

# A Semi-Supervised Maximum Margin Metric Learning Approach for Small Scale Person Re-identification

T M Feroz Ali

Indian Institute of Technology Bombay

ferozalitm@ee.iitb.ac.in

Subhasis Chaudhuri

Indian Institute of Technology Bombay

sc@ee.iitb.ac.in

## Abstract

*In video surveillance, person re-identification is the task of searching person images in non-overlapping cameras. Though supervised methods for person re-identification have attained impressive performance, obtaining large scale cross-view labeled training data is very expensive. However, unlabelled data is available in abundance. In this paper, we propose a semi-supervised metric learning approach that can utilize information in unlabelled data with the help of a few labelled training samples. We also address the small sample size problem that inherently occurs due to the few labeled training data. Our method learns a discriminative space where within class samples collapse to singular points, achieving the least within class variance, and then use a maximum margin criterion over a high dimensional kernel space to maximally separate the distinct class samples. A maximum margin criterion with two levels of high dimensional mappings to kernel space is used to obtain better cross-view discrimination of the identities. Cross-view affinity learning with reciprocal nearest neighbor constraints is used to mine new pseudo-classes from the unlabelled data and update the distance metric iteratively. We attain state-of-the-art performance on four challenging datasets with a large margin.*

## 1. Introduction

Person re-identification (re-ID) is the task of searching previously unseen persons across non-overlapping cameras. It is a challenging problem in video surveillance due to the large variations in illumination and viewpoint of personal appearances across cameras (refer Fig. 1).

Most of the existing works on person re-ID generally focus on the design of efficient features and representations [27, 26, 19] or distance metric learning [37, 2, 19, 42]. The features are designed to be robust against the viewpoints and illumination changes of persons. In order to attain better discrimination of the persons, distance metric learn-



Figure 1: Challenges in person re-identification: large variations in viewpoint and illumination across camera views (each row represents different camera view).

ing methods are used such that similar class samples come closer and dissimilar class samples get well separated.

Supervised methods for person re-ID have achieved tremendous improvement in performance. However, for the supervision, obtaining large cross-view annotated data for each of the targeted camera pairs is practically very expensive. Nevertheless, unlabelled data is easily available in abundance and hence methods that can efficiently utilize information from unlabelled data with minimal supervision becomes highly relevant for practical re-ID applications. Therefore, we look at the following problem: Given a few cross-view labelled and a major unlabelled training data (with input features), can we learn a good semi-supervised distance metric for small scale person re-identification, especially when there exists small sample size (SSS) problem, i.e. the number of training samples being very small compared to the feature dimension.

For metric learning methods, the small sample size (SSS) problem can create serious degradation since it creates singularity in the inherent matrices, making it non-invertible [2, 37, 42]. In such cases, application of regularization on the matrices or unsupervised dimensionality reduction on the input features makes it sub-optimal due to the loss of discriminative information. The SSS problem becomes more important for small scale semi-supervised case due

to the following reason: Though the deeply learned features [1, 36, 33, 35, 8, 38, 5] perform well with large training data, they struggle to perform with few training samples [2, 27, 41, 37] and hence hand-engineered features [27, 26, 19] are usually employed for small scale person re-identification [27, 19, 26]. However, these hand-engineered features are very high dimensional, typically in ten thousands, much higher than the number of training samples. Furthermore, for a semi-supervised case, the amount of labeled data available for training becomes even lesser compared to the supervised case.

In this paper, we propose a semi-supervised metric learning approach where it first learns a primary cross-view invariant discriminative subspace using the labelled data, by collapsing the same-class samples to singular points in a discriminative nullspace and then maximizing the margin between the classes in a high dimensional kernel space. A secondary discriminative subspace is then learned that maximizes the margin of distinct identities belonging to the unlabelled training data. New pseudo-classes are mined using cross-view affinity learning, which are then used for recursive update of the distance metric to further extract the information in the unlabelled data. Experimental results on four challenging datasets show that our approach outperforms the state-of-the-art methods, with a huge margin.

## 2. Related Methods

**Supervised methods:** There exists extensive work on supervised metric learning methods [2, 37, 42, 19, 39, 3, 7, 14]. Using the input features, they learn a Mahalanobis distance based similarity metric [14, 19, 20] or a discriminative space [2, 37, 39, 42] that well separates the dissimilar class samples and bring same class samples closer. Few metric learning methods [42, 2, 37] addressed the small sample size problem in re-ID. Zhang *et al.* [42] learned a discriminative nullspace to minimize the within class variance to the least. NK3ML [2] learned a more discriminative subspace which simultaneously collapses the within classes samples to singular points as well as maximizes the margin between distinct class samples. IRS [37] used an identity regression space for embedding each identity using one-hot feature coding.

Recently, the deep learning based methods [34, 18, 40, 30, 31, 32, 4, 6] have shown impressive performance on large scale labelled data. They generally learn discriminative features and distance metric jointly. However, they have limited performance on small scale training data. [2, 37].

**Semi-supervised methods:** In the literature, there exist few image based semi-supervised methods based on dictionary learning [24, 13] as well as metric learning [16, 42]. SSCDL[24] jointly learned coupled dictionaries from both

labelled and unlabelled images. Kodirov *et al.* [13] used iterative dictionary learning with graph Laplacian regularization. On the other hand, Li *et al.* [16] proposed a region metric learning approach with graph based label propagation to estimate positive neighbors from imbalanced unlabelled data. Zhang *et al.* [42] used a self-training approach using nullspace based distance metric to learn pseudo-classes from the unlabelled data and addressed the small sample size problem simultaneously. Dictionary learning and metric learning were jointly used in SBAL[23]. It used a Bayesian latent factor based attribute learning framework to determine a dictionary of attributes adaptively.

