UC Riverside
UC Riverside Previously Published Works

Title

Unsupervised Adaptive Re-Identification in Open World Dynamic Camera Networks

Permalink

https://escholarship.org/uc/item/5jz4k7r5

Authors

Panda, Rameswar
Bhuiyan, Amran
Murino, Vittorio
et al.

Publication Date

2017-07-01

DOI

10.1109/cvpr.2017.151

Peer reviewed

Unsupervised Adaptive Re-identification in Open World Dynamic Camera Networks

Rameswar Panda[1,*]   Amran Bhuiyan[2,*,†]   Vittorio Murino[2]   Amit K. Roy-Chowdhury[1]

[1] Department of ECE   [2]Pattern Analysis and Computer Vision (PAVIS)
UC Riverside   Istituto Italiano di Tecnologia, Italy

{rpand002@,amitrc@ece.}ucr.edu   {amran.bhuiyan,vittorio.murino}@iit.it

Abstract

Person re-identification is an open and challenging problem in computer vision. Existing approaches have concentrated on either designing the best feature representation or learning optimal matching metrics in a static setting where the number of cameras are fixed in a network. Most approaches have neglected the dynamic and open world nature of the re-identification problem, where a new camera may be temporarily inserted into an existing system to get additional information. To address such a novel and very practical problem, we propose an unsupervised adaptation scheme for re-identification models in a dynamic camera network. First, we formulate a domain perceptive re-identification method based on geodesic flow kernel that can effectively find the best source camera (already installed) to adapt with a newly introduced target camera, without requiring a very expensive training phase. Second, we introduce a transitive inference algorithm for re-identification that can exploit the information from best source camera to improve the accuracy across other camera pairs in a network of multiple cameras. Extensive experiments on four benchmark datasets demonstrate that the proposed approach significantly outperforms the state-of-the-art unsupervised learning based alternatives whilst being extremely efficient to compute.

1. Introduction

Person re-identification (re-id), which addresses the problem of matching people across non-overlapping views in a multi-camera system, has drawn a great deal of attention in the last few years [25, 59, 74]. Much progress has been made in developing methods that seek either the best feature representations (e.g., [64, 39, 3, 42]) or propose to learn optimal matching metrics (e.g., [37, 52, 37, 67]). While they have obtained reasonable performance on com-
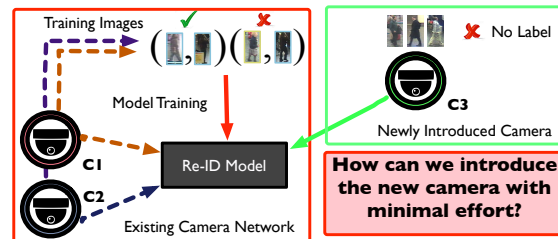

Figure 1. Consider an existing network with two cameras C_1 and C_2 where we have learned a re-id model using pair-wise training data from both of the cameras. During the operational phase, a new camera C_3 is introduced to cover a certain area that is not well covered by the existing 2 cameras. Most of the existing methods do not consider such dynamic nature of a re-id model. In contrast, we propose to adapt the existing re-id model in an unsupervised way: *what is the best source camera to pair with the new camera and how can we exploit the best source camera to improve the matching accuracy across the other cameras.*

monly used benchmark datasets (e.g., [51, 16, 12, 73]), we believe that these approaches have not yet considered a fundamental related problem: *Given a camera network where the inter-camera transformations/distance metrics have been learned in an intensive training phase, how can we incorporate a new camera into the installed system with minimal additional effort?* This is an important problem to address in realistic open-world re-identification scenarios, where a new camera may be temporarily inserted into an existing system to get additional information.

To illustrate such a problem, let us consider a scenario with $\mathcal{N}$ cameras for which we have learned the optimal pairwise distance metrics, so providing high re-id accuracy for all camera pairs. However, during a particular event, a new camera may be temporarily introduced to cover a certain related area that is not well-covered by the existing network of $\mathcal{N}$ cameras (See Fig. 1 for an example). Despite the dynamic and open nature of the world, almost all work in re-identification assumes a *static* and *closed* world model of the re-id problem where the number of cameras is fixed in a network. Given a newly introduced camera, traditional re-

---

id methods will try to relearn the inter-camera transformations/distance metrics using a costly training phase. This is impractical since labeling data in the new camera and then learning transformations with the others is time-consuming, and defeats the entire purpose of temporarily introducing the additional camera. Thus, there is a pressing need to develop *unsupervised* learning models for re-identification that can work in such dynamic camera networks.

Domain adaptation [15, 32] has recently been successful in many classical vision problems such as object recognition [24, 56, 19] and activity classification [48, 69] with multiple classes or domains. The main objective is to scale learned systems from a source domain to a target domain without requiring prohibitive amount of training data in the target domain. Considering a newly introduced camera as target domain, we pose an important question in this paper: *Can unsupervised domain adaptation be leveraged upon for re-identification in a dynamic camera network?*

Unlike classical vision problems, e.g., object recognition [56], domain adaptation for re-id has additional challenges. A central issue in domain adaptation is that from *which source to transfer*. When there is only one source of information available which is highly relevant to the task of interest, then domain adaptation is much simpler than in the more general and realistic case where there are multiple sources of information of greatly varying relevance. Re-identification in a dynamic network falls into the latter, more difficult, case. Specifically, given multiple source cameras (already installed) and a target camera (newly introduced), *how can we select the best source camera to pair with the target camera?* The problem can be easily extended to multiple additional cameras being introduced.

