number of queries.

Privacy Attack. To analyze attacks on the proposed private descriptor representation, we provide the adversary with multiple tools. We assume that they have access to a database V of 128,000 real-world descriptors built using the same procedure as the lifting database (described above). Further, the attacker has unrestricted access to the lifting algorithm and is able to use it on-demand. Finally, they have access to extensive training data (the MegaDepth [54] dataset) as well as the architecture and loss from Pittaluga *et al*. [30] allowing them to train new feature inversion networks.

First, we consider a nearest neighbor attack (NNA) where each subspace is approximated by its closest correspondence from a database of real-world descriptors. Formally, for each private representation $\mathcal{D}$ associated to a descriptor d, the database V is used to retrieve the closest element to the subspace $\tilde{d} = \arg\min_{v \in V} \text{dist}(\mathcal{D}, v)$. Next, the approximated descriptors $\tilde{d}$ can be fed to a regular feature inversion network to reconstruct the appearance of the original image.

Second, we consider a direct inversion attack (DIA) where the affine subspaces are provided as input to a CNN. To

Attack	Lifting	Dim.	MAE (↓)	SSIM (↑)	PSNR (↑)
	raw	0	0.105	0.755	17.937
NNA	random	2	0.112	0.738	17.448
NNA	sub-hybrid	2	0.206	0.530	12.288
DIA	sub-hybrid	2	0.176	0.648	13.447
DIA	sub-hybrid	4	0.179	0.594	13.531
DIA	sub-hybrid	6	0.194	0.559	12.823

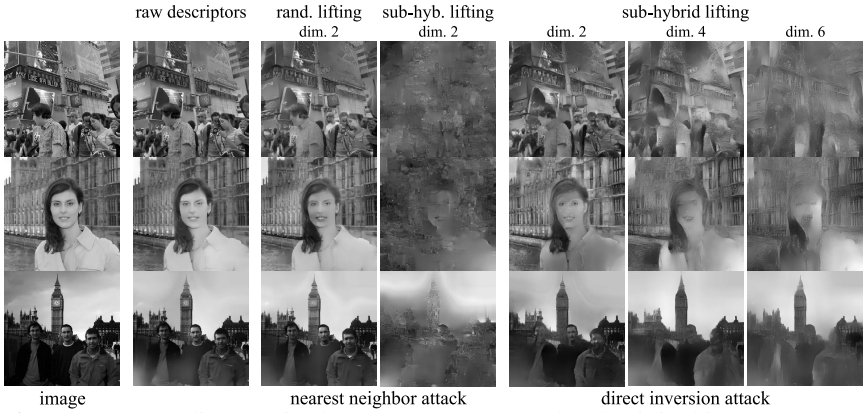

Figure 3: **Image reconstruction.** On the left, we report quality metrics between reconstructed and original images. On the right, we show several qualitative examples: first original image, then reconstructions from the raw descriptors and using the proposed privacy attacks on different lifting methods and dimensions. Image credit (top to bottom): *pagedooley* (Kevin Dooley), *laylamoran4battersea* (Layla Moran), *martinalvarez* (Martin Alvarez Espinar).

this end, we train multiple feature inversion networks from Difference-of-Gaussians (DoG) keypoints and private descriptors lifted to 2, 4, and 6 dimensions, respectively. Note that the architectures proposed in previous works [30, 10] are very compute and memory intensive – training them on higher dimensional subspaces would be a challenge in itself.

We run the proposed privacy attacks on 10 images[1] using HardNet descriptors and present the results in Figure 3. On the left, we quantitatively report image reconstruction quality metrics such as mean absolute error (MAE), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR); on the right, we show qualitative image reconstructions. Please refer to the supplementary material for more examples. Using the raw descriptors, one can reconstruct the original image with very high fidelity (note the readability of text in the first example). The nearest neighbor attack is successful on private features using random lifting, but not when using sub-hybrid lifting due to the adversarial samples. For all reconstructions, the general outline of the buildings is recovered mainly due to the lack of features in the sky (*e.g.*, third example). The direct inversion attack is able to reconstruct some parts of the original image, but the quality is significantly deteriorated. Furthermore, distinguishing details such as faces or text are heavily perturbed and become non-existent for higher lifting dimensions.

4.3. Face Descriptors

For this evaluation, we use a state-of-the-art deep face descriptor – the best performing ArcFace [27] model with a ResNet-101 [55] backbone trained on MS-Celeb-1M [56].