The most similar work to ours is [42]. However, unlike our approach, it does not use any margin maximization between distinct class samples as well as cross-view reciprocal constraints for same-class samples to extract pseudo-classes, and has lesser generalization to test data.

## 3. Our Approach

### 3.1. Main idea

The key idea of our approach is to automatically mine new classes from unlabelled data with the help of available cross-view labelled data. As the labelled data carry information from multiple cameras, they have cross-view identity discrimination information. Hence, we first learn a primary cross-view invariant discriminative space by first collapsing the within class samples to singular points in a discriminative nullspace and then maximizing the margin between the singular points using a maximum margin criterion over a high dimensional kernel space. Next, we obtain within view identity information of the unlabelled data automatically. Then we learn a secondary cross-view invariant distance metric that maximizes the margin between the within view distinct identities of the unlabelled data more efficiently. Over this secondary space, we then use cross-view reciprocal nearest neighbor constraints to discover new possible pseudo-classes. Most confident pseudo-classes are filtered and augmented to the available labelled training data, which is then used to learn a better cross-view discriminative space iteratively.

### 3.2. Nullspace Kernel Maximum Margin Metric Learning

We first use the given cross-view labelled samples of the training data to learn a cross-view invariant distance metric based on Nullspace Kernel Maximum Margin Metric learning (NK3ML)[2], which is one of the state-of-the-art metric learning methods that address small sample size problem in person re-identification. NK3ML first learns a discriminative nullspace that collapses the samples of the same class into a singular point, thus achieving the least value of zero within class variance. Using the nullspace, NK3ML then

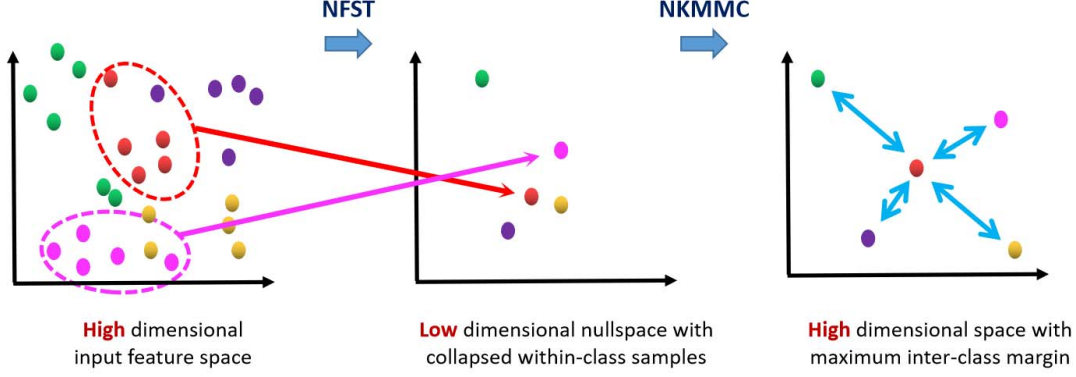


Figure 2: Illustration of Nullspace Kernel Maximum Margin Metric Learning (NK3ML). Each color corresponds to distinct classes. The figure is adapted from [2].

learns another discriminative space that maximizes the margin between the dissimilar class samples using a maximum margin criterion. It also uses mapping to very high dimensional space using kernels to obtain maximum inter-class discrimination. NK3ML is illustrated in Fig. 2. We briefly explain NK3ML, originally published in [2], in the next two subsections for the completeness of the descriptions.

### 3.2.1 Null-Foley Sammon Transform

To collapse the within-class samples to singular points, NK3ML makes use of Null Foley Sammon Transform (NFST)[10]. For given  $n$  training samples  $\mathbf{x} \in \mathbb{R}^d$  belonging to  $c$  classes, the objective of NFST is to find discriminants  $\mathbf{w} \in \mathbb{R}^d$  that maximize Fisher criterion  $\mathcal{J}_F(\mathbf{w})$  with orthonormal constraints.

$$\underset{\mathbf{w}^T \mathbf{w}_i = 0, \|\mathbf{w}_j\|=1}{\text{maximize}} \quad \mathcal{J}_F(\mathbf{w}_j) \quad i = 1, \dots, j-1, \quad (1)$$

where Fisher criterion  $\mathcal{J}_F(\mathbf{w})$  is defined as the ratio of between class scatter and within class scatter.

$$\mathcal{J}_F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}. \quad (2)$$

$\mathbf{S}_b$  is the between class scatter matrix and  $\mathbf{S}_w$  is the within class scatter matrix. NFST optimizes the cost function  $\mathcal{J}_F(\mathbf{w})$  by forcing the denominator term  $\mathbf{w}^T \mathbf{S}_w \mathbf{w}$  to zero while ensuring the numerator term  $\mathbf{w}^T \mathbf{S}_b \mathbf{w}$  is positive.

$$\mathbf{w}^T \mathbf{S}_w \mathbf{w} = 0, \quad (3)$$

$$\mathbf{w}^T \mathbf{S}_b \mathbf{w} > 0. \quad (4)$$

Thus it attempts to make  $\mathcal{J}_F(\mathbf{w}) \rightarrow \infty$ . The least value of zero within class variance is attained which results in collapsing the samples of each class to singular points. The total class scatter matrix  $\mathbf{S}_t$  is defined as  $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$ .