Moreover, once the best source camera is identified, *how can we exploit this information to improve the re-identification accuracy of other camera pairs?* For instance, let us consider $\mathtt{C}_1$ being the best source camera for the newly introduced camera $\mathtt{C}_3$ in Fig. 1. Once the pair-wise distance metric between $\mathtt{C}_1$ and $\mathtt{C}_3$ is obtained, can we exploit this information to improve the re-id accuracy across $(\mathtt{C}_2-\mathtt{C}_3)$? This is an especially important problem because it will allow us to now match data in the newly inserted target camera $\mathtt{C}_3$ with all the previously installed cameras.

1.1. Overview of Solution Strategy

To address the first challenge, we propose an unsupervised re-identification method based on geodesic flow kernel [18, 19] that can effectively find the best source camera to adapt with a target camera. Given camera pairs, each consisting of 1 (out of $\mathcal{N}$) source camera and a target camera, we first compute a kernel over the subspaces representing the data of both cameras and then use it to find the kernel distance across the source and target camera. Then, we rank the source cameras based on the average distance and choose the one with lowest distance as the best source cam-

era to pair with the target camera. This is intuitive since a camera which is closest to the newly introduced camera will give the best re-id performance on the target camera and hence, is more likely to adapt better than others. In other words, a source camera with lowest distance with respect to a target camera indicates that both of the sensors could be similar to each other and their features may be similarly distributed. Note that we learn the kernel with the labeled data from the source camera only.

To address the second challenge, we introduce a transitive inference algorithm for re-identification that can exploit information from best source camera to improve accuracy across other camera pairs. With regard to the example in Fig. 1 in which source camera $\mathtt{C}_1$ best matches with target camera $\mathtt{C}_3$, our proposed transitive algorithm establishes a path between camera pair ($\mathtt{C}_2-\mathtt{C}_3$) by marginalization over the domain of possible appearances in best source camera $\mathtt{C}_1$. Specifically, $\mathtt{C}_1$ plays the role of a "connector" between $\mathtt{C}_2$ and $\mathtt{C}_3$. Experiments show that this approach indeed increases the overall re-id accuracy in a network by improving matching performance across camera pairs, while exploiting side information from best source camera.

1.2. Contributions

We address a novel, and very practical, problem in this paper — how to add one or more cameras temporarily to an existing network and exploit it for re-identification, without also adding a very expensive training phase. To the best of our knowledge, this is the first time such a problem is being addressed in re-identification research. Towards solving this problem, we make the following contributions:
(i) An unsupervised re-identification approach based on geodesic flow kernel that can find the best source camera to adapt with a newly introduced target camera in a dynamic camera network; (ii) a transitive inference algorithm to exploit the side information from the best source camera to improve the matching accuracy across other camera pairs; (iii) rigorous experiments validating the advantages of our approach over existing alternatives on multiple benchmark datasets with variable number of cameras.

2. Related Work

Person re-identification has been studied from different perspectives (see [74] for a recent survey). Here, we focus on some representative methods closely related to our work.

Supervised Re-identification. Most existing person re-identification techniques are based on supervised learning. These methods either seek the best feature representation [64, 39, 3, 49, 9, 55, 6, 5] or learn discriminant metrics/dictionaries [37, 76, 53, 36, 38, 58, 23, 26, 72] that yield an optimal matching score between two cameras or between a gallery and a probe image. Recently, deep learning methods have been applied to person re-identification [70, 68, 65, 10, 63, 43]. Combining feature representation and met-

ric learning with end-to-end deep neural networks is also a recent trend in re-identification [1, 35, 66]. Considering that a modest-sized camera network can easily have hundreds of cameras, these supervised re-id models will require huge amount of labeled data which is difficult to collect in real-world settings. In an effort to bypass tedious labeling of training data in supervised re-id models, there has been recent interest in using active learning for re-identification that intelligently selects unlabeled examples for the experts to label in an interactive manner [40, 13, 50, 60, 14]. However, all these approaches consider a static camera network unlike the problem domain we consider.

Unsupervised Re-identification. Unsupervised learning models have received little attention in re-identification because of their weak performance on benchmarking datasets compared to supervised methods. Representative methods along this direction use either hand-crafted appearance features [47, 41, 46, 11] or saliency statistics [71] for matching persons without requiring huge amount of labeled data. Recently, dictionary learning based methods have also been utilized for re-identification in an unsupervised setting [28, 44, 29]. Although being scalable in real-world settings, these approaches have not yet considered the dynamic nature of the re-id problem, where new cameras can be introduced at any time to an existing network.

Open World Re-Identification. Open world recognition has been introduced in [4] as an attempt to move beyond the dominant static setting to a dynamic and open setting where the number of training images and classes are not fixed in recognition. Inspired by such approaches, recently there have been few works in re-identification [77, 8] which try to address the open world scenario by assuming that gallery and probe sets contain different identities of persons. Unlike such approaches, we consider yet another important aspect of open world re-identification where the camera network is dynamic and the system has to incorporate a new camera with minimal additional effort.

Domain Adaptation. Domain adaptation, which aims to adapt a source domain to a target domain, has been successfully used in many areas of computer vision and machine learning. Despite its applicability in classical vision tasks, domain adaptation for re-identification still remains as a challenging and under addressed problem. Only very recently, domain adaptation for re-id has begun to be considered [34, 75, 62, 45]. However, these studies consider only improving the re-id performance in a static camera network with fixed number of cameras.

To the best of our knowledge, this is the first work to address the intrinsically dynamic nature of re-identification in unconstrained open world settings, i.e., scenarios where new camera(s) can be introduced to an existing network, and which, the system will have to incorporate for re-identification with minimal to no human supervision.

3. Methodology

To adapt re-id models in a dynamic camera network, we first formulate a domain perceptive re-identification approach based on geodesic flow kernel which can effectively find the best source camera (out of multiple installed ones) to pair with a newly introduced target camera with minimal additional effort (Section 3.2). Then, to exploit information from the best source camera, we propose a transitive inference algorithm that improves the matching performance across other camera pairs in a network (Section 3.3).

