Privacy-Preserving Image Features via Adversarial Affine Subspace Embeddings Supplementary Material

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

This document contains the following supplementary information. First, we describe the dual formulation for the point-to-subspace and subspace-to-subspace distances. Next, we discuss the space and time complexity of our matching algorithm. Finally, we show more quantitative and qualitative results of privacy attacks on local features in the scenario of an image-based localization service.

1. Dual Formulation

Alternative to our formulation in the main paper, an m-dimensional linear subspace of $\mathbb{R}^n$ can also be interpreted as the intersection of $n - m$ hyperplanes. Under this formulation, an affine subspace can be defined by the sum of a translation vector a_0 and the orthogonal subspace of the linear span of $a_1, \ldots, a_{n-m}$, *i.e.*, $\mathcal{A} = a_0 + \text{span}(a_1, \ldots, a_{n-m})^\perp$. Throughout the entire section, we suppose that $(a_1, \ldots, a_{n-m})$ is orthonormal, *i.e.*, that $A = [a_1 \ldots a_{n-m}]^T$ satisfies $AA^T = I$.

We consider two affine subspaces $\mathcal{D}, \mathcal{E}$ under this representation. Let $(x^*, y^*) \in \mathcal{D} \times \mathcal{E}$ be a solution of the subspace-to-subspace distance, *i.e.*, $\|y^* - x^*\| = \min_{x \in \mathcal{D}, y \in \mathcal{E}} \|y - x\|$. As before, 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 $\mathcal{E}$, *i.e.*, there exist $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^{n-m}$ such that $y^* - x^* = \sum_{j=1}^{n-m} \mu_j d_j = \sum_{j=1}^{n-m} \nu_j e_j$. Finally, $x^*, y^*, \boldsymbol{\mu}, \boldsymbol{\nu}$ must satisfy the following constraints:

$$\begin{cases} y^* - x^* = \sum_{j=1}^{n-m} \mu_j d_j = \sum_{j=1}^{n-m} \nu_j e_j \\ d_i^T (x^* - d_0) = 0 \\ e_i^T (y^* - e_0) = 0 \end{cases} \tag{1}$$

which can be rewritten as:

$$\begin{cases} \sum_{j=1}^{n-m} \mu_j d_j - \sum_{j=1}^{n-m} \nu_j e_j = \mathbf{0} \\ d_i^T x^* = d_i^T d_0 \\ e_i^T x^* + \sum_{j=1}^{n-m} \nu_j e_i^T e_j = e_i^T e_0 \end{cases} . \tag{2}$$

This system can be represented under the following form:

$$\begin{bmatrix} D & \mathbf{0}_{(n-m)^2} & \mathbf{0}_{(n-m)^2} \\ E & \mathbf{0}_{(n-m)^2} & I \\ \mathbf{0}_{n^2} & D^T & -E^T \end{bmatrix} \begin{bmatrix} x^* \\ \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} Dd_0 \\ Ee_0 \\ \mathbf{0} \end{bmatrix} , \tag{3}$$

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

In this case, finding the subspace-to-subspace distance can be reduced to solving a linear system with $3n - 2m$ unknowns and equations. This formulation is thus preferable when $m > \frac{3}{4}n$.

For the point-to-subspace distance between a private descriptor under this representation $\mathcal{D}$ and an original descriptor e, the system can be simplified to:

$$\begin{cases} e - x^* = \sum_{j=1}^{n-m} \mu_j d_j \\ d_i^T (x^* - d_0) = 0 \end{cases} \tag{4}$$

$$\Leftrightarrow \sum_{j=1}^{n-m} \mu_j d_i^T d_j = d_i^T (e - d_0) \tag{5}$$

$$\Leftrightarrow \mu_i = d_i^T (e - d_0) \ , \tag{6}$$

since $DD^T = I$. Thus,

$$\text{dist}(\mathcal{D}, e) = \| \sum_{j=1}^{n-m} d_j^T (e - d_0) d_j \| \tag{7}$$

$$= \| p_\perp^{\text{span}(d_1, \ldots, d_{n-m})} (e - d_0) \| \ . \tag{8}$$

This formulation is more advantageous when $m \geq \frac{1}{2}n$ as it only requires $n - m$ dot product evaluations instead of m.

2. Complexity Analysis

Time Complexity. The complexity of lifting to an m-dimensional subspace is $\mathcal{O}(mn)$ under the supposition that the lifting database offers $\mathcal{O}(1)$ access to a random element (*e.g.*, array, hashtable).