If  $\mathbf{Z}_t$  and  $\mathbf{Z}_w$  are the nullspaces of  $\mathbf{S}_t$  and  $\mathbf{S}_w$  respectively, and  $\mathbf{Z}_t^\perp$  is the orthogonal compliment of  $\mathbf{Z}_t$ , then it has been shown that  $\mathbf{w} \in (\mathbf{Z}_t^\perp \cap \mathbf{Z}_w)$  [10]. The discriminants  $\mathbf{w}$  are called Null Projecting Directions (NPDs). When small sample size (SSS) problem occurs, there exists exactly  $c-1$  NPDs, where  $c$  is the number of classes.

The procedure for obtaining the Null Projecting Directions (NPDs) are as follows. Please refer [2, 10] for the detailed theory.

1. Take the zero mean data  $\mathbf{Y}_t = \{\mathbf{x}_1 - \mathbf{m}, \dots, \mathbf{x}_n - \mathbf{m}\}$ , where  $\mathbf{m}$  is the mean of the samples.
2. Find orthonormal basis for  $\mathbf{Y}_t$  to get  $\mathbf{U} = (\theta_1, \dots, \theta_{n-1})$  using Gram Schmidt orthogonalization procedure.
3. Find nullspace of  $\mathbf{U}^T \mathbf{S}_w \mathbf{U}$  to get orthonormal basis vectors  $\beta$ , where each basis vector is of form  $\beta = (b_1, \dots, b_{n-1})$ . Since the dimension of the nullspace is  $c-1$ , there would be  $c-1$  solutions for  $\beta$ .
4. Each  $\beta$  corresponds to an NPD  $\mathbf{w}$ , which is obtained as

$$\mathbf{w} = b_1 \theta_1 + \dots + b_{n-1} \theta_{n-1} = \mathbf{U} \beta. \quad (5)$$

The  $c-1$  NPDs constitute the discriminative nullspace where the samples of each class collapse to one singular point. Hence we have  $c$  singular points in the  $c-1$  dimensional discriminative nullspace. A general input sample with a high dimensional feature can be mapped to the low dimensional discriminative nullspace using a projection matrix  $\mathbf{W}_N \in \mathbb{R}^{d \times (c-1)}$ , which is formed by the  $c-1$  NPDs as its columns.

### 3.2.2 Kernel Maximum Margin Criterion

Though NFST is optimal in minimizing the within class variance, it fails to maximize the between class variance. As a result, singular points of distinct classes can lie closer in the nullspace. This lack of between class discrimination can

result in low generalization for the test data. In order to address this issue, NK3ML uses a *Maximum Margin Criterion* (MMC)[15] to learn another discriminative space using the nullspace that separates the singular points with maximum margin. In order to attain further separation of the singular points, NK3ML make benefit of kernel based technique that maps the singular points to a very high dimensional space (using an appropriate kernel), such that they are more likely to get well separated with maximum margin.

The Maximum Margin Criterion for  $c$  classes is defined as

$$\mathcal{J}_M = \frac{1}{2} \sum_{i=1}^c \sum_{j=1}^c p_i p_j d(C_i, C_j), \quad (6)$$

where the inter-class margin (or distance)  $d(C_i, C_j)$  between the classes  $C_i$  and  $C_j$  is defined as

$$d(C_i, C_j) = d(\mathbf{m}_i, \mathbf{m}_j) - s(C_i) - s(C_j). \quad (7)$$

$\mathbf{m}_i$  represents the mean of the class  $i$ , and  $d(\mathbf{m}_i, \mathbf{m}_j)$  is the distance between the class means. The scatter  $s(C_i)$  of class  $i$  is estimated as  $s(C_i) = \text{tr}(\mathbf{S}_i)$ , where  $\mathbf{S}_i$  is its within class scatter matrix. The margin is solved to get  $d(C_i, C_j) = \text{tr}(\mathbf{S}_b - \mathbf{S}_w)$ .

A set of discriminants  $\{\mathbf{v}_k | k = 1, \dots, l\}$  can be than found that maximizes the margin, using the below optimization problem.

$$\begin{aligned} & \underset{\mathbf{v}_k}{\text{maximize}} && \sum_{k=1}^l \mathbf{v}_k^T (\mathbf{S}_b - \mathbf{S}_w) \mathbf{v}_k, \\ & \text{subject to} && \mathbf{v}_k^T \mathbf{v}_k = 1, \quad k = 1, \dots, l. \end{aligned} \quad (8)$$

The optimal solutions are composed of the first  $l$  eigenvectors of  $\mathbf{S}_b - \mathbf{S}_w$ .

Instead of using the above Maximum Margin Criterion (MMC) directly, NK3ML utilizes the following *Normalized Kernel Maximum Margin Criterion* (NKMMC) to take benefit of kernel based methods. NKMMC learns non-linear discriminants by mapping input sample  $\mathbf{y}$  to very high dimensional feature space  $\mathcal{F}$  using a non-linear mapping  $\phi(\mathbf{y})$ , which is implicitly determined by a kernel function  $k(\mathbf{y}_i, \mathbf{y}_j) = \langle \phi(\mathbf{y}_i), \phi(\mathbf{y}_j) \rangle$ . Given  $n$  training samples, the kernel matrix  $\mathbf{K} \in \mathbb{R}^{n \times n}$  can be calculated. Then a discriminant vector  $\mathbf{v}_k \in \mathcal{F}$  can be expressed as  $\mathbf{v}_k = \sum_{j=1}^n \alpha_k^{(j)} \phi(\mathbf{y}_j)$ , where  $\alpha_k^{(j)}$  is the  $j$ th component of the expansion coefficient vector  $\alpha_k \in \mathbb{R}^n$ . Then (8) can be kernelized to obtain the optimization problem corresponding to NKMMC as:

$$\begin{aligned} & \underset{\alpha_k}{\text{maximize}} && \sum_{k=1}^l \alpha_k^T (\mathbf{P} - \mathbf{Q}) \alpha_k, \\ & \text{subject to} && \alpha_k^T \mathbf{K} \alpha_k = 1, \end{aligned} \quad (9)$$

where  $\mathbf{Q} := \sum_{i=1}^c \frac{1}{n} \mathbf{K}_i (\mathbf{I}_{n_i} - \frac{1}{n_i} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T) \mathbf{K}_i^T$ . The matrix  $\mathbf{K}_i \in \mathbb{R}^{n \times n_i}$  corresponding to the  $i$ th class with  $n_i$  samples is defined as  $\mathbf{K}_i(p, q) := k(\mathbf{y}_p, \mathbf{y}_q^{(i)})$ .  $\mathbf{I}_{n_i}$  is  $n_i$  dimensional identity matrix,  $\mathbf{1}_{n_i}$  is  $n_i$  dimensional column vector of ones,  $\mathbf{P} = \sum_{i=1}^c \frac{1}{n_i} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^T$ ;  $\tilde{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^c n_i \tilde{\mathbf{m}}_i$  and  $\tilde{\mathbf{m}}_i^{(j)} = \frac{1}{n_i} \sum_{\mathbf{y} \in C_i} k(\mathbf{y}, \mathbf{y}_j)$ . The optimal solutions  $\alpha_k$  are obtained by finding the *generalized* eigenvectors corresponding to the  $l$  largest *generalized* eigenvalues of  $(\mathbf{P} - \mathbf{Q})$  and  $\mathbf{K}$ . In order to obtain maximum margin between the samples, all the eigenvectors with positive eigenvalues are chosen as the discriminants [2].

In effect, the NK3ML first maps the input samples from the very high dimensional input feature space to a very low dimensional space, where samples of each class collapse to singular points. Then it maps the data from the low dimensional space to a very high dimensional space using appropriate kernels to obtain maximum margin between the singular points. We refer discriminative space of NK3ML, obtained using the labelled training data as the *primary* discriminative space.

### 3.3. Pseudo Class Mining

We next utilize the *unlabelled* data and the primary discriminative space to mine new pseudo-classes in following *five* stages.

In the *first* stage, we map each of the unlabelled training data  $\mathbf{z} \in \mathbb{R}^d$  to the primary space of NK3ML, to obtain  $\mathbf{u} \in \mathbb{R}^l$  in two steps: (1) projecting  $\mathbf{z}$  onto the NFST nullspace as

$$\mathbf{y} = \mathbf{W}_{N\mathbf{z}}^T \mathbf{z} \quad (10)$$

and then mapping it to each of the discriminant  $\mathbf{v}_k$  of NKMMC.

$$\mathbf{v}_k^T \phi(\mathbf{y}) = \left( \sum_{j=1}^n \alpha_k^{(j)} \phi(\mathbf{y}_j) \right)^T \phi(\tilde{\mathbf{y}}) \quad (11)$$

$$= \sum_{j=1}^n \alpha_k^{(j)} k(\mathbf{y}_j, \tilde{\mathbf{y}}) \quad (12)$$

**Within-view identity labeling:** In the second stage, we propose to use the original input images/videos of the unlabelled samples from each of the given camera to automatically identify the images that belong to distinct persons. Though the true identities of the images are not known, standard intra-camera tracking models with spatial and temporal constraints of the cameras can be used to extract person tracklets and annotate every tracklet with a distinct stamp [17]. Thus within view identity information can be extracted to group the samples to distinct classes. Note that the cross-view identity association is still not known and need to be estimated.

### 3.3.1 Anchor-view identity discrimination

In the *third* stage, we choose the camera with the maximum number of distinct person appearances as an *anchor camera*. Then we learn a *secondary* discriminative space that better maximizes the distance between the within view identities of the anchor camera. It should be noted that though the primary space is already learned using the initial labelled training data  $\mathbf{x}$  to discriminate distinct identities, it could not use the data  $\mathbf{z}$  for its training. Hence certain distinct identities from data  $\mathbf{z}$  may still be not well separated. Hence we learn the secondary discriminative space that is more suitable in separating the distinct identities in  $\mathbf{z}$  coming from the anchor camera.

To that end, we use the same Normalized Kernel Maximum Margin Criterion (NKMMC) in (9), however, by using the anchor camera samples  $\mathbf{u}^A$  lying in the primary discriminative space as the training data. As seen previously, the Normalized Kernel Maximum Margin Criterion (NKMMC) can ensure separation of distinct training classes with maximum margin. Given  $\tilde{n}$  samples belonging to  $\tilde{c}$  classes from the anchor camera, we map the its data samples  $\mathbf{u}^A$  to a secondary kernel space  $\tilde{\mathcal{F}}$ , using a secondary non-linear mapping function  $\tilde{\phi}(\mathbf{u}^A)$ . Using this space, it learns  $f$  discriminant vectors  $\{\tilde{\mathbf{v}}_k\}_{k=1}^f$  that maximize  $\mathcal{J}_M$ . Since each discriminant  $\tilde{\mathbf{v}}_k$  lies in the span of the mapped samples, we have  $\tilde{\mathbf{v}}_k = \sum_{j=1}^{\tilde{n}} \epsilon_k^{(j)} \tilde{\phi}(\mathbf{u}_j^A)$  where  $\epsilon_k^{(j)}$  is the  $j$ th element of the expansion vector  $\epsilon_k \in \mathcal{R}^{\tilde{n}}$ . Let  $\tilde{\mathbf{K}} \in \mathcal{R}^{\tilde{n} \times \tilde{n}}$  be the kernel matrix computed using the secondary kernel function  $\tilde{k}(\mathbf{u}_i^A, \mathbf{u}_j^A)$ . Then the optimization problem based on NKMMC to learn the optimal discriminants  $\tilde{\mathbf{v}}_k$  is