3.1. Initial Setup

Our proposed framework starts with an installed camera network where the discriminative distance metrics between each camera pairs is learned using a off-line intensive training phase. Let there be $\mathcal{N}$ cameras in a network and the number of possible camera pairs is $\binom{\mathcal{N}}{2}$. Let $\{(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_i^{\mathcal{B}})\}_{i=1}^m$ be a set of training samples, where $\mathbf{x}_i^{\mathcal{A}} \in \mathbb{R}^D$ represents feature representation of training a sample from camera view $\mathcal{A}$ and $\mathbf{x}_i^{\mathcal{B}} \in \mathbb{R}^D$ represents feature representation of the same person in a different camera view $\mathcal{B}$. We assume that the provided training data is for the task of single-shot person re-identification, i.e., there exists only two images of the same person – one image taken from camera view $\mathcal{A}$ and another image taken from camera view $\mathcal{B}$.

Given the training data, we follow KISS metric learning (KISSME) [30] and compute the pairwise distance matrices such that distance between images of the same individual is less than distance between images of different individuals. The basic idea of KISSME is to learn the Mahalanobis distance by considering a log likelihood ratio test of two Gaussian distributions. The likelihood ratio test between dissimilar pairs and similar pairs can be written as

$$\mathcal{R}(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_j^{\mathcal{B}}) = \log \frac{\frac{1}{\mathcal{C}_{\mathcal{D}}} \exp(-\frac{1}{2}\mathbf{x}_{ij}^T \Sigma_{\mathcal{D}}^{-1} \mathbf{x}_{ij})}{\frac{1}{\mathcal{C}_{\mathcal{S}}} \exp(-\frac{1}{2}\mathbf{x}_{ij}^T \Sigma_{\mathcal{S}}^{-1} \mathbf{x}_{ij})} \quad (1)$$

where $\mathbf{x}_{ij} = \mathbf{x}_i^{\mathcal{A}} - \mathbf{x}_j^{\mathcal{B}}$, $\mathcal{C}_{\mathcal{D}} = \sqrt{2\pi|\Sigma_{\mathcal{D}}|}$, $\mathcal{C}_{\mathcal{S}} = \sqrt{2\pi|\Sigma_{\mathcal{S}}|}$, $\Sigma_{\mathcal{D}}$ and $\Sigma_{\mathcal{S}}$ are covariance matrices of dissimilar and similar pairs respectively. With simple manipulations, (1) can be written as

$$\mathcal{R}(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_j^{\mathcal{B}}) = \mathbf{x}_{ij}^T \mathbf{M} \mathbf{x}_{ij} \quad (2)$$

where $\mathbf{M} = \Sigma_{\mathcal{S}}^{-1} - \Sigma_{\mathcal{D}}^{-1}$ is the Mahalanobis distance between covariances associated to a pair of cameras. We follow [30] and clip the spectrum by an eigen-analysis to ensure $\mathbf{M}$ is positive semi-definite.

Note that our approach is agnostic to the choice of metric learning algorithm used to learn the optimal metrics across camera pairs in an already installed network. We adopt KISSME in this work since it is simple to compute and has shown to perform satisfactorily on the person re-id problem. We will also show an experiment using Logistic Discriminant-based Metric Learning (LDML) [21] instead of KISSME later in Section 4.4.

3.2. Discovering the Best Source Camera

Objective. Given an existing camera network where optimal matching metrics across all possible camera pairs are computed using the above training phase, our first objective is to select the best source camera which has the lowest kernel distance with respect to the newly inserted camera. Labeling data in the new camera and then learning distance metrics with the already existing $\mathcal{N}$ cameras is practically impossible since labeling all the samples may often require tedious human labor. To overcome such an important problem, we adopt an unsupervised strategy based on geodesic flow kernel [18, 19] to learn the metrics without requiring any labeled data from the newly introduced camera.

Approach Details. Our approach for discovering the best source camera consists of the following steps: (i) compute geodesic flow kernels between the new (target) camera and other existing cameras (source); (ii) use the kernels to determine the distance between them; (iii) rank the source cameras based on distance with respect to the target camera and choose the one with the lowest as best source camera.

Let $\{\mathcal{X}^s\}_{s=1}^{\mathcal{N}}$ be the $\mathcal{N}$ source cameras and $\mathcal{X}^{\mathcal{T}}$ be the newly introduced target camera. To compute the kernels in an unsupervised way, we extend a previous method [18] that adapts classifiers in the context of object recognition to re-identification in a dynamic camera network. The main idea of our approach is to compute the low-dimensional subspaces representing data of two cameras (one source and one target) and then map them to two points on a Grassmannian[1]. Intuitively, if these two points are close on the Grassmannian, then the computed kernel would provide high matching performance on the target camera. In other words, both of the cameras could be similar to each other and their features may be similarly distributed over the corresponding subspaces. For simplicity, let us assume we are interested in computing the kernel matrix $\mathbf{K}^{\mathcal{ST}} \in \mathbb{R}^{D \times D}$ between the source camera $\mathcal{X}^{\mathcal{S}}$ and a newly introduced target camera $\mathcal{X}^{\mathcal{T}}$. Let $\tilde{\mathcal{X}}^{\mathcal{S}} \in \mathbb{R}^{D \times d}$ and $\tilde{\mathcal{X}}^{\mathcal{T}} \in \mathbb{R}^{D \times d}$ denote the d-dimensional subspaces, computed using Partial Least Squares (PLS) and Principal Component Analysis (PCA) on the source and target camera, respectively. Note that we cannot use PLS on the target camera since it is a supervised dimension reduction technique and requires label information for computing the subspaces.