In general, for matching two features lifted to m-dimensional affine subspaces under the primal representation, we require a matrix multiplication $(m \times n)(n \times m)$ (*i.e.*, $M = -DE^T$), the resolution of a system with $2m$ unknowns and equations, and a constant number of additional matrix multiplications between $m \times m$ matrices. Thus, the complexity is $\mathcal{O}(m^2 n + m^3)$. Similarly, for the dual representation, the complexity is $\mathcal{O}((3n - 2m)^3)$.

Attack	Lifting	Dim.	MAE (↓)	SSIM (↑)	PSNR (↑)
	raw	0	0.092	0.778	18.645
NNA	random	2	0.111	0.740	17.386
NNA	sub-hybrid	2	0.181	0.519	13.434
DIA	sub-hybrid	2	0.150	0.653	14.959
DIA	sub-hybrid	4	0.160	0.611	14.471
DIA	sub-hybrid	6	0.166	0.585	14.154

Table 1: **Image reconstruction – SIFT statistics.** We report quality metrics between reconstructed and original images for SIFT descriptors.

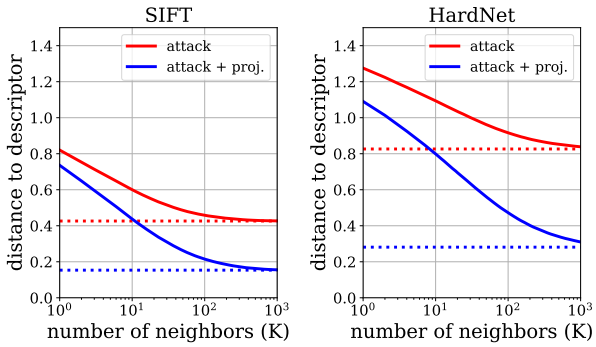

Figure 1: **Image reconstruction – quantitative.** We plot the average distance of the reconstructed descriptors to the original one for different values of K, the number of nearest neighbors considered during the attack – descriptors were lifted using sub-hybrid lifting to planes. The dotted lines are the limits of their solid conterparts.

To match two images with N_1 and N_2 local features respectively, we use exhaustive matching which requires computing distances between all pairs of features, *i.e.*, a time complexity of $\mathcal{O}(N_1 N_2 C)$, where C is the complexity of matching two features as defined above.

Space Complexity. For the primal representation, we require one translation vector and m basis vectors totaling $\mathcal{O}((m+1)n)$ floating point variables instead of $\mathcal{O}(n)$ for the original features. For the dual representation, we require storing $\mathcal{O}((n-m+1)n)$ floating point variables.

3. Privacy Attacks on Local Features

In this section, we first provide additional results of the proposed privacy attacks on local features. We then study a new oracle based attack underlining the effectiveness of adversarial lifting.

Additional Results. We run the privacy attacks described in Section 4.2, paragraph *Privacy Attack* of the main paper on both SIFT and HardNet private features with different lifting methods and dimensions. To recall, we proposed a nearest neighbor (NNA) and a direct inversion attack (DIA). In Table 1, 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) for SIFT descriptors. We show additional qualitative results of the attack for SIFT and HardNet descriptors in Figures 2 and 3, respectively. All images were published on Flickr under a CC BY 4.0 License. Image credit (top-to-bottom): `twang_dunga` (Twang Dunga), `scaredykat` (krista), `bab4lity` (wwikgren), `herry` (Herry Lawford), `smemon` (Sean MacEntee), `laylamoran4battersea` (Layla Moran), `shankaronline` (Shankar S.), `martinalvarez` (Martin Alvarez Espinar), `pagedooley` (Kevin Dooley), `nukeit1` (James McCauley).

Oracle Attack. In this section, we also provide the adversary with a fictional oracle that, given a list of possible attack descriptors for a private feature, returns the closest one to the original descriptor. We propose the following attack methodology: for each private representation $\mathcal{D}$ associated to a descriptor d, the database V of 128,000 real-world descriptors is used to retrieve the K closest elements to the subspace $\tilde{d}_1, \ldots, \tilde{d}_K$. Next, these attack descriptors are provided to the oracle, which returns the closest one to the original descriptor d, *i.e.*, $j = \arg\min_{i \in \{1,\ldots,K\}} \|\tilde{d}_i - d\|$. The descriptor $\tilde{d}_j$ is then used as an approximation to the original descriptor. We also consider a version where the reconstructed descriptor is obtained by orthogonal projection of $\tilde{d}_j$ to the subspace $\mathcal{D}$ (denoted by *proj.*). Finally, a feature inversion network can be used to reconstruct the original image from the approximated descriptors. In practice, the attacker does not have access to the original descriptor d, so implementing an oracle would be extremely challenging.