$$\begin{aligned} & \underset{\epsilon_k}{\text{maximize}} && \sum_{k=1}^f \epsilon_k^T (\tilde{\mathbf{P}} - \tilde{\mathbf{Q}}) \epsilon_k \\ & \text{subject to} && \epsilon_k^T \tilde{\mathbf{K}} \epsilon_k = 1 \end{aligned} \quad (13)$$

where  $\tilde{\mathbf{Q}} = \sum_{i=1}^{\tilde{c}} \frac{1}{\tilde{n}} \tilde{\mathbf{K}}_i (\mathbf{I}_{\tilde{n}_i} - \frac{1}{\tilde{n}_i} \mathbf{1}_{\tilde{n}_i} \mathbf{1}_{\tilde{n}_i}^T) \tilde{\mathbf{K}}_i^T$ ,  $\tilde{\mathbf{P}} = \sum_{i=1}^{\tilde{n}} \frac{1}{\tilde{c}_i} (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^T$ ,  $\tilde{\mathbf{m}} = \frac{1}{\tilde{c}} \sum_{i=1}^{\tilde{c}} \tilde{c}_i \tilde{\mathbf{m}}_i$ ,  $\tilde{\mathbf{m}}_i^{(j)} = \frac{1}{\tilde{c}_i} \sum_{\mathbf{u}^A \in C_i} \tilde{k}(\mathbf{u}^A, \mathbf{u}_j^A)$ . Similar to (9), the optimal solutions are composed of the generalized eigen vectors corresponding to the leading  $f$  generalized eigenvalues of  $\tilde{\mathbf{P}} - \tilde{\mathbf{Q}}$  and  $\tilde{\mathbf{K}}$ . We use all the eigenvectors with positive eigenvalues to attain a good inter-class margin.

We then map all the unlabelled samples  $\mathbf{u}$  from all the cameras to this secondary space. A general sample  $\mathbf{u}$  can be projected onto the discriminant vector  $\tilde{\mathbf{v}}_k$  as

$$\tilde{\mathbf{v}}_k^T \tilde{\phi}(\mathbf{u}) = \left( \sum_{j=1}^{\tilde{n}} \epsilon_k^{(j)} \tilde{\phi}(\mathbf{u}_j^A) \right)^T \tilde{\phi}(\mathbf{u}) = \sum_{j=1}^{\tilde{n}} \epsilon_k^{(j)} \tilde{k}(\mathbf{u}_j^A, \mathbf{u}). \quad (14)$$

All the distinct identity samples coming from the anchor camera are now guaranteed to lie well separated with

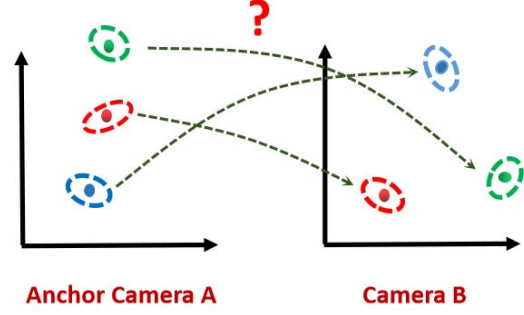


Figure 3: Cross-view affinity learning of the identities from the anchor camera with respect to another camera (each color represents distinct identities).

maximum margin in the secondary space. It should be noted that for both NK3ML and anchor-view identity discrimination, we used mapping to kernel space. Hence there occur two levels of subsequent high dimensional kernel mappings in our approach, which helps in improving the separation of the underlying identities with maximum margin.

### 3.3.2 Cross-view affinity learning

As the secondary space is learned over the view invariant primary space, the secondary space has two important properties: (1) anchor-view identity discrimination property, as the secondary space is better trained with respect to the unlabelled training data identities seen in the anchor camera; (2) view invariant property, which naturally occurs as the secondary metric is learned over the view invariant primary space. As a consequence of the two properties existing together in the secondary space, the cross-view unlabelled samples from other cameras tend to lie closer to their corresponding true identities from the anchor camera, as well as all the distinct identity samples get well separated with maximum margin. This results in localization of the cross-view same class samples and hence better discrimination of the cross-view data is achieved in the secondary space.

In the *fourth* stage, we utilize this secondary space to perform cross-view affinity learning for each of the identities of the anchor camera against all the identities in other cameras, as illustrated in Fig. 3. We then mine new pseudo-classes across views by using *cross-view k-reciprocal-nearest-neighbor constraint* over the affinities.