Face Verification. We report face verification accuracy on multiple datasets: LFW [57], CFP [58] (both frontal-frontal denoted FF and frontal-profile denoted FP), and AgeDB-30 [59]. We follow the regular evaluation protocol, notably 10-fold cross validation where, for each fold, the training split is used to determine a distance threshold that separates between same / different identity and the accuracy is computed on the validation split. Finally, the mean classification accuracy over the 10 folds is reported in Figure 4.

We evaluate two scenarios: point-to-subspace (*p-to-s*) matching, where one of the images is represented using the original descriptor and the other one is lifted to a subspace, and subspace-to-subspace (*s-to-s*) matching, where both descriptors are private. As expected, the point-to-subspace matching performs better across the board. For the subspace-to-subspace distance, the performance on the simple datasets (LFW and CFP-FF) only drops by a few percents. For more complex datasets (frontal-profile matching in CFP-FP, large age differences in AgeDB-30), the performance drop is more significant. Nevertheless, the simpler datasets are still very representative of common authentication systems (Microsoft Windows Hello [60], Apple Face ID [61]), making our approach highly relevant for such scenarios.

Privacy Attack. The privacy attack we are concerned with involves inferring distinguishing properties (gender, race) from only the ArcFace [27] descriptors. For this purpose, we used FairFace [62], a face dataset consisting of 97,698 images with balanced gender (2 classes) and race (7 classes) annotations. We randomly selected 10,000 training images for the database needed by our lifting method. The remaining 76,744 training images were used for the attack. The validation set of 10,954 images is used for evaluation.

We attack an ArcFace descriptor using a K-nearest neighbors (K-NN) classifier [44] to predict the gender and race of the person. We do this both on the original feature as well as the lifted feature for $K = 10$. We also implemented a variant of our hybrid lifting method (denoted *hybrid+*) that

[1]We manually selected 2 holiday images from Hong Kong, London, New York, Paris, and Tokyo published on Flickr under a CC BY 4.0 License.

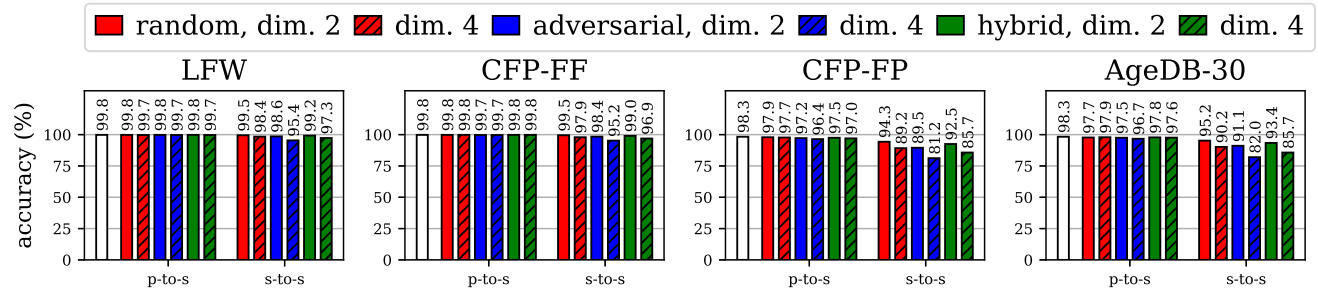

Figure 4: **Face verification.** We show the accuracy on different face verification datasets. The white bar represents the reference accuracy of raw ArcFace descriptors. The point-to-subspace distance (*p-to-s*), performs within at most 2% of the original descriptors. For the subspace-to-subspace distance (*s-to-s*), the performance drop is more significant in the difficult scenarios (CFP-FP and AgeDB-30), but frontal authentication (LFW and CFP-FF) is still very accurate (95% at worst).

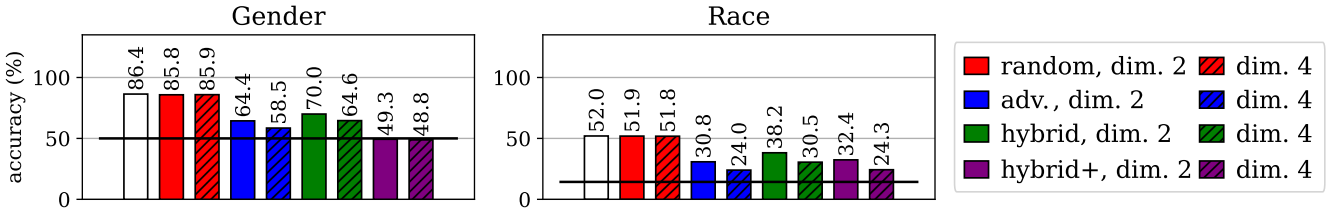

Figure 5: **FairFace.** We report the accuracy of a K-NN classifier trained to predict the gender and race of a subject from their ArcFace descriptor. The black line represents the approximate accuracy of a random classifier. The white bars represent the accuracy on raw ArcFace descriptors. Private representations using a database for lifting successfully conceal information.

exploits the gender / race of each person. In this variant, each feature is lifted by sampling database entries with a different gender / of a different race to obtain a balanced subspace, which better conceals these attributes.

