Learning a Discriminative Null Space for Person Re-identification

Li Zhang Tao Xiang Shaogang Gong
Queen Mary University of London
{david.lizhang, t.xiang, s.gong}@qmul.ac.uk

Abstract

Most existing person re-identification (re-id) methods focus on learning the optimal distance metrics across camera views. Typically a person's appearance is represented using features of thousands of dimensions, whilst only hundreds of training samples are available due to the difficulties in collecting matched training images. With the number of training samples much smaller than the feature dimension, the existing methods thus face the classic small sample size (SSS) problem and have to resort to dimensionality reduction techniques and/or matrix regularisation, which lead to loss of discriminative power. In this work, we propose to overcome the SSS problem in re-id distance metric learning by matching people in a discriminative null space of the training data. In this null space, images of the same person are collapsed into a single point thus minimising the within-class scatter to the extreme and maximising the relative between-class separation simultaneously. Importantly, it has a fixed dimension, a closed-form solution and is very efficient to compute. Extensive experiments carried out on five person re-identification benchmarks including VIPeR, PRID2011, CUHK01, CUHK03 and Market1501 show that such a simple approach beats the state-of-the-art alternatives, often by a big margin.

1. Introduction

The problem of person re-identification (re-id) has attracted great attention in the past five years [34, 10]. When a person is captured by multiple non-overlapping views, the objective is to match him/her across views among a large number of imposters. Despite the best efforts from the computer vision researchers, re-id remains a largely unsolved problem. This is because that a person's appearance often undergoes dramatic changes across camera views due to changes in view angle, body pose, illumination and background clutter. Furthermore, since people are mainly distinguishable by their clothing under a surveillance setting, many passers-by can be easily confused with the target person because they wear similar clothes.

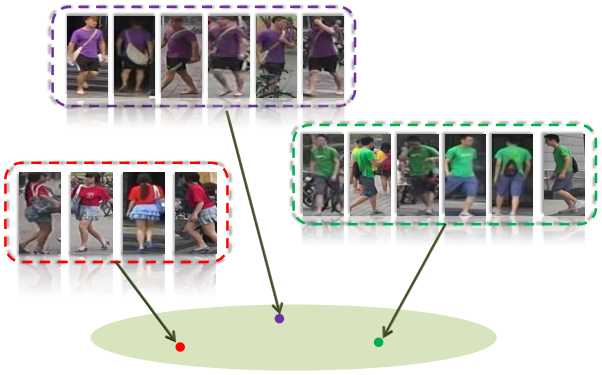

Figure 1. Training images of same identity are projected to a single point in a learned discriminative null space.

Existing approaches focus on developing discriminative feature representations that are robust against the view/pose/illumination/background changes [12, 38, 6, 18, 26, 41, 22], or learning a distance metric [12, 17, 31, 44, 28, 33, 30, 21, 40, 39, 36, 27, 23, 22], or both jointly [20, 1]. Among them, the distance metric learning methods are most popular and are the focus of this paper. Given any feature representation and a set of training data consisting of matching image pairs across camera views, the objective is to learn the optimal distance metric that gives small values to images of the same person and large values for those of different people. Distance metric learning has been extensively studied in machine learning [37], and existing metric learning methods employed for re-id are either originated elsewhere or extensions of existing methods with modifications to address the additional challenges arising from the re-id task. Although they have been shown to be effective in improving the existing re-id benchmarks over the past five years, all these models are still limited by some of classical problems in model learning.

Specifically, a key challenge for distance metric learning when applied to person re-id is the small sample size (SSS) problem [4]. Specifically, to capture rich person appearance whilst being robust against those condition changes mentioned above, the feature representations used by most

recent re-id works are of high dimension – typically in the order of thousands or tens of thousands. In contrast, the number of training samples is typically small, normally in hundreds. This is because that collecting training samples of matched person pairs across views is labour intensive and tedious. As a result the sample size is much smaller (often in an order of magnitude) than the feature dimension, a problem known as the SSS problem. Metric learning methods suffer from the SSS problem because they essentially aim to minimise the within-class (intra-person) variance (distance), whilst maximising the inter-class (inter-person) variance (distance). With a small sample size, the within-class scatter matrix becomes singular [4]; to avoid it, unsupervised dimensionality reduction or regularisation are required. This in turn makes the learned distance metric sub-optimal and less discriminative [4, 43, 13].

In this paper, we argue that the SSS problem in person re-id distance metric learning can be best solved by learning a discriminative null space of the training data. In particular, instead of minimising the within-class variance, data points of the same classes are *collapsed*, by a transform, into a single point in a new space (see Fig. 1). By keeping the between-class variance non-zero, this automatically maximises the Fisher discriminative criterion and results in a discriminative subspace. The null space method, also known as the null Foley-Sammon transfer (NFST) [13] is specifically designed for the small sample case, with rigorous theoretical proof on the resulting subspace dimension. Importantly, it has a closed-form solution, no parameter to tune, requires no pre-precessing steps to reduce the feature dimension, and can be computed efficiently. Furthermore, to deal with the non-linearity of the person's appearance, a kernel version can be developed easily to further boost the matching performance within the null space. It therefore offers a perfect solution to the challenging person re-id problem. In addition to formulating the NSFT model as a fully supervised model to solve the person re-id problem, we also extend it to the semi-supervised setting to further alleviate the effects of the SSS problem by exploiting unlabelled data abundant in re-id applications.