Figure 1 shows quantitative results of the oracle attack on the 10 Flickr holiday images totaling around 40,000 features with SIFT and HardNet descriptors. We plot the average distance between the original and the reconstructed descriptor as a function of the number of neighbors K. For this experiment, we used sub-hybrid lifting to planes ($m = 2$). The projected version is always closer, but it is not necessarily on the unit hyper-sphere. The dotted lines represent the asymptotic values of each respective curve, *i.e.*, the value for $K = 128{,}000$. A first important observation is that, despite only using one adversarial sample during the subspace construction, there is a significant number of confounding real-world descriptors in the neighborhood of the subspace. Note that SIFT descriptors only take positive values (*i.e.*, in $\mathbb{R}_+^{128}$), which explains the smaller distance between reconstructed and original when compared to HardNet descriptors taking values in $\mathbb{R}^{128}$. We also show qualitative examples in Figures 4 and 5. Even for large numbers of neighbors and access to an imaginary oracle, the reconstructed image remains far from the original.

Figure 2: **Image reconstruction – SIFT.** We show qualitative examples: first original image, then reconstructions from the raw descriptors and using the proposed privacy attacks on different lifting methods and dimensions.

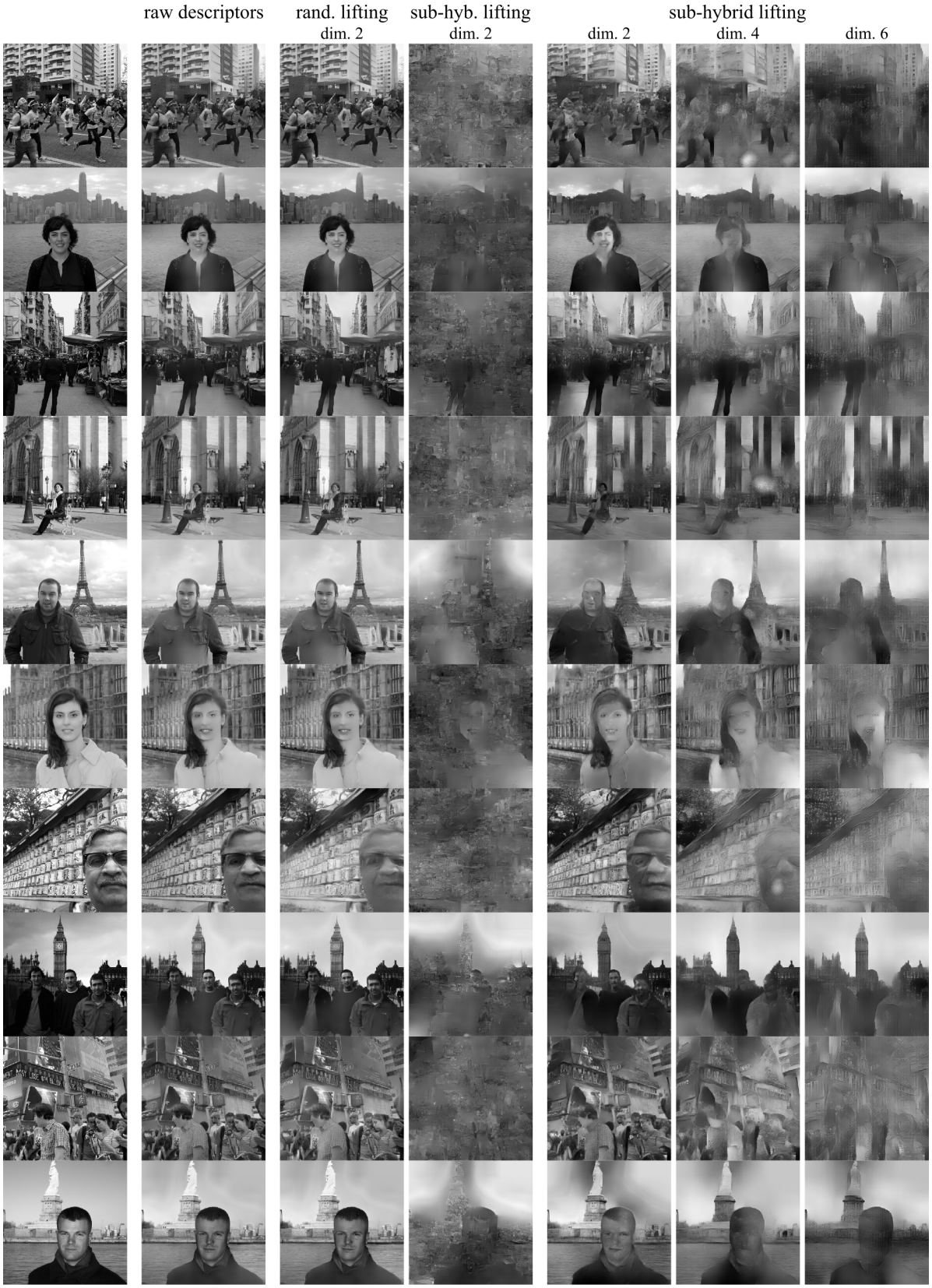

Figure 3: **Image reconstruction – HardNet.** We show qualitative examples: first original image, then reconstructions from the raw descriptors and using the proposed privacy attacks on different lifting methods and dimensions.