The results are reported in Figure 5. The black vertical lines denote the approximate performance of a random classifier. Similar to image matching, pure random lifting is again not effective at concealing the private attributes. Adversarial lifting has the best results in terms of privacy, but its face verification accuracy is also the worst. Hybrid lifting offers a trade-off between random lifting (high performance) and adversarial lifting (good for privacy). Finally, the hybrid+ version is most effective at concealing the gender.

5. Limitations and Future Work

Speciale *et al*. [17] showed that solving the target task of camera localization reveals the concealed location of some features in the query image. Similarly, in our 3D reconstruction task, the pair of closest points on two matched affine subspaces provides a way to estimate the concealed feature descriptors. This implies that features associated with 3D points triangulated from multiple views are likely to be revealed. By inverting the estimated descriptors, an adversary might be able to approximately reconstruct the appearance of the stationary part of the scene. However, this is not a serious limitation, as feature descriptors extracted from image regions depicting people or other transient objects will generally not be matched in multiple overlapping images and therefore their appearance is unlikely to be revealed.

For face verification, it is possible to infer the face descriptor after repeated authentications of a person if a history of the private descriptors is stored. One potential mitigation is to generate near parallel subspaces for a particular individual, although it is unclear how this approach behaves with respect to the manifold of face descriptors. A potential option would be adding a trusted third-party in the system that receives private descriptors from both client and server and computes the distances without storing any data.

Apart from addressing these limitations, other directions for future work include training descriptors more suitable for lifting and implementing scalable matching inspired by prior work on subspace representations [63] to enable large-scale applications such as place recognition.

6. Conclusion

We have proposed a novel privacy-preserving feature representation by embedding feature descriptors into affine subspaces containing adversarial samples. To find similar features, nearest neighbor computation is enabled through point-to-subspace or subspace-to-subspace distance. We experimentally demonstrate the high practical relevance of our approach for crowd-sourced visual localization and mapping as well as face authentication, while rendering it difficult to recover sensitive information.

Acknowledgements. This work was supported by the Microsoft Mixed Reality & AI Zürich Lab PhD scholarship.

References

[1] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski. Building Rome in a Day. *Communications of the ACM*, 2011.

[3] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

[4] Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 1991.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[6] Y-Lan Boureau, Francis Bach, Yann Le Cun, and Jean Ponce. Learning mid-level features for recognition. In *Proc. CVPR*, 2010.

[7] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. In *Proc. CVPR Workshops*, 2014.

[8] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *Proc. ECCV*, 2014.

[9] Philippe Weinzaepfel, Hervé Jégou, and Patrick Pérez. Reconstructing an image from its local descriptors. In *Proc. CVPR*, 2011.

[10] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proc. CVPR*, 2016.

[11] Andrey Zhmoginov and Mark Sandler. Inverting face embeddings with convolutional neural networks. *arXiv*, 2016.

[12] Guangcan Mai, Kai Cao, Pong C. Yuen, and Anil K. Jain. On the reconstruction of face images from deep face templates. *IEEE PAMI*, 2018.

[13] Paillier, Pascal. Public-key Cryptosystems Based on Composite Degree Residuosity Classes. In *Proc. EuroCrypt*, 1999.

[14] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. TCC*, 2006.

[15] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and Open Problems in Federated Learning. *arXiv*, 2019.

[16] Pablo Speciale, Johannes L. Schonberger, Sing Bing Kang, Sudipta N. Sinha, and Marc Pollefeys. Privacy Preserving Image-Based Localization. In *Proc. CVPR*, 2019.

[17] Pablo Speciale, Johannes L. Schonberger, Sudipta N. Sinha, and Marc Pollefeys. Privacy Preserving Image Queries for Camera Localization. In *Proc. ICCV*, 2019.

[18] Michael Calonder, Vincent Lepetit, Mustafa Ozuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua. BRIEF: Computing a local binary descriptor very fast. *IEEE PAMI*, 2011.

[19] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proc. BMVC.*, 2016.