The contributions of this work are as follows: (1) We identify the small sample size (SSS) problem suffered by all existing metric learning based re-id methods and argue that their solutions to this problem is suboptimal. (2) For the first time, we propose to overcome the SSS problem in person re-id by learning a discriminative null space of the training data. (3) We develop a novel semi-supervised learning method in the null space to exploit the abundant unlabelled data to further alleviate the effects of the SSS problem. Extensive experiments carried out on five person re-identification benchmarks including VIPeR [11], PRID2011 [14], CUHK01 [19], CUHK03 [19] and Market1501 [42] show that such a simple and computationally very efficient approach beats all state-of-the-art methods presented to date, often by a large margin.

2. Related Work

Existing works on person re-id can be roughly categorised into three groups. The first group of methods design invariant and discriminant features [20, 12, 6, 18, 26, 41, 24, 38, 22]. The general trend is that the dimensions of the proposed features are getting higher. For instance the dimensions of two representations, recently proposed in [24] and [22] and used in our experiments, are 5,138 and 26,960 respectively. However, no matter how robust the designed features are, they are unlikely to be completely invariant to the often drastic cross-view pose/illumination/background changes. Therefore, the second group of methods focus on learning robust and discriminative distance metrics or subspaces for matching people across views [12, 17, 31, 44, 28, 33, 30, 21, 40, 39, 36, 27, 24, 23, 22, 29]. Recently, the third group of methods start to appear which are based on deep learning [20, 1]. However, person re-id seems to be one of the few vision problems that deep learning has not been able to shine due to the small training sample size problem addressed in this work.

Apart from a few exceptions [12, 31] based on ranking or boosting, the second groups of methods can be further divided into two major sub-groups: those on learning distance metrics [17, 44, 28, 33] and those on learning discriminative subspaces [30, 24, 36, 22]. Seemingly different, these two sub-groups are closely related [9]. Specifically, most metric learning methods focus on Mahalanobis form metrics. If the linear projection of a feature vector $\mathbf{x}_i$ in a learned discriminative subspace is denoted as $\mathbf{y}_i$, we have $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$. The Euclidean distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ is exactly a Mahalanobis distance $||\mathbf{y}_i - \mathbf{y}_j||^2 = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)$ where $\mathbf{A} = \mathbf{W}^T\mathbf{W}$ is a positive semidefinite matrix. In other words, learning a discriminative subspace followed by computing Euclidean distance is equivalent to computing a discriminative Mahalanobis distance over feature vectors in the original space. By making this connection, it is not difficult to see why both methods suffer from the same SSS problem typically associated with the subspace learning methods [4, 43, 13]. Most existing methods need to work with a reduced dimensionality [30], achieved typically by PCA whose dimension has to be carefully tuned for each dataset. Some works additionally require introducing matrix regularisation term if the intra-class scatter matrix is used in the formulation, in order to prevent matrix singularity [30, 24, 36, 22], again with free parameters to tune. Critically, they suffer from the degenerate eigenvalue problem (i.e. several eigenvectors share the same eigenvalue), which makes the solution sub-optimal resulting in loss of discriminant ability [43]. In contrast, for our discriminative null space based approach, neither dimensionality reduction be-

fore model learning nor regularisation term is required, and it has no parameters to tune.

As a solution proposed specifically to address the SSS problem, the null Foley-Sammon transfer (NFST) method has been around for a long time [13], but received very little attention apart from a recent application to the novelty detection problem [2]. A possible reason is that by restricting the learned discriminative projecting directions to the null projecting directions (NPDs), on which within-class distance is always zero and between-class distance is positive, the model is *extreme*, leaving little space for further extension with clear added-value. For example, the more relaxed Fisher discriminative analysis (FDA) can be extended, gaining notable advantage, by exploit graph laplacian to preserve local data structure, known as LFDA[32], which has been successfully applied to re-id [30]. However, a similar graph laplacian extension to NFST does not apply due to its single point per class nature. Despite the restrictions, given the acute SSS problem in re-id distance metric learning, the basic idea of learning a null space for overcoming this problem becomes very attractive. The general concept of collapsing same-class data points to a single point has been exploited in a Mahalanobis distance learning framework, known as maximally collapsing metric learning (MCML) [9]. However, MCML does not exploit a null space. Instead, the MCML model must make approximations with plenty of free parameters to tune and no closed-form solution.

In this work, we exploit the original null Foley-Sammon transfer (NFST) method [13] with its conventional supervised learning approach, benefiting from its attractive closed-form solution and no parameters tuning required. Moreover, we extend the original fully supervised null space model to a semi-supervised learning setting. This is to explore, in addition to a few labelled data, a larger quantities of unlabelled data typically available in person re-id scenarios for model learning. The problem of semi-supervised re-id has attracted interest lately due to its potential to overcome the lack of training data problem. One approach is by dictionary learning for sparse coding [25, 16], which has an unsupervised nature, thus can be learned with both labelled and unlabelled data. In this work, we compare the new semi-supervised null space model against dictionary learning based methods and demonstrate the superior performance from the new model.

3. Methodology

3.1. Problem Definition

Given a set of N training data denoted as $\mathbf{X} \in \mathbb{R}^{d \times N}$. Each column of the data descriptor matrix $\mathbf{X}$, $\mathbf{x}_i$ is a feature vector representing the i-th training sample. In the case of person re-id, this feature vector is extracted from a person detection box and contains appearance information about the person, and its dimension d is typically very high. We assume that each data point belongs to one of C classes, i.e. C different identities. The objective of learning a discriminative null space is to learn a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times m}$ to project the original high-dimensional feature vector $\mathbf{x}_i$ into a lower-dimensional one $\mathbf{y}_i \in \mathbb{R}^m$ with $m < d$. Person re-id can then be performed by computing the Euclidean distance between two projected vectors in the learned discriminative null space.