Given both of the subspaces, the closed loop solution to the geodesic flow kernel between the source and target camera is defined as

$$\mathbf{x}_i^{\mathcal{S}^T} \mathbf{K}^{\mathcal{ST}} \mathbf{x}_j^{\mathcal{T}} = \int_0^1 (\psi(\mathbf{y})^T \mathbf{x}_i^{\mathcal{S}})^T (\psi(\mathbf{y}) \mathbf{x}_j^{\mathcal{T}}) \, d\mathbf{y} \quad (3)$$

where $\mathbf{x}_i^{\mathcal{S}}$ and $\mathbf{x}_j^{\mathcal{T}}$ represent feature descriptor of i-th and j-th sample in source and target camera respectively. $\psi(\mathbf{y})$

---

[1]Let d being the dimension of the subspace, the collection of all d-dimensional subspaces form the Grasssmannian.

is the geodesic flow parameterized by a continuous variable $\mathbf{y} \in [0, 1]$ and represents how to smoothly project a sample from the original D-dimensional feature space onto the corresponding low dimensional subspace. The geodesic flow $\psi(\mathbf{y})$ over two cameras can be defined as [18],

$$\psi(\mathbf{y}) = \begin{cases} \tilde{\mathcal{X}}^{\mathcal{S}} & \text{if } \mathbf{y} = 0 \\ \tilde{\mathcal{X}}^{\mathcal{T}} & \text{if } \mathbf{y} = 1 \\ \tilde{\mathcal{X}}^{\mathcal{S}} \mathcal{U}_1 \mathcal{V}_1(\mathbf{y}) - \tilde{\mathcal{X}}_o^{\mathcal{S}} \mathcal{U}_2 \mathcal{V}_2(\mathbf{y}) & \text{otherwise} \end{cases} \quad (4)$$

where $\tilde{\mathcal{X}}_o^{\mathcal{S}} \in \mathbb{R}^{D \times (D-d)}$ is the orthogonal matrix to $\tilde{\mathcal{X}}^{\mathcal{S}}$ and $\mathcal{U}_1, \mathcal{V}_1, \mathcal{U}_2, \mathcal{V}_2$ are given by the following pairs of SVDs,

$$\mathcal{X}^{\mathcal{S}^T} \mathcal{X}^{\mathcal{T}} = \mathcal{U}_1 \mathcal{V}_1 \mathcal{P}^T, \ \mathcal{X}_o^{\mathcal{S}^T} \mathcal{X}^{\mathcal{T}} = -\mathcal{U}_2 \mathcal{V}_2 \mathcal{P}^T. \quad (5)$$

With the above defined matrices, $\mathbf{K}^{\mathcal{ST}}$ can be computed as

$$\mathbf{K}^{\mathcal{ST}} = \begin{bmatrix} \tilde{\mathcal{X}}^{\mathcal{S}} \mathcal{U}_1 & \tilde{\mathcal{X}}_o^{\mathcal{S}} \mathcal{U}_2 \end{bmatrix} \mathcal{G} \begin{bmatrix} \mathcal{U}_1^T \mathcal{X}^{\mathcal{S}^T} \\ \mathcal{U}_2^T \mathcal{X}_o^{\mathcal{S}^T} \end{bmatrix} \quad (6)$$

where $\mathcal{G} = \begin{bmatrix} \text{diag}[1 + \frac{\sin(2\theta_i)}{2\theta_i}] & \text{diag}[\frac{(\cos(2\theta_i)-1)}{2\theta_i}] \\ \text{diag}[\frac{(\cos(2\theta_i)-1)}{2\theta_i}] & \text{diag}[1 - \frac{\sin(2\theta_i)}{2\theta_i}] \end{bmatrix}$ and $[\theta_i]_{i=1}^d$ represents the principal angles between source and target camera. Once we compute all pairwise geodesic flow kernels between a target camera and source cameras using (6), our next objective is to find the distance across all those pairs. A source camera which is closest to the newly introduced camera is more likely to adapt better than others. We follow [54] to compute distance between a target camera and a source camera pair. Specifically, given a kernel matrix $\mathbf{K}^{\mathcal{ST}}$, the distance between data points of a source and target camera is defined as

$$\mathbf{D}^{\mathcal{ST}}(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_j^{\mathcal{T}}) = \mathbf{x}_i^{\mathcal{S}^T} \mathbf{K}^{\mathcal{ST}} \mathbf{x}_i^{\mathcal{S}} + \mathbf{x}_j^{\mathcal{T}^T} \mathbf{K}^{\mathcal{ST}} \mathbf{x}_j^{\mathcal{T}} - 2 \mathbf{x}_i^{\mathcal{S}^T} \mathbf{K}^{\mathcal{ST}} \mathbf{x}_j^{\mathcal{T}}$$
$$(7)$$

where $\mathbf{D}^{\mathcal{ST}}$ represents the kernel distance matrix defined over a source and target camera. We compute the average of a distance matrix $\mathbf{D}^{\mathcal{ST}}$ and consider it as the distance between two camera pairs. Finally, we chose the one that has the lowest distance as the best source camera to pair with the newly introduced target camera.

Remark 1. Note that we do not use any labeled data from the target camera to either compute the geodesic flow kernels in (6) or the kernel distance matrices in (7). Hence, our approach can be applied to adapt re-id models in a large-scale camera network with minimal additional effort.

Remark 2. We assume that the newly introduced camera will be close to at least one of the installed ones since we consider them to be operating in the same time window and thus have similar environmental factors. Moreover, our approach is not limited to a single camera and can be easily extended to even more realistic scenarios where multiple cameras are introduced to an existing network at the same time. One can easily find a common best source camera based on lowest average distance to pair with all the new cameras, or

multiple best source cameras, one for each target camera, in an unsupervised way similar to the above approach. Experiments on a large-scale network of 16 cameras show the effectiveness of our approach in scenarios where multiple cameras are introduced at the same time (See Section 4.2).