[20] Kun He, Yan Lu, and Stan Sclaroff. Local Descriptors Optimized for Average Precision. In *Proc. CVPR*, 2018.

[21] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in NeurIPS*, 2017.

[22] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010.

[23] Johannes Lutz Schönberger, Filip Radenović, Ondrej Chum, and Jan-Michael Frahm. From Single Image Query to Detailed 3D Reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 2016.

[25] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014.

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *Proc. CVPR*, 2017.

[27] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proc. CVPR*, 2019.

[28] Zhenhua Wang, Bin Fan, and Fuchao Wu. Affine Subspace Representation for Feature Description. In *Proc. ECCV*, 2014.

[29] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in NeurIPS*, 2016.

[30] Francesco Pittaluga, Sanjeev J Koppal, Sing Bing Kang, and Sudipta N Sinha. Revealing Scenes by Inverting Structure From Motion Reconstructions. In *Proc. CVPR*, 2019.

[31] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 1977.

[32] Cynthia Dwork. Differential Privacy: A Survey of Results. In *Proc. TAMC*, 2008.

[33] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *arXiv*, 2016.

[34] Zhan Qin, Jingbo Yan, Kui Ren, Chang Wen Chen, and Cong Wang. Towards efficient privacy-preserving image feature extraction in cloud computing. In *Proc. ACMM*, 2014.

[35] Chao-Yung Hsu, Chun-Shien Lu, and Soo-Chang Pei. Image feature extraction in encrypted domain with privacy-preserving SIFT. *IEEE Transactions on Image Processing*, 2012.

[36] Linzhi Jiang, Chunxiang Xu, Xiaofang Wang, Bo Luo, and Huaqun Wang. Secure outsourcing SIFT: Efficient and privacy-preserving image feature extraction in the encrypted domain. *IEEE Transactions on Dependable and Secure Computing*, 2017.

[37] Taeyun Kim, Yongwoo Oh, and Hyoungshick Kim. Efficient Privacy-Preserving Fingerprint-Based Authentication System Using Fully Homomorphic Encryption. *Security and Communication Networks*, 2020.

[38] Vishnu Naresh Boddeti. Secure face matching using fully homomorphic encryption. In *International Conference on Biometrics Theory, Applications and Systems*, 2018.

[39] Joshua J Engelsma, Anil K Jain, and Vishnu Naresh Boddeti. Hers: Homomorphically encrypted representation search. *arXiv*, 2020.

[40] Marcel Geppert, Viktor Larsson, Pablo Speciale, Johannes Lutz Schönberger, and Marc Pollefeys. Privacy Preserving Structure-from-Motion. In *Proc. ECCV*, 2020.

[41] Mikiya Shibuya, Shinya Sumikura, and Ken Sakurada. Privacy preserving visual SLAM. In *Proc. ECCV*, 2020.

[42] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proc. CVPR*, 2017.

[43] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE PAMI*, 2017.

[44] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[45] Johannes L. Schönberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proc. CVPR*, 2017.

[46] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Proc. CVPR*, 2019.

[47] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proc. CVPR*, 2016.

[48] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proc. CVPR*, 2016.

[49] Johannes Lutz Schönberger, True Price, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. A Vote-and-Verify Strategy for Fast Spatial Verification in Image Retrieval. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[50] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DoF outdoor visual localization in changing conditions. In *Proc. CVPR*, 2018.

[51] Tilman Reinhardt. Google Visual Positioning Service. https://ai.googleblog.com/2019/02/using-global-localization-to-improve.html, 2019.

[52] Neena Kamath. Announcing Azure Spatial Anchors for collaborative, cross-platform mixed reality apps. https://azure.microsoft.com/en-us/blog/announcing-azure-spatial-anchors-for-collaborative-cross-platform-mixed-reality-apps/, 2019.

[53] Long-Term Visual Localization Benchmark. https://www.visuallocalization.net.

[54] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR*, 2018.

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[56] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World. *Journal of Electronic Imaging*, 2016.

[57] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[58] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *Proc. WACV*, 2016.

[59] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: The First Manually Collected, In-The-Wild Age Database. In *Proc. CVPR Workshops*, 2017.

[60] Windows Hello. https://blogs.windows.com/windowsexperience/2015/07/25/say-hello-to-windows-hello-on-windows-10/.

[61] Face ID. https://support.apple.com/en-us/HT208108.

[62] Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv*, 2019.

[63] Ronen Basri, Tal Hassner, and Lihi Zelnik-Manor. Approximate Nearest Subspace Search with Applications to Pattern Recognition. In *Proc. CVPR*, 2007.