3.2. Foley-Sammon Transform

The learned null Foley-Sammon transform (NFST) space is closely related to linear discriminant analysis (LDA), also known as Foley-Sammon transform (FST) [8]. So before we formulate NFST, let us first briefly revisit FST.

The objective of FST is to learn a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times m}$ so that each column, denoted as $\mathbf{w}$, is an optimal discriminant direction that maximises the Fisher discriminant criterion:

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_b \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}, \tag{1}$$

where $\mathbf{S}_b$ is the between-class scatter matrix and $\mathbf{S}_w$ is the within-class scatter matrix. The optimisation of Eq. (1) can be done by solving the following generalised eigenproblem:

$$\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}. \tag{2}$$

If $\mathbf{S}_w$ is non-singular, $C-1$ eigenvectors $\mathbf{w}^{(1)}, ..., \mathbf{w}^{(c-1)}$ can be computed corresponding to the $C-1$ largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$. Using them as the columns, the projection matrix $\mathbf{W}$ can project the original data into a $C-1$ dimensional discriminative subspace where the C classes become maximally separable. However, in the small sample size case, we have $d > N$; as a result, $\mathbf{S}_w$ is singular. FST thus runs in numerical problems and common solutions include reducing d by PCA or adding a regularisation term to $\mathbf{S}_w$. In [13], a more principled way to overcome the SSS problem in FST is proposed, termed as Null Foley-Sammon transform (NFST).

3.3. Null Foley-Sammon transform

NFST aims to learn a discriminative subspace where the training data points of each of the C classes are collapsed to a single point, resulting in C points in the space. In order to make this subspace discriminative, these C points should not further collapse to a single point. Formally, we aim to learn the optimal projection matrix $\mathbf{W}$ so that each of its column $\mathbf{w}$ satisfies the following two conditions:

$$\mathbf{w}^\top \mathbf{S}_w \mathbf{w} = 0, \tag{3}$$

$$\mathbf{w}^\top \mathbf{S}_b \mathbf{w} > 0. \tag{4}$$

That is, it satisfies zero within-class scatter and positive between-class scatter. This guarantees the best separability of the training data in the sense of Fisher discriminant criterion. Such a linear projecting direction $\mathbf{w}$ is called Null Projecting Direction (NPD) [13].

Next, we show that a NPD must lie in the null space of $\mathbf{S}_w$. In particular, we have the following Lemma:

Lemma 1. *Let $\mathbf{W}$ be a projection matrix which maps a sample $\mathbf{x}$ into the null space of $\mathbf{S}_w$, where the null space is spanned by the orthonormal set of $\mathbf{W}$, that is, $\mathbf{S}_w\mathbf{W} = 0$. If all samples are mapped into the null space of $\mathbf{S}_w$ through $\mathbf{W}$, the within-class scatter matrix $\widehat{\mathbf{S}}_w$ of the mapped samples is a complete zero matrix.*

Proof. Let $\mathbf{x}_n^c$ be the n^{th} sample of the $c^{th} \in \{1, ...C\}$ class which has N_c samples in total. $\mathbf{y}_n^c$ denote the mapped feature vector through $\mathbf{W}$. We have:

$$\begin{aligned}
\widehat{\mathbf{S}}_w &= \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{y}_n^c - \overline{\mathbf{y}}^c)(\mathbf{y}_n^c - \overline{\mathbf{y}}^c)^\top \\
&= \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{W}^\top\mathbf{x}_n^c - \mathbf{W}^\top\mu^c)(\mathbf{W}^\top\mathbf{x}_n^c - \mathbf{W}^\top\mu^c)^\top \\
&= \mathbf{W}^\top \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{x}_n^c - \mu^c)(\mathbf{x}_n^c - \mu^c)^\top \mathbf{W} \\
&= \mathbf{W}^\top \mathbf{S}_w \mathbf{W} = \mathbf{0}
\end{aligned}$$

where $\mathbf{y}_n^c = \mathbf{W}^\top\mathbf{x}_n^c$, $\overline{\mathbf{y}}^c = \mathbf{W}^\top\mu^c$, $\mu^c = \frac{1}{N_c}\sum_{n=1}^{N_c}\mathbf{x}_n^c$, N_c is the number of samples in class c, and μ^c is the mean vector of all data belonging to the class c.

Now with Lemma 1, we know that Eq. (3) holds as long as $\mathbf{w}$ is from the null space of $\mathbf{S}_w$. Next we take a look the condition in the inequality (4). It is easy to see that when Eq. (3) holds, (4) also holds if:

$$\mathbf{w}^\top \mathbf{S}_t \mathbf{w} > 0, \tag{5}$$

where $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$ is the total scatter matrix. We now denote the null space of the $\mathbf{S}_t$ and $\mathbf{S}_w$as:

$$\mathbf{Z}_t = \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{S}_t\mathbf{z} = 0\right\}, \tag{6}$$

$$\mathbf{Z}_w = \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{S}_w\mathbf{z} = 0\right\}, \tag{7}$$

and their orthogonal complements as $\mathbf{Z}_t^\perp$ and $\mathbf{Z}_w^\perp$ respectively. Now since $\mathbf{S}_b$ is non-negative definite, we can see that in order for the NPDs to satisfy both Eqs. (3) and (4) simultaneously, they must lie in the shared space between $\mathbf{Z}_w$ and $\mathbf{Z}_t^\perp$, that is:

$$\mathbf{w} \in (\mathbf{Z}_t^\perp \cap \mathbf{Z}_w). \tag{8}$$

It has been proved in [13] that there are precisely $C-1$ NPDs $\mathbf{w}$ that satisfy both Eq. (3) and (4). In other words, the discriminative null space we are looking for has $m = C-1$ dimensions.