Figure 4: **Image reconstruction (oracle) – SIFT.** We show qualitative examples: first original image, then reconstructions from the raw descriptors and using the oracle privacy attack for different values of K. Descriptors are lifted to planes ($m = 2$).

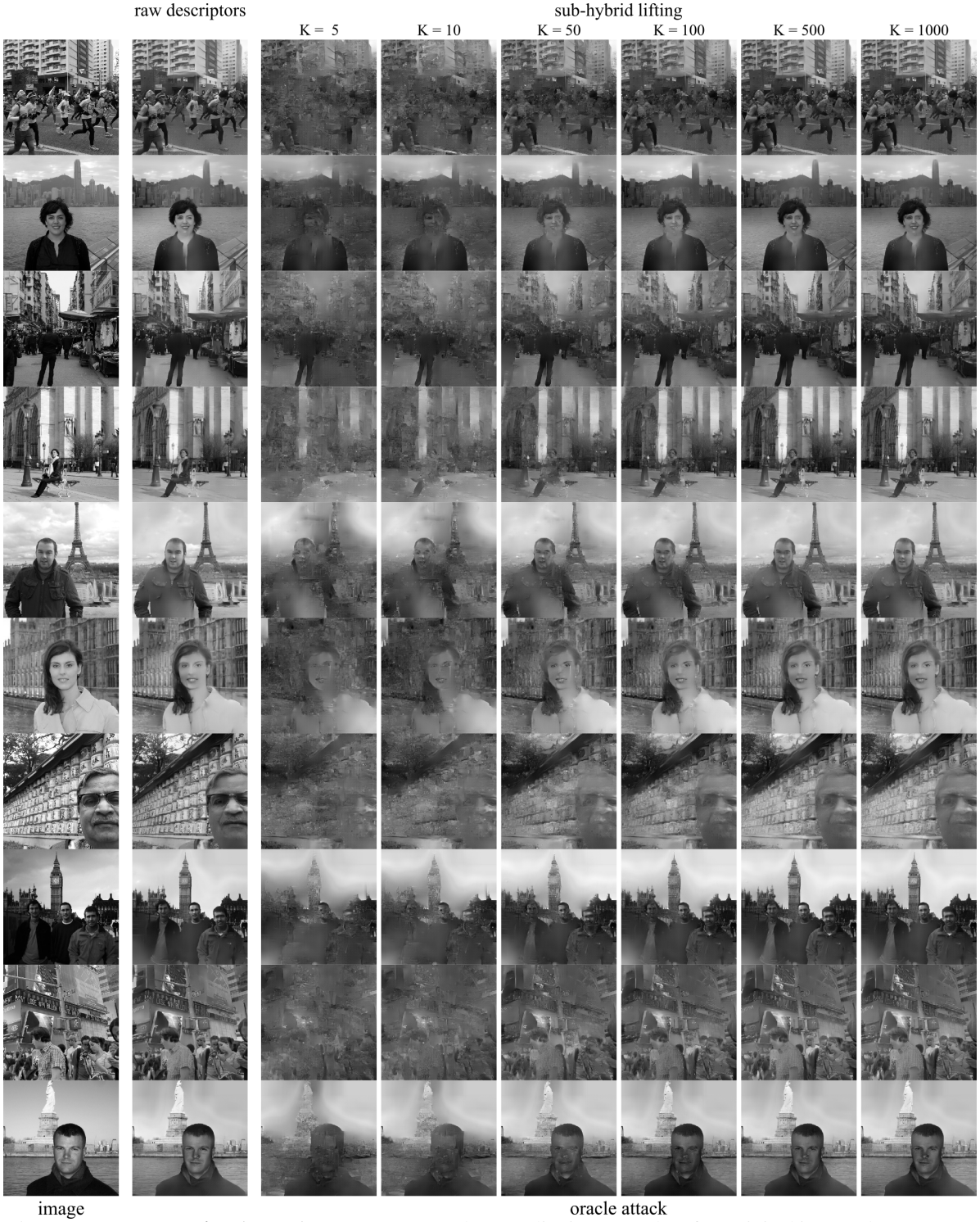

Figure 5: **Image reconstruction (oracle) – HardNet.** We show qualitative examples: first original image, then reconstructions from the raw descriptors and using the oracle privacy attack for different values of K. Descriptors are lifted to planes ($m = 2$).