**Reciprocal-nearest-neighbor constraint:** One way to mine new pseudo-classes in the secondary space is to retrieve nearest neighbors across views and perform k-nearest



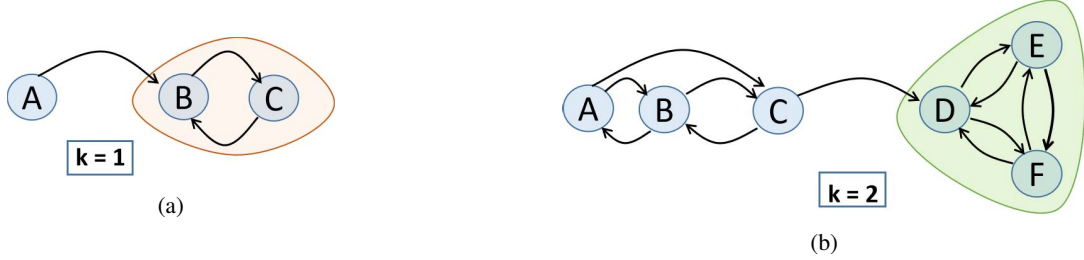


Figure 4: Illustration of  $k$ -reciprocal nearest neighbor constraint: The unidirectional  $k$ -nearest neighbors of each node is shown using the arrows. (a) For  $k=1$ , the node  $A$  does not have any reciprocal neighbors as its unidirectional neighbor  $B$  does not have  $A$  as the unidirectional neighbor. However, nodes  $B$  and  $C$  are  $k$ -reciprocal neighbors as each of them is a  $k$ -nearest neighbor of the other. (b) For  $k=2$ , nodes  $A$  and  $C$  are  $k$ -reciprocal neighbors of  $B$ , while  $A$  and  $C$  are not  $k$ -reciprocal neighbors of each other. However, nodes  $C$ ,  $D$  and  $E$  are all  $k$ -reciprocal nearest neighbors of each other. Figure 4(b) is adapted from [28].

neighbor clustering for each of the identities of the anchor camera against the other views. Though we have localized the identities in the secondary space, however, due to the large variations in illumination and viewpoint of persons across views, instead of the true matches, irrelevant objects may exist among the  $k$ -nearest neighbors, which can degrade the mined pseudo-classes. In order to avoid such degradation and to achieve a higher and robust performance, we propose to use  $k$ -reciprocal-nearest-neighbor constraint [12, 28], across views, for mining new classes. It helps in finding cross-view samples which are very likely to be related to the queries, by measuring mutual similarity.

The  $k$ -nearest neighbors of a query sample  $q$  is defined as

$$N_k(q) = \{g_1, g_2, \dots, g_k\}, \quad |N_k(q)| = k \quad (15)$$

where  $|X|$  represents the cardinality of the set  $X$ , and  $g_1, g_2, \dots, g_k$  are the top- $k$  search retrieval results for  $q$ . Then  $k$ -reciprocal nearest neighbors of  $q$  are defined as

$$R_k(q) = \{g_i \mid (g_i \in N_k(q)) \wedge (q \in N_k(g_i))\}. \quad (16)$$

The concept of  $k$ -reciprocal nearest neighbors is illustrated in Fig. 4. Two samples are  $k$ -reciprocal nearest neighbors, if for each of the samples as queries, the other sample exists in the former’s top  $k$ -retrieved closest neighbors. It is a stronger constraint of closeness compared to the conventional  $k$ -nearest neighbor relationship. The  $k$ -reciprocal nearest neighbor constraint considers local densities around the queries and their retrieved matches. Hence it can deal with uneven distribution of samples in the space, thus becoming more robust to false retrieval results and outliers.

For cross-view affinity learning for the anchor camera, we use the *cross-view*  $k$ -reciprocal-nearest-neighbor constraint for searching the most likely matches. The cross-view  $k$ -reciprocal-nearest-neighbors for any identity  $q^A$  of the anchor camera with respect to another camera  $B$  is de-

finied as

$$R_k^*(q^A) = \{g_i \mid g_i \in R_k(q^A), R_k(q^A) \in C_B\}, \quad (17)$$

where  $C_B$  is the set of all identities in camera  $B$ .

### 3.3.3 Recursive metric learning

In the *fifth* stage, the newly mined cross-view pseudo-classes are ranked according to the increasing order of their within class affinities (refer Fig. 5). The pseudo-classes with affinities above the upper quartile are filtered and augmented to the existing labelled training data. Next, we learn a new primary discriminative space using NK3ML with the new set of labelled training classes. All the above steps are repeated recursively till there are no more cross-view pseudo class available in the unlabelled training data that satisfy the  $k$ -reciprocal nearest neighbor constraint.

The strategy of selecting a fraction of the top ranked pseudo-classes in each iteration is to ensure that the pseudo-classes in the initial iterations are the best, which in turn help in mining more reliable pseudo-classes in the subsequent iterations. This also ensures that model drifting due to errors in the mined pseudo-classes is better avoided, which is of utmost importance in self-training [43] based methods.

The proposed method thus utilizes the information from both labelled and unlabelled data to learn a discriminative distance metric effectively. The approach ensures that within class variance is minimized to the least and the distinct identity samples get well separated with maximum margin to get better generalization for test data. Note that both NK3ML and NKMMC that is used for learning the primary and secondary discriminative spaces have closed form solutions, which does not need regularization or unsupervised dimensionality reduction. Hence our proposed method efficiently handles the small sample size (SSS) problem in person re-identification.

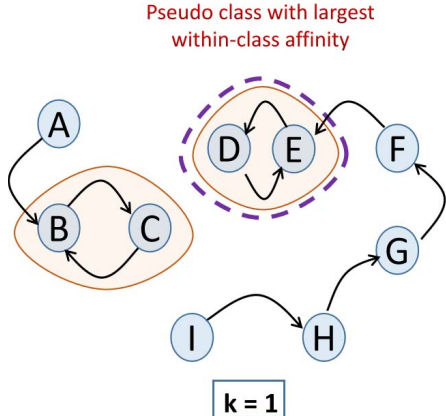


Figure 5: Illustration of the pseudo-class mining: At each iteration, the pseudo-classes with largest intra-class affinities are given new labels and augmented to the labelled training data.