3.4. Learning the Discriminative Null Space

Let $\mathbf{X}_w$ be the matrix consisting of vectors $\mathbf{x}_i^c - \mu^c$. $\mathbf{X}_t$ be the matrix consisting of vectors $\mathbf{x}_i - \mu$ with $\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$. We then have,

$$\mathbf{S}_w = \frac{1}{N}\mathbf{X}_w\mathbf{X}_w^\top, \quad \mathbf{S}_t = \frac{1}{N}\mathbf{X}_t\mathbf{X}_t^\top \tag{9}$$

Now we know where to look for the NPDs – the shared space between $\mathbf{Z}_w$ and $\mathbf{Z}_t^\perp$. Next, we shall see how to compute them. Let us first take a look at how to compute $\mathbf{w}$ that satisfies $\mathbf{w} \in \mathbf{Z}_t^\perp$. First we notice that:

$$\begin{aligned}
\mathbf{Z}_t &= \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{S}_t\mathbf{z} = 0\right\} = \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{z}^\top\mathbf{S}_t\mathbf{z} = 0\right\} \\
&= \left\{\mathbf{z} \in \mathbb{R}^d \mid (\mathbf{X}_t^\top\mathbf{z})^\top\mathbf{X}_t^\top\mathbf{z} = 0\right\} \\
&= \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{X}_t^\top\mathbf{z} = 0\right\}.
\end{aligned}$$

Hence, $\mathbf{Z}_t^\perp$ is the subspace spanned by zero-mean data $\mathbf{x}_i - \mu$. We can obtain the orthonormal basis $\mathbf{U} = [\mathbf{u}^{(1)}, ..., \mathbf{u}^{(N-1)}]$ of the zero-mean data using Gram-Schmidt orthonormalisation, then represent each solution $\mathbf{w}$ as:

$$\mathbf{w} = \beta_1\mathbf{u}^{(1)} + ... + \beta_{N-1}\mathbf{u}^{(N-1)} = \mathbf{U}\beta, \tag{10}$$

Note that there are $N-1$ basis vectors because the rank of $\mathbf{S}_t$ is $N-1$.

So now after expressing $\mathbf{w}$ using Eq. (10), it must satisfy $\mathbf{w} \in \mathbf{Z}_t^\perp$. The next step is the make it also satisfy $\mathbf{w} \in \mathbf{Z}_w$. This can be achieved by substituting Eq. (10) into Eq. (3) and solve the following eigen-problem:

$$(\mathbf{U}^\top\mathbf{S}_w\mathbf{U})\beta = \mathbf{0}, \tag{11}$$

for which we know that $C-1$ solutions $\beta^{(1)}, ..., \beta^{(C-1)}$ exist, giving $C-1$ NPDs, $\mathbf{U}\beta$.

In summary, the problem of learning the discriminative null space boils down to solving an eigen-problem which has a closed-form solution and can be solved very efficiently. Importantly, the whole optimisation algorithm has no free parameter to tune.

3.5. Kernelisation

The NFST model is a linear model. It has been demonstrated [36] that many distance metric learning or discriminative subspace based methods for person re-id benefit from kernelisation because of the non-linearity in person's appearance. In the following we describe how the discriminative null space can be kernelised.

Given a kernel function $k(\mathbf{x}_i, \mathbf{x}_i) = \langle\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)\rangle$, where $\Phi(\mathbf{x}_i)$ maps $\mathbf{x}_i$ to an implicit higher dimensional space, we can compute the data kernel matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$ for training data $\mathbf{X}$ as $\mathbf{K} = \Phi(\mathbf{X})^\top\Phi(\mathbf{X})$. Now the within-class scatter matrix $\mathbf{S}_w$ and total-class scatter matrix $\mathbf{S}_t$ can

be kernelised as:

$$\mathbf{K}_w = \mathbf{K}(\mathbf{I}-\mathbf{L})(\mathbf{I}-\mathbf{L})^\top\mathbf{K},$$
$$\mathbf{K}_t = \mathbf{K}(\mathbf{I}-\mathbf{M})(\mathbf{I}-\mathbf{M})^\top\mathbf{K},$$

where $\mathbf{I}$ is a $N \times N$ identity matrix, $\mathbf{L}$ is a block diagonal matrix with block sizes equal to the number of data points N_c for each class $c \in \{1, ...C\}$ and $\mathbf{M}$ is a $N \times N$ matrix with all entries equal to $\frac{1}{N}$.

Now to write Eq. (11) in its kernelised form, we need to replace $\mathbf{S}_w$ with $\mathbf{K}_w$, and compute the orthonormal basis of $\mathbf{K}_t$ to replace $\mathbf{U}$. The orthonormal basis of $\mathbf{K}_t$ can be computed using kernel PCA. First, we compute the centred kernel matrix $\widetilde{\mathbf{K}}$. Second, the eigendecomposition of $\widetilde{\mathbf{K}}$ is written as $\mathbf{K}_t = \mathbf{V}\mathbf{E}\mathbf{V}^\top$ with $\mathbf{E}$ being the diagonal matrix containing $N-1$ non-zero eigenvalues and $\mathbf{V}$ containing the corresponding eigenvectors in its columns. Now the scaled eigenvectors $\widetilde{\mathbf{V}} = \mathbf{V}\mathbf{E}^{-1/2}$ contain coefficients for the kernelised orthonormal basis used to replace $\mathbf{U}$ in Eq. (11). Let