3.3. Transitive Inference for Re-identification

Objective. In the previous section we have presented a domain perceptive re-identification approach that can effectively find a best source camera to pair with the target camera in a dynamic camera network. Once the best source camera is identified, the next question is: *can we exploit the best source camera information to improve the re-identification accuracy of other camera pairs?*

Approach Details. Let $\{\mathbf{M}^{ij}\}_{i,j=1,i<j}^{\mathcal{N}}$ be the optimal pair-wise metrics learned in a network of $\mathcal{N}$ cameras following Section 3.1 and $\mathcal{S}^\star$ be the best source camera for a newly introduced target camera $\mathcal{T}$ following Section 3.2.

Motivated by the effectiveness of Schur product in operations research [31], we develop a simple yet effective transitive algorithm for exploiting information from the best source camera. Schur product (a.k.a. Hadamard product) has been an important tool for improving the matrix consistency and reliability in multi-criteria decision making. Our problem naturally fits to such decision making systems since our goal is to establish a path between two cameras via the best source camera. Given the best source camera $\mathcal{S}^\star$, we compute the kernel matrix between remaining source cameras and the target camera as follows,

$$\tilde{\mathbf{K}}^{\mathcal{ST}} = \mathbf{M}^{\mathcal{SS}^\star} \odot \mathbf{K}^{\mathcal{S}^\star \mathcal{T}}, \ \forall [\mathcal{S}]_{i=1}^{\mathcal{N}}, \ \mathcal{S} \neq \mathcal{S}^\star \quad (8)$$

where $\tilde{\mathbf{K}}^{\mathcal{ST}}$ represents the updated kernel matrix between source camera $\mathcal{S}$ and target camera $\mathcal{T}$ by exploiting information from best source camera $\mathcal{S}^\star$. The operator $\odot$ denotes Schur product of two matrices. Eq. 8 establishes an indirect path between camera pair $(\mathcal{S}, \mathcal{T})$ by marginalization over the domain of possible appearances in best source camera $\mathcal{S}^\star$. In other words, camera $\mathcal{S}^\star$ plays a role of connector between the target camera $\mathcal{T}$ and all other source cameras.

Summarizing, to adapt re-id models in a dynamic network, we use the kernel matrix $\mathbf{K}^{\mathcal{S}^\star \mathcal{T}}$ computed using (6) to obtain the re-id accuracy across the newly inserted target camera and best source camera, whereas we use the updated kernel matrices, computed using (8) to find the matching accuracy across the target camera and remaining source cameras in an existing network.

Remark 3. While there are more sophisticated strategies to incorporate the side information, the reason to adopt a simple weighting approach as in (8) is that by doing so we can scale the re-identification models easily to a large scale network involving hundreds of cameras in real-time applications. Furthermore, the proposed transitive algorithm performs satisfactorily in several dynamic camera networks as illustrated in Section 4.

3.4. Extension to Semi-supervised Adaptation

Although our framework is designed for unsupervised adaptation of re-id models, it can be easily extended if labeled data from the newly introduced target camera becomes available. Specifically, the label information from target camera can be encoded while computing subspaces. That is, instead of using PCA for estimating the subspaces, we can use Partial Least Squares (PLS) to compute the discriminative subspaces on the target data by exploiting the labeled information. PLS has shown to be effective in finding discriminative subspaces by projecting data with labeled information to a common subspace [17, 57]. This essentially leads to semi-supervised adaptation of re-id models in a dynamic camera network (See experiments in Sec 4.3).

4. Experiments

Datasets. We conduct experiments on four different publicly available datasets to verify the effectiveness of our framework, namely WARD [51], RAiD [12], SAIVT-SoftBio [7] and Shinpuhkan2014 [27]. Although there are number of other datasets (*e.g.* ViPeR [20], PRID2011 [22] and CUHK [33] etc.) for evaluating the performance in re-id, these datasets do not fit our purposes since they have only two cameras or are specifically designed for video-based re-identification [61]. The number of cameras in WARD, RAiD and SAIVT-SoftBio are 3, 4, and 8 respectively. Shinpuhkan2014 is one of the largest publicly available dataset for re-id with 16 cameras. Detailed description of these datasets is available in the supplementary material.

Feature Extraction and Matching. The feature extraction stage consists of extracting Local Maximal Occurrence (LOMO) feature proposed in [39] for person representation. The descriptor has 26,960 dimensions. We follow [30, 52] and apply principal component analysis to reduce the dimensionality to 100 in all our experiments. Without low-dimensional feature, it is computationally infeasible to inverse covariance matrices of both similar and dissimilar pairs as discussed in [30, 52]. To compute distance between cameras, as well as, re-id matching score, we use kernel distance [54] (Eq. 7) for a given projection metric.

Performance Measures. We show results in terms of recognition rate as Cumulative Matching Characteristic (CMC) curves and normalized Area Under Curve (nAUC) values, as is common practice in re-id literature [26, 12, 50, 71, 28]. Due to space constraint, we only report average CMC curves for most experiments and leave the full CMC curves in the supplementary material.