## 4. Experiments

**Datasets and Evaluation Protocol:** We use four small scale datasets including VIPeR [9], PRID450S [29], GRID [25] and PRID2011 [11] for our experiments. They respectively contain 632, 450, 250 and 200 persons, captured from two disjoint cameras, with one image per person in each camera. Following the standard protocol [42, 13, 24, 23], the dataset identities are divided equally into two halves; one forming the training set and the other forming the test set. A random one third of the training set is labelled and the rest is unlabelled. The test images are divided into half forming the query(probe) set and the gallery set. Images from one camera form the query set and the rest form the gallery. The gallery of GRID and PRID2011 datasets have an additional 775 and 549 person images, respectively, which are different from their query set identities. Each image of the query set is matched against the gallery. The distance scores are ranked and sorted. Rank-N matching performance is then calculated as the percentage of true match being present in the first N search results. The above procedure is repeated 10 times and the average performance is reported.

**Features:** In our method, we use standard feature descriptors including LOMO[19], GOG[26] and WHOS[22]. They are of dimensions 26960, 27622 and 2960 respectively. In addition, we use a new feature descriptor named LOMO<sup>†</sup>, which is the LOMO feature obtained without using Retinex[19] transformation of the images, in order to take advantage of color diversity, similar to GOG. Re-ID datasets have large variations in illumination and background. Hence we use specific features for each dataset, to better capture their intrinsic characteristics. For VIPeR

Methods	Rank1	Rank5	Rank10	Rank20
KISSME[14]	18.50	43.70	57.90	74.5
kLFDA[39]	27.50	56.00	69.60	82.60
kCCA[21]	24.60	56.20	71.70	85.60
XQDA[19]	27.63	55.76	68.67	81.90
MLAPG[20]	25.82	48.54	59.24	70.22
kNFST[42]	26.46	54.72	67.66	81.30
SSCDL[24]	25.60	53.70	68.20	83.60
IterativeLap[13]	30.06	46.84	56.74	67.31
semi-kNFST[42]	31.68	59.40	72.78	84.91
SBAL [23]	33.60	-	-	-
Ours	64.75	88.83	94.46	97.78

Table 1: Performance comparison with state-of-the-art results on VIPeR dataset.

Methods	Rank1	Rank5	Rank10	Rank20
KISSME[14]	5.10	15.20	24.10	40.10
kLFDA[39]	14.10	33.70	44.00	56.20
kCCA[21]	5.30	15.70	25.80	37.00
XQDA[19]	13.50	30.60	40.20	52.50
kNFST[42]	12.40	28.40	37.50	49.40
IterativeLap[13]	22.10	45.30	56.50	66.30
semi-kNFST[42]	24.70	46.80	58.20	68.20
SBAL [23]	24.40	-	-	-
Ours	29.90	53.10	61.60	72.20

Table 2: Performance comparison with state-of-the-art results on PRID2011 dataset.

Methods	Rank1	Rank5	Rank10	Rank20
XQDA[19]	9.92	22.08	29.92	39.92
kNFST[42]	9.60	20.08	27.60	37.60
IterativeLap[13]	10.96	28.64	37.92	48.24
Ours	23.36	39.76	52.32	63.84

Table 3: Performance comparison with state-of-the-art results on GRID dataset.

dataset, we use concatenated LOMO and WHOS features, while for other datasets, we use concatenated LOMO, LOMO<sup>†</sup> and GOG features. Naturally, small sample size problem exists due to the large dimension of the features.

**Parameters:** We use RBF kernel for both NK3ML, as well as for NKMMC in anchor-view identity discrimination. The width of the RBF kernel is automatically set using a mean function of the training samples, similar to [42] and [39]. We set  $k=1$  for cross-view reciprocal nearest neighbor constraint.

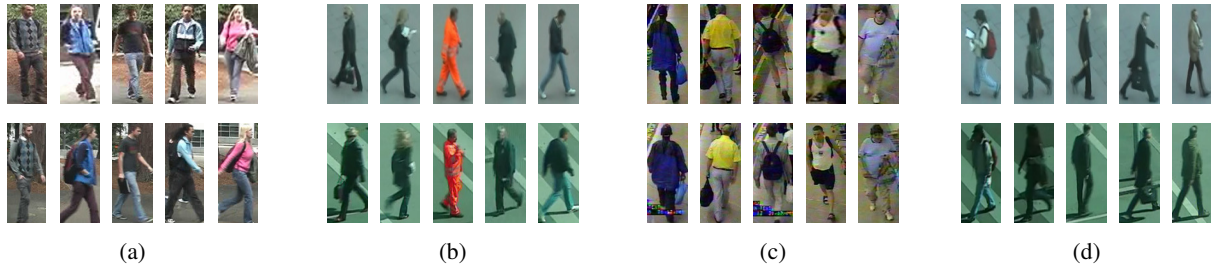


Figure 6: Sample images from (a) VIPeR, (b) PRID2011, (c) GRID and (d) PRID450S datasets.

Methods	Rank1	Rank5	Rank10	Rank20
IterativeLap[13]	24.31	42.13	51.78	61.82
XQDA[19]	45.07	71.64	81.78	90.13
MLAPG[20]	39.91	67.47	78.62	88.62
kNFST[42]	41.87	70.04	81.07	90.80
Ours	63.78	86.49	92.53	96.80

Table 4: Performance comparison with state-of-the-art results on PRID450S dataset.

Methods	VIPeR	PRID2011	GRID	PRID450S
IterativeLap[13]	1680	975	1214	871
Ours	48	21	33	58

Table 5: Execution time comparison (in seconds) for various datasets.