$$\mathbf{H} = ((\mathbf{I}-\mathbf{M})\widetilde{\mathbf{V}})^\top\mathbf{K}(\mathbf{I}-\mathbf{L}), \tag{12}$$

and with Eq. (9), we can rewrite Eq. (11) as:

$$\mathbf{H}\mathbf{H}^\top\beta = \mathbf{0}. \tag{13}$$

By solving the eigen-problem Eq. (13), we obtain the final $C-1$ null projection directions (NPDs) as:

$$\mathbf{w}^{(i)} = ((\mathbf{I}-\mathbf{M})\widetilde{\mathbf{V}})^\top\beta^{(i)} \quad \forall i = 1, ..., C-1. \tag{14}$$

3.6. Semi-supervised Learning

The NFST method is a fully supervised method. When applied to the problem of re-id, the labelled training set is used to learn the projection $\mathbf{W}$. The test data are then projected into the same subspace and matched by computing the Euclidean distance between a query sample and a set of gallery samples.

In a real-world application scenario, the labelled training data are scarce but there are often plenty of unlabelled data (person images collected from different views) that can be used to alleviate the small sample size problem. To this end, the NFST method is extended to the semi-supervised setting. More specifically, given a training set $\mathbf{X}$ contains a labelled subset $\mathbf{X}^l$ of N^l samples and an unlabelled subset $\mathbf{X}^u$ of N^u samples. Using the NFST method described above, we can first learn an initial projection matrix $\mathbf{W}^0$ using $\mathbf{X}^l$ only. Then $\mathbf{X}^u$ is projected to the lower-dimensional subspace through $\mathbf{W}^0$ and becomes $\mathbf{Y}^u_{\mathbf{W}^0}$. To utilise the unlabelled data $\mathbf{X}^u$, we use their projections $\mathbf{Y}^u_{\mathbf{W}^0}$ to build a cross-view correspondence matrix $\mathbf{A} \in \mathbb{R}^{N_u \times N_u}$ which captures the identity relationship for the unlabelled people across views. Note, since the data are unlabelled, the true cross-view correspondence relationship is unknown. We therefore use $\mathbf{A}$ to represent a soft cross-view correspondence relationship. That is, each person in one view can correspond to multiple people in another view depending on their visual similarity in the learned discriminative subspace parameterised by $\mathbf{W}^0$. To this end, we first construct a k-nearest-neighbour (k-nn) graph G across camera views with N_u vertices, where each vertex represents a unlabelled data point. $\mathbf{A}$ is then computed as the weight matrix of G using a heat kernel. With this k-nn graph, we then create pseudo-classes, each consisting one vertex from one view and its k-nearest-neighbours from the other view. Next these pseudo-classes are augmented with the labelled classes in $\mathbf{X}^l$ to create a new training set, denoted $\mathbf{P}$, on which a new project matrix $\mathbf{W}^1$ is computed using NFST. Re-learning the projection matrix runs iteratively till the average distance for the k-nearest-neighbours stop decreasing. In our experiments, we found that the algorithm converges rapidly.

Algorithm 1 Semi-supervised null space learning

Input: $\mathbf{X}^l$, $\mathbf{X}^u$, k, $\mathbf{P}^0 = \mathbf{0}$.
Output: The learned projection $\mathbf{W}$.

```
Estimate W^0 using X^l;
t = 0;
while not converged do
    project X^u through W^t to obtain Y^u_{W^t}
    build k-nn graph G with Y^u_{W^t}
    take top f percent to create the pseudo-classes P^{t+1}
    learn W^{t+1} with X^l + P^{t+1}
    t = t + 1
end while
```

This semi-supervised learning is essentially based on self-training, a popular strategy taken by many semi-supervised learning methods [45]. For any self-training based methods, preventing model drift is of paramount importance. Apart from examining the average distance for the k-nearest-neighbours, another measure taken is to rank the k-nearest-neighbours and take only the top f percent with the smallest distance to create the pseudo-classes. The complete semi-supervised null space learning algorithm is summarised in Alg. 1.

4. Experiments

4.1. Datasets and Settings

Datasets Five widely used datasets are selected for experiments, including the three largest benchmarks available (CUHK01, CUHK03, and Market1501).

VIPeR [11] contains 632 identities and each has two images captured outdoor from two views with distinct view angles. All images are scaled to 128 × 48 pixels. The

632 people's images are randomly divided into two equal halves, one for training and the other for testing. This is repeated for 10 times and the averaged performance is reported.

PRID2011 [14] consists of person images recorded from two cameras. Specifically, it has two camera views. View A captures 385 people, whilst View B contains 749 people. Only 200 people appear in both views. The single shot version of the dataset is used in our experiments as in [15]: In each data split, 100 people with one image from each view are randomly chosen from the 200 present in both camera views for the training set, while the remaining 100 of View A are used as the probe set, and the remaining 649 of View B are used as gallery. Experiments are repeated over the 10 splits provided in [15].

CUHK01 [19] contains 971 identities with each person having two images in each camera view. All the images are normalised to 160 $\times$ 60 pixels. Following the standard setting, images from camera A are used as probe and those from camera B as gallery. We randomly partition the dataset into 485 people for training and 486 for testing (multi-shot) following [22, 41], again over 10 trials.