Experimental Settings. We maintain following conventions during all our experiments: All the images are normalized to 128×64 for being consistent with the evaluations carried out by state-of-the-art methods [3, 12, 11]. Following the literature [12, 30, 39], the training and testing sets are kept disjoint by picking half of the available data for

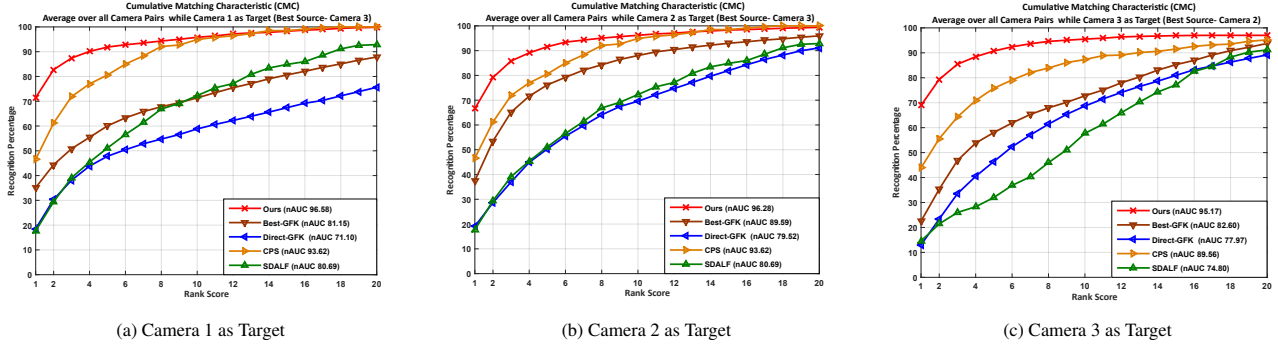(a) Camera 1 as Target   (b) Camera 2 as Target   (c) Camera 3 as Target

Figure 2. CMC curves for WARD dataset with 3 cameras. Plots (a, b, c) show the performance of different methods while introducing camera 1, 2 and 3 respectively to a dynamic network. Please see the text in Section 4.1 for the analysis of the results. Best viewed in color.
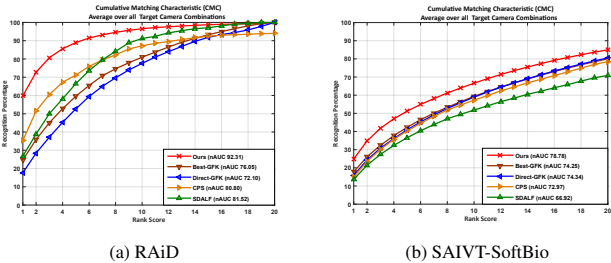

(a) RAiD   (b) SAIVT-SoftBio

Figure 3. CMC curves averaged over all target camera combinations, introduced one at a time. (a) Results on RAiD dataset with 4 cameras (b) Results on SAVIT-SoftBio dataset with 8 cameras. Please see the text in Section 4.1 for the analysis of the results.

training set and rest of the half for testing. We repeated each task 10 times by randomly picking 5 images from each identity both for train and test time. The subspace dimension for all the possible combinations are kept at 50.

4.1. Re-identification by Introducing a New Camera

Goal. The goal of this experiment is to analyze the performance of our unsupervised framework while introducing a single camera to an existing network where optimal distance metrics are learned using an intensive training phase.

Compared Methods. We compare our approach with several unsupervised alternatives which fall into two categories: (i) hand-crafted feature-based methods including CPS [11] and SDALF [3], and (ii) two domain adaptation based methods (Best-GFK and Direct-GFK) based on geodesic flow kernel [18]. For Best-GFK baseline, we compute the re-id performance of a camera pair by applying the kernel matrix, $\mathbf{K}^{\mathcal{S}^* \mathcal{T}}$ computed between best source and target camera [18], whereas in Direct-GFK baseline, we use the kernel matrix computed directly across source and target camera using (6). The purpose of comparing with Best-GFK is to show that the kernel matrix computed across the best source and target camera does not produce optimal re-id performance in computing matching performance across other source cameras and the target camera.

Implementation Details. We use publicly available codes for CPS and SDALF and tested on our experimented datasets. We also implement both Best-GFK and

Direct-GFK baselines under the same experimental settings as mentioned earlier to have a fair comparison with our proposed method. For all the datasets, we considered one camera as newly introduced target camera and all the other as source cameras. We considered all the possible combinations for conducting experiments. We first pick which source camera matches best with the target one, and then, use the proposed transitive algorithm to compute the re-id performance across remaining camera pairs.

Results. Fig. 2 shows the results for all possible 3 combinations (two source and one target) on the 3 camera WARD dataset, whereas Fig. 3 shows the average performance over all possible combinations by inserting one camera, on RAiD and SAIVT-SoftBio dataset, respectively. From all three figures, the following observations can be made: (i) the proposed framework for re-identification consistently outperforms all compared unsupervised methods on all three datasets by a significant margin. (ii) among the alternatives, CPS baseline is the most competitive. However, the gap is still significant due to the two introduced components working in concert: discovering the best source camera and exploiting its information for re-identification. The rank-1 performance improvements over CPS are 23.44%, 24.50% and 9.98% on WARD, RAiD and SAVIT-SoftBio datasets respectively. (iii) Best-GFK works better than Direct-GFK in most cases, which suggests that kernels computed across the best source camera and target camera can be applied to find the matching accuracy across other camera pairs in re-identification. (iv) Finally, the performance gap between our method and Best-GFK (maximum improvement of 17% in nAUC on RAiD) shows that the proposed transitive algorithm is effective in exploiting information from the best source camera while computing re-id accuracy across camera pairs.

4.2. Introducing Multiple Cameras

Goal. The aim of this experiment is to validate the effectiveness of our proposed approach while introducing multiple cameras at the same time in a dynamic camera network.