### 4.1. Comparative Results

The experimental results of our method evaluated on datasets VIPeR, PRID2011, GRID and PRID450S are provided in Table 1, 2, 3 and 4, respectively. We compare our results with state-of-the-art semi-supervised methods including SBAL [23], semi-kNFST[42], IterativeLap [13] and SSCDL[24]. RegionMetric [16] uses pre-training using other datasets and hence is not directly comparable with our method. We also compare with supervised methods including KISSME[14], kLFDA[39], kCCA[21], XQDA[19], and kNFST[42], whose results are provided in separate rows. These methods are trained using only the labelled training data. The results of KISSME, kLFDA, and kCCA are obtained from [13]. For all the datasets, our approach outperforms all the compared methods, with high margin, at all ranks. Especially for VIPeR dataset, we attain an improvement of more than 30% in rank-1 accuracy, compared to the second best method SBAL[23]. Similarly, a rank-1 improvement margin of more than 5%, 12%, and 22% is attained for PRID2011, GRID and PRID450S datasets, respectively. It is worth noting that SBAL[23] uses deep learning based features and our method still outperforms

it with a good margin. This emphasizes the limitation of deep learning methods for semi-supervised person re-identification with minimal training data. Our method is able to utilize the information from both the available labelled and unlabelled data, while efficiently handling the small sample size problem in person re-identification.

### 4.2. Execution Time

We evaluate the total run time of our method in Table 5. IterativeLap[13] is the only semi-supervised method whose source code we could find publicly available and hence we use the same for comparison. Both the methods are implemented in MATLAB and evaluated in a PC with 3.4GHz CPU and 32GB memory. It can be seen that our method has low time complexity and is more suitable for practical implementation.

## 5. Conclusion

In this paper, we proposed a semi-supervised approach for person re-identification that learns an efficient cross-view invariant distance metric for handling small training sample size problem. Using a few labelled training samples, the method first learns a primary cross-view discriminative space by collapsing the training classes into singular points and then maximizing the margin between the singular points. Later a secondary discriminative space is learned that separates the distinct identities in the unlabelled data. Cross-view affinity learning with k-reciprocal nearest neighbor constraints is used to mine new pseudo-classes, which are augmented to the labelled training data to learn a better cross-view discriminative distance metric recursively. The method does not need regularization or unsupervised dimensionality reduction and handles small sample size problem efficiently. Experiments on challenging datasets confirm the superiority of our approach compared to the state-of-the-art methods.

**Acknowledgment:** This research work is supported under Visvesvaraya PhD Scheme for Electronics and IT, by Ministry of Electronics and Information Technology (MeitY), Government of India.



## References

- [1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. *CVPR*, 2015.
- [2] T M Feroz Ali and Subhasis Chaudhuri. Maximum margin metric learning over discriminative nullspace for person re-identification. *ECCV*, 2018.
- [3] T M Feroz Ali, Kalpesh K Patel, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multiple kernel fisher discriminant metric learning for person re-identification. *ICVGIP*, 2018.
- [4] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018.
- [5] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *CVPR*, 2017.
- [6] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2017.
- [7] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *PAMI*, 2017.
- [8] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. *CVPR*, 2016.
- [9] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *ECCV*, 2008.
- [10] Y.F. Guo, L. Wu, H. Lu, Z. Feng, and X. Xue. Null foley-sammon transform. *Pattern Recognition*, 2006.
- [11] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. *Image analysis*.
- [12] Herve Jegou, Hedi Harzallah, and Cordelia Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *CVPR*, 2007.
- [13] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. 2015.
- [14] M. Kstinger, M. Hirzer and P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. *CVPR*, 2012.
- [15] H. Li, T. Jiang, and K. Zheng. Efficient and robust feature extraction by maximum margin criterion. *NIPS*, 2004.
- [16] Jiawei Li, Andy J Ma, and Pong C Yuen. Semi-supervised region metric learning for person re-identification. *IJCV*, 2018.
- [17] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. 2018.
- [18] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.
- [19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. *CVPR*, 2015.
- [20] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. *ICCV*, 2015.
- [21] G. Lisanti, I. Masi, and A. D. Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras. ACM*, 2014.
- [22] G. Lisanti, I. Masi, and A. D. Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE TPAMI*, 2014.
- [23] Wenhe Liu, Xiaojun Chang, Ling Chen, and Yi Yang. Semi-supervised bayesian attribute learning for person re-identification. *AAAI*, 2018.
- [24] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. 2014.
- [25] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. *CVPR*, 2009.
- [26] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. *CVPR*, 2016.
- [27] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptors with application to person re-identification. *PAMI*, 2019.
- [28] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR*, 2011.
- [29] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, 2014.
- [30] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. *CVPR*, 2018.
- [31] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, 2018.
- [32] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *CVPR*, 2018.
- [33] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. *ECCV*, 2016.
- [34] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.
- [35] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human reidentification. *ECCV*, 2016.
- [36] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human reidentification. *ECCV*, 2016.
- [37] Hanxiao Wang, Xiatian Zhu, Shaogang Gong, and Tao Xiang. Person re-identification in identity regression space. *IJCV*, 2018.
- [38] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *CVPR*, 2016.
- [39] F. Xiong, M. Gou, O. Camps, and M. Sznajder. Person re-identification using kernel-based metric learning methods. *ECCV*, 2014.

- [40] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. *CVPR*, 2018.
- [41] Yang Yang, Zhen Lei, Jinqiao Wang, and Stan Z Li. In defense of color names for small-scale person re-identification.
- [42] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. *CVPR*, 2016.
- [43] X. Zhu. Semi-supervised learning literature survey. 2005.