CUHK03 [20] contains 13,164 images of 1,360 identities, captured by six surveillance cameras with each person only appearing in two views. It provides both manually labelled pedestrian bounding boxes and bounding boxes automatically detected by the deformable-part-model (DPM) detector [7]. A real-world re-id system has to rely on a person detector; the latter version of the data is thus ideal for testing performance given detector errors. We report results on both of the manually labelled and detected person images. The 20 training/test splits provided in [20] is used under and the single-shot setting as in [22] – two images are randomly chosen for testing; one is for probe and the other for gallery.

Market1501 [42] is the biggest re-id benchmark dataset to date, containing 32,668 detected person bounding boxes of 1,501 identities. Each identity is captured by six cameras at most, and two cameras at least. During testing, for each identity, one query image in each camera is selected, therefore multiple queries are used for each identity. Note that, the selected 3,368 queries in [42] are hand-drawn, instead of DPM-detected as in the gallery. Each identity may have multiple images under each camera. We use the provided fixed training and test set, under both the single-query and multi-query evaluation settings.

Feature Representations By default the recently proposed Local Maximal Occurrence (LOMO) features [22] are used for person representation. The descriptor has 26,960 dimensions. To test our method's ability to fuse different representations, we also consider another histogram-based image descriptor proposed in [24]. These include colour histogram, HOG and LBP which are concatenated resulting in 5138 dimensions.

Evaluation metrics We use Cumulated Matching Characteristics (CMC) curve to evaluate the performance of person re-identification methods for all datasets in this paper. Due to space limitation and for easier comparison with published results, we only report the cumulated matching accuracy at selected ranks in tables rather than plotting the actual curves. Note that for the Market1501 dataset, since there are on average 14.8 cross-camera ground truth matches for each query, we additionally use mean average precision(mAP) as in [42] to evaluate the performance.

Parameter setting There is no free parameter to tune for our model. However, with the kernelisation, kernel selection is necessary. Unless stated otherwise, RBF kernel is used with the kernel width determined automatically using the mean pairwise distance of samples. For other compared methods, different model specific parameters have to be tuned carefully to report the highest results. Note that under the semi-supervised null space learning algorithm, there are free parameters: the value of k in the k-nn graph is fixed to 3 for all experiments. The percentage of neighbours f kept for creating pseudo classes are fixed at 40%. We found that the results are not sensitive to the values of these parameters.

4.2. Fully Supervised Learning Results

For the fully supervised setting, all the labels of the training data are used for model learning. For different datasets, we select different most representative and competitive alternative methods for comparison.

Results on VIPeR We first evaluate our method against the state-of-the-art on VIPeR. We compare with 17 existing methods. Among them, the distance metric learning based methods are RPLM [15], MtMCML [27], Mid-level Filter [41], SCNCD [38], Similarity Learning [3], LADF [21], ITML [5], LMNN [35], KISSME [17], and MCML [9], whilst the others are discriminative subspace learning based methods including kCCA [24], MFA [36], kLFDA [36], and XQDA [22]. Note that XQDA can be considered as hybrid between metric learning and subspace learning. In addition, deep learning based model is also compared [1]. For fair comparison, whenever possible (i.e. code is available and features can be replaced), we compare with these methods using the same LOMO features. Otherwise, the reported results are presented.

From the results shown in Table 1, we can make the following observations: (1) Our method achieves the highest performance when a single type of features are used (Rank 1 of 42.28% compared to the closest competitor XQDA [22] which gives 40.00%). (2) For fair comparison against methods which fuse more than one types of features [29] or more than one models [41], we also present our method's result obtained by a simple score-level fusion using the two types

Table 1. Fully supervised results on VIPeR

Rank	1	5	10	20
RPLM [15]	27.00	55.30	69.00	83.00
MtMCML [27]	28.83	59.34	75.82	88.51
MCML [9]	20.19	47.31	63.96	77.69
Mid-level [41]	29.11	52.34	65.95	79.87
SCNCD [38]	37.80	68.50	81.20	90.40
LADF [21]	30.22	64.70	78.92	90.44
Improved Deep [1]	34.81	63.61	75.63	84.49
Similarity Learning [3]	36.80	70.40	**83.70**	91.70
ITML (LOMO) [5]	24.65	49.78	63.04	78.39
LMNN (LOMO) [35]	29.43	59.78	73.51	84.91
KISSME (LOMO) [17]	34.81	60.44	77.22	86.71
kCCA (LOMO) [24]	30.16	62.69	76.04	86.80
MFA (LOMO) [36]	38.67	69.18	80.47	89.02
kLFDA (LOMO) [36]	38.58	69.15	80.44	89.15
XQDA (LOMO) [22]	40.00	68.13	80.51	91.08
Ours (LOMO)	**42.28**	**71.46**	82.94	**92.06**
Mid-level+LADF [41]	43.39	73.04	84.87	93.70
Metric Ensembles [29]	45.90	77.50	88.90	95.80
Ours (Fusion)	**51.17**	**82.09**	**90.51**	**95.92**

of features described earlier. Our method (Ours (Fusion)) beats the nearest rival [29] by over 5% on Rank 1. (3) The discriminative subspace learning based methods seem to be more competitive compared with the distance metric learning based methods. Note that all of them have been kernelised and we observe a significant drop in performance without kernelisation. This confirms the conclusion drawn in [36] that kernelisation is critical for addressing the non-linearity problem in re-id. (4) The most related methods MCML [9] and MtMCML [27] yeild much poorer results[1], indicating that the principle of collapsing same-class samples is better realised in a subspace learning framework which provides an exact and closed-form solution. (5) The deep learning based method [1] does not fare well on this small dataset despite the fact that the model has been pre-trained on the far-larger CUHK01+CUHK03 datasets. This suggests that the model learned from other datasets are not transferable by the simple model fine-tuning strategy and small sample size remains a bottle-neck for applying deep learning to re-id.