Implementation Details. We conduct this experiment

(a) Camera 2, 9 as Targets   (b) Camera 3, 9, 13 as Targets   (c) Camera 2, 5, 7, 8, 14 as Targets
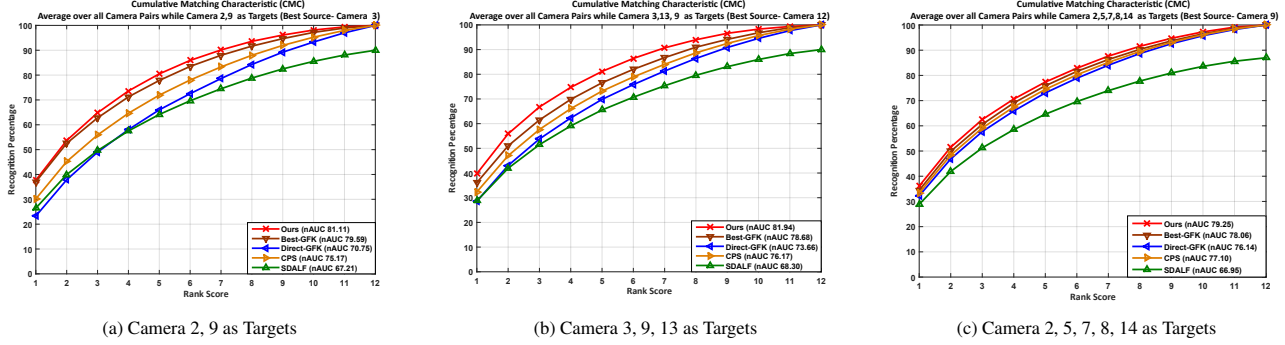
Figure 4. CMC curves for Shinpuhkan2014 dataset with 16 cameras. Plots (a, b, c) show the performance of different methods while introducing 2, 3 and 5 cameras respectively at the same time. Please see the text in Section 4.2 for the analysis of the results.

on Shinpuhkan2014 dataset [27] with 16 cameras. We randomly chose 2, 3 and 5 cameras as the target cameras while remaining cameras are possible source cameras. For each case, we pick the common best source camera based on the average distance and follow the same strategy as in Section 4.1. Results with multiple best source cameras, one for each target camera, are included in the Supplementary.

Results. Fig. 4 shows results of our method while randomly introducing 2, 3 and 5 cameras respectively on Shinpuhkan2014 dataset. From Fig. 4 (a, b, and c), the following observations can be made: (i) Similar to the results in Section 4.1, our approach outperforms all compared methods in all three scenarios. This indicates that the proposed method is very effective and can be applied to large-scale dynamic camera networks where multiple cameras can be introduced at the same time. (ii) The gap between ours and `Best-GFK` is moderate but still we improve by 4% in nAUC values, which corroborates the effectiveness of transitive inference for re-identification in a large-scale camera network.

4.3. Extension to Semi-supervised Adaptation

Goal. The proposed method can be easily extended to semi-supervised settings when labeled data from the target camera become available. The objective of this experiment is to analyze the performance of our approach in such settings by incorporating labeled data from the target domain.

Compared Methods. We compare the proposed unsupervised approach with four variants of our method where 10%, 25%, 50% and 100% of the labeled data from target camera are used for estimating kernel matrix respectively.

Implementation Details. We follow same experimental strategy in finding average re-id accuracies over a camera network. However, we use PLS instead of PCA, to compute the discriminative subspaces in target camera by considering 10%, 25%, 50% and 100% labeled data respectively.

Results. We have the following key findings from Fig. 5: (i) As expected, the semi-supervised baseline `Ours-Semi-100%`, works best since it uses all the labeled data from target domain to compute the kernel matrix for finding the best source camera. (ii) Our method re-
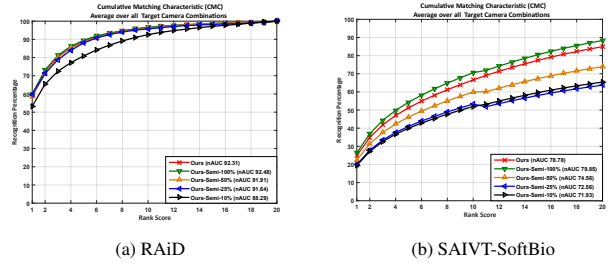

(a) RAiD   (b) SAIVT-SoftBio

Figure 5. Semi-supervised adaptation with labeled data. Plots (a,b) show CMC curves averaged over all target camera combinations, introduced one at a time, on RAiD and SAVIT-SoftBio respectively. Please see the text in Section 4.3 for analysis of the results.

mains competitive to `Ours-Semi-100%` on both datasets (Rank-1 accuracy: 60.04% vs 59.84% on RAiD and 26.41% vs 24.92% on SAVIT-SoftBio). However, it is important to note that collecting labeled samples from the target camera is very difficult in practice. (iii) Interestingly, the performance gap between our unsupervised method and other three semi-supervised baselines (`Ours-Semi-50%`, `Ours-Semi-25%`, and `Ours-Semi-10%`) are moderate on RAiD (Fig. 5-a), but on SAVIT-SoftBio, the gap is significant (Fig. 5-b). We believe this is probably due to the lack of enough labeled data in the target camera to give a reliable estimate of PLS subspaces.

4.4. Re-identification with LDML Metric Learning

Goal. The objective of this experiment is to verify the effectiveness of our approach by changing the initial setup presented in Section 3.1. Specifically, our goal is to show the performance of the proposed method by replacing KISSME [30] with LDML metric learning [21]. Ideally, we would expect similar performance improvement by our method, irrespective of the metric learning used to learn the optimal distance metrics in an existing network of cameras.