Results on PRID2011 We compare the state-of-the-art [15, 29] results reported on PRID2011 in Table 2. With access to the implementation codes, we also compare with the methods in [22, 36, 24] using the same LOMO features. The results show clearly with a single feature type, our method is the state-of-the-art; when fusing two types of features, the result is improved dramatically (over 10% increase on both Rank 1 and 5), and significantly higher than the reported results of the feature fusion method in [29], which fuses four different types of features including the deep convolutional neural network (CNN) features.

Table 2. Fully supervised results on PRID2011

Rank	1	5	10	20
RPLM [15]	15.00	32.00	42.00	54.00
kCCA (LOMO) [24]	14.30	37.40	47.60	62.50
MFA (LOMO) [36]	22.30	45.60	57.20	68.20
kLFDA (LOMO) [36]	22.40	46.50	58.10	68.60
XQDA (LOMO) [22]	26.70	49.90	61.90	73.80
Ours (LOMO)	**29.80**	**52.90**	**66.00**	**76.50**
Metric Ensembles [29]	17.90	39.00	50.00	62.00
Ours (Fusion)	**40.90**	**64.70**	**73.20**	**81.00**

Results on CUHK01 & CUHK03 Compared with VIPeR and PRID2011, these two datasets are much bigger with thousands of training samples. However, the sample size is still much smaller than the feature dimension, i.e. the SSS problem still exists. Table 3 shows that on CUHK01, our method beats all compared existing methods at low ranks and when two types of features are fused, the margin is significant. As for CUHK03, there are two versions: the one with manually cropped person images, and the one with bounding boxes produced by a detector. The latter obviously is harder as reflected by the decrease of matching accuracy for all compared methods. But it is also a better indicator of real-world performance. It can be seen from Table 4 that, as expected, on this much larger dataset, the deep learning based model [1] with its millions of parameters becomes much more competitive – with manually cropped images, our result with single feature type is higher on Rank 1 but lower on other ranks. However, with the detector boxes, our method is less affected and outperforms the deep model in [1] by a big margin. In addition, our performance is further boosted by fusing two types of features.

Table 3. Fully supervised results on CUHK01

Rank	1	5	10	20
SalMatch [39]	28.45	45.85	55.67	67.95
Mid-level Filter [41]	34.30	55.06	64.96	74.94
Improved Deep [1]	47.53	71.60	80.25	87.45
kCCA (LOMO) [24]	56.30	80.66	87.94	93.00
MFA (LOMO) [36]	54.79	80.08	87.26	92.72
kFLDA (LOMO) [36]	54.63	80.45	86.87	92.02
XQDA (LOMO) [22]	63.21	83.89	**90.04**	94.16
Ours (LOMO)	**64.98**	**84.96**	89.92	**94.36**
Metric Ensembles [29]	53.40	76.40	84.40	90.50
Ours (Fusion)	**69.09**	**86.87**	**91.77**	**95.39**

Results on Market1501 This dataset is the largest and most realistic dataset with natural detector errors abundant

[1]The result of MCML is from [27] using different features. We did have access to the code of MCML. However, no matter how hard we try, it would not converge to a meaningful solution using the higher-dimensional LOMO features.

Table 4. Fully supervised results on CUHK03. '-' means that no reported results is available.

Dataset	CUHK03 (Manual)				CUHK03 (Detected)			
Rank	1	5	10	20	1	5	10	20
DeepReID [20]	20.65	51.50	66.50	80.00	19.89	50.00	64.00	78.50
Improved Deep [1]	54.74	**86.50**	**93.88**	**98.10**	44.96	76.01	83.47	93.15
XQDA (LOMO) [22]	52.20	82.23	92.14	96.25	46.25	78.90	88.55	94.25
Ours (LOMO)	**58.90**	85.60	92.45	96.30	**53.70**	**83.05**	**93.00**	**94.80**
Metric Ensembles [29]	62.10	89.10	94.30	97.80	-	-	-	-
Ours (Fusion)	**62.55**	**90.05**	**94.80**	**98.10**	**54.70**	**84.75**	**94.80**	**95.20**

in the provided data as they were collected in front of a busy supermarket. Since it is new, few reported results are available. The baseline presented in [42] is not competitive because it is based on a weaker BoW features and L2-Norm distance. We compare our method with four alternatives with the same LOMO features. The results in Table 5 again show that our method significantly outperforms the alternatives, under both the single query and multi-query settings and with both evaluation metrics. This is despite the fact that with 12,936 training samples, the SSS problem is the least severe in this dataset.

Table 5. Fully supervised results on Market1501

Query	*singleQ*		*multiQ*	
Evaluation metrics	Rank-1	mAP	Rank-1	mAP
Baseline [42]	34.38	14.10	42.64	19.47
Baseline (+HS) [42]	-	-	47.25	21.88
KISSME (LOMO) [17]	40.50	19.02	-	-
MFA (LOMO) [36]	45.67	18.24	-	-
kLFDA (LOMO) [36]	51.37	24.43	52.67	27.36
XQDA (LOMO) [22]	43.79	22.22	54.13	28.41
Ours (LOMO)	**55.43**	**29.87**	**67.96**	**41.89**
Ours (Fusion)	**61.02**	**35.68**	**71.56**	**46.03**

4.3. Semi-supervised Learning Results

For semi-supervised setting, we use the VIPeR and PRID2011 datasets. The same data splits are used as in the fully-supervised setting. The difference is that only one third of the training data are labelled following the setting in [25, 16]. For comparison, apart from the state-of-the-art methods in [25, 16], we also choose three subspace learning based methods trained on the labelled data only.

The results in Table 6 show that the performance of our method is clearly superior to that of the compared alternatives. The advantage is more significant on PRID2011. This dataset has only 100 pairs or 200 training samples; with only one third of them labelled, the SSS problem becomes the most acute than any experiment we conducted before. Comparing Table 6 with Table 2, it is apparent that the performance of all three compared subspace learning methods, kCCA, kLFDA, and XQDA degrades drastically. In contrast, the performance of our method decrease much more gracefully from 29.80% to 24.70% on Rank 1. This is partly because our self-training based method can exploit the unlabelled data. It also shows that it can better cope with the SSS problem in its extreme.

Table 6. Semi-supervised Re-ID results on VIPeR and PRID2011

Dataset	VIPeR				PRID2011			
Rank	1	5	10	20	1	5	10	20
SSCDL [25]	25.60	53.70	68.20	83.60	-	-	-	-
kCCA (LOMO) [24]	13.64	37.97	53.77	69.94	5.80	16.00	24.70	36.00
kLFDA (LOMO) [36]	25.47	53.25	66.49	80.13	12.00	27.10	37.80	50.30
XQDA (LOMO) [22]	28.04	56.30	69.65	81.74	12.60	29.40	40.20	53.00
IterativeLap (LOMO) [16]	29.43	49.05	59.18	69.62	18.70	34.60	43.50	52.30
Ours (LOMO)	**31.68**	**59.40**	**72.78**	**84.91**	**24.70**	**46.80**	**58.20**	**68.20**
Ours (Fusion)	**41.01**	**69.81**	**81.61**	**91.04**	**35.80**	**58.10**	**69.10**	**78.90**

4.4. Running Cost

We compare the run time of our method with XQDA, kLFDA and MFA on Market1501. We calculate the overall training time over 12,936 samples and test time over 3,368 queries. All algorithms are implemented in Matlab and run on a server with 2.6GHz CPU cores and 384GB memory. Table 7 shows that for training, our method is the most efficiently, whilst on testing it is much slower than XQDA, but faster than kLFDA and MFA. Considering the test time is over 3,368 queries, it is more than adequate for real-time applications.

Table 7. Run time comparison on Market1501 (in seconds)

Method	Ours	XQDA [22]	kLFDA [36]	MFA [36]
Training	393.1	3233.8	995.2	437.8
Testing	31.3	1.6	43.4	43.2

5. Conclusion

We proposed to solve the person re-id problem by learning a discriminative null space of the training samples. Compared with existing re-id models, the employed NFST model is much simpler, with a closed-form solution and no parameters to tune. Yet, it is very effective in dealing with the SSS problem faced by the re-id methods. Extensive experiments on five benchmarks show that our method achieves the state-of-the-art performance on all of them under both fully supervised and semi-supervised settings.

Acknowledgement

This work was funded in part by the European FP7 Project SUNNY (grant agreement no. 313243).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 2, 6, 7, 8

[2] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler. Kernel null space methods for novelty detection. In *CVPR*, 2013. 3

[3] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015. 6, 7

[4] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 2000. 1, 2

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007. 6, 7

[6] M. Farenzena, L. Bazzani, A. Perina, M. Cristani, and V. Murino. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 1, 2

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 6

[8] D. Foley and J. Sammon. An optimal set of discriminant vectors. *IEEE Trans. Computers*, 1975. 3

[9] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005. 2, 3, 6, 7

[10] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person Re-Identification*. Springer, 2014. 1

[11] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on PETS*, 2007. 2, 5

[12] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 1, 2

[13] Y.-F. Guo, L. Wu, H. Lu, Z. Feng, and X. Xue. Null foley–sammon transform. *Pattern recognition*, 2006. 2, 3, 4

[14] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*. 2011. 2, 6

[15] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 6, 7

[16] E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015. 3, 8

[17] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 2, 6, 7, 8

[18] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE TPAMI*, 2013. 1, 2

[19] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 2, 6

[20] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 6, 8

[21] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 1, 2, 6, 7

[22] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 2, 6, 7, 8

[23] G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE TPAMI*, 2014. 1, 2

[24] G. Lisanti, I. Masi, and A. Del Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*. ACM, 2014. 2, 6, 7, 8

[25] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014. 3, 8

[26] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV Workshop*, 2012. 1, 2

[27] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE TIP*, 2014. 1, 2, 6, 7

[28] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 1, 2

[29] S. Paisitkriangkrai, C. Shen, and A. ven den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 2, 6, 7, 8

[30] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 1, 2, 3

[31] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010. 1, 2

[32] M. Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *ICML*, 2006. 3

[33] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li. Person re-identification by regularized smoothing kiss metric learning. *IEEE TCSVT*, 2013. 1, 2

[34] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Comput. Surv.*, 2013. 1

[35] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005. 6, 7

[36] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 1, 2, 4, 6, 7, 8

[37] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2006. 1

[38] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*, 2014. 1, 2, 6, 7

[39] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 1, 2, 7

[40] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2

[41] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 1, 2, 6, 7

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 6, 8

[43] W. Zheng, L. Zhao, and C. Zou. Foley-sammon optimal discriminant vectors using kernel approach. *IEEE TNN*, 2005. 2

[44] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE TPAMI*, 35(3), 2013. 1, 2

[45] X. Zhu. Semi-supervised learning literature survey. 2005. 5