Results. Fig. 6 shows results on WARD and RAiD respectively. Following are the analysis of the figures: (i) Our approach outperforms all compared methods in both datasets which suggests that the proposed adaptation technique works significantly well irrespective of the metric

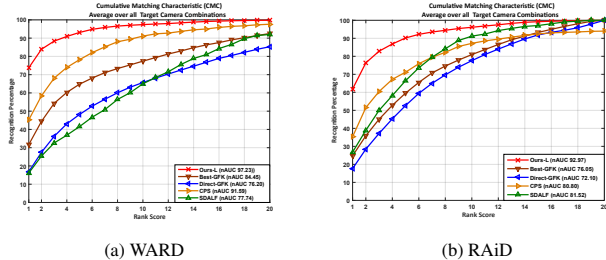| | (a) WARD | (b) RAiD |

Figure 6. Re-id performance with LDML as initial setup. Plots (a,b) show CMC curves averaged over all target camera combinations, introduced one at a time, on WARD and RAiD respectively.

learning method used in the existing camera network. (ii) The proposed approach works slightly better with LDML compared to KISSME on WARD dataset (73.77 vs 68.99 in rank-1 accuracy). However, the margin becomes smaller on RAiD (61.87 vs 59.84) which is relatively a complex re-id dataset with 2 outdoor and 2 indoor cameras. (iii) Although performance of LDML is slightly better than KISSME, it is important to note that KISSME is about 40% faster than that of LDML in learning the metrics in WARD dataset. KISSME is computationally efficient and hence more suitable for learning metrics in a large-scale camera network.

4.5. Comparison with Supervised Re-identification

Goal. The objective of this experiment is to compare the performance of our approach with supervised alternatives in a dynamic camera network.

Compared Methods. We compare with several supervised alternatives which fall into two categories: (i) feature transformation based methods including FT [49], ICT [2], WACN [51], that learn the way features get transformed between two cameras and then use it for matching, (ii) metric learning based methods including KISSME [30], LDML [21], XQDA [39] and MLAPG [38]. As mentioned earlier, our model can operate with any initial network setup and hence we show our results with both KISSME and LDML, denoted as Ours-K and Ours-L, respectively. Note that we could not compare with recent deep learning based methods as they are mostly specific to a static setting and also their pairwise camera results are not available on the experimented datasets. We did not re-implement such methods in our dynamic setting as it is very difficult to exactly emulate all the implementation details.

Implementation Details. To report existing feature transformation based methods results, we use prior published performances from [12]. For metric learning based methods, we use publicly available codes and test on our experimented datasets. Given a newly introduced camera, we use the metric learning based methods to relearn the pairwise distance metrics using the same train/test split, as mentioned earlier in experimental settings. For each datasets, we show the average performance over all possible combinations by introducing one camera at a time.

Table 1. Comparison with supervised methods. Numbers show rank-1 recognition scores in % averaged over all possible combinations of target cameras, introduced one at a time.

Methods	WARD	RAiD	Reference
FT	49.33	39.81	TPAMI2015 [49]
ICT	42.51	25.31	ECCV2012 [2]
WACN	37.53	17.71	CVPRW2012 [51]
KISSME	66.95	55.68	CVPR2012 [30]
LDML	58.66	61.52	ICCV2009 [21]
XQDA	**77.20**	**77.81**	TPAMI2015 [39]
MLAPG	72.26	77.68	ICCV2015 [38]
Ours-K	*68.99*	*59.84*	Proposed
Ours-L	*73.77*	*61.87*	Proposed

Results. Table 1 shows the rank-1 accuracy averaged over all possible target cameras introduced one at a time in a dynamic network. We have the following key findings from Table 1: (i) Both variants of our unsupervised approach (Ours-K and Ours-L) outperforms all the feature transformation based approaches on both datasets by a big margin. (ii) On WARD dataset with 3 cameras, our approach is very competitive on both settings: Ours-K outperforms KISSME and LDML whereas Ours-L overcomes MLAPG. This result suggests that our approach is more effective in matching persons across a newly introduced camera and existing source cameras by exploiting information from best source camera via a transitive inference. (iii) On the RAiD dataset with 4 cameras, the performance gap between our method and metric-learning based methods begins to increase. This is expected as with a large network involving a higher number of camera pairs, an unsupervised approach can not compete with a supervised one, especially, when the latter one is using an intensive training phase. However, we would like to point out once more that in practice collecting labeled samples from a newly inserted camera is very difficult and unrealistic in actual scenarios.

We refer the reader to the supplementary material for more detailed results (individual CMC curves) along with qualitative matching results on all datasets.

5. Conclusions

We presented an unsupervised framework to adapt re-identification models in a dynamic network, where a new camera may be temporarily inserted into an existing system to get additional information. We developed a domain perceptive re-identification method based on geodesic flow kernel to find the best source camera to pair with a newly introduced target camera, without requiring a very expensive training phase. In addition, we introduced a simple yet effective transitive inference algorithm that can exploit information from best source camera to improve the accuracy across other camera pairs. Extensive experiments on several benchmark datasets well demonstrate the efficacy of our method over state-of-the-art methods.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 3

[2] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning implicit transfer for person re-identification. In *ECCV*, 2012. 8

[3] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *CVIU*, 2013. 1, 2, 5, 6

[4] A. Bendale and T. Boult. Towards open world recognition. In *CVPR*, 2015. 3

[5] A. Bhuiyan, A. Perina, and V. Murino. Person re-identification by discriminatively selecting parts and features. In *ECCV*, 2014. 2

[6] A. Bhuiyan, A. Perina, and V. Murino. Exploiting multiple detections to learn robust brightness transfer functions in re-identification systems. In *ICIP*, 2015. 2

[7] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *DICTA*, 2012. 5

[8] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, R. Radke, Z. Wu, and F. Xiong. From the lab to the real world: Re-identification in an airport camera network. *TCSVT*, 2016. 3

[9] A. Chakraborty, A. Das, and A. K. Roy-Chowdhury. Network consistent data association. *TPAMI*, 2016. 2

[10] 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. In *CVPR*, 2016. 2

[11] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 3, 5, 6

[12] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*, 2014. 1, 5, 8

[13] A. Das, R. Panda, and A. Roy-Chowdhury. Active image pair selection for continuous person re-identification. In *ICIP*, 2015. 3

[14] A. Das, R. Panda, and A. K. Roy-Chowdhury. Continuous adaptation of multi-camera person identification models through sparse non-redundant representative selection. *CVIU*, 2017. 3

[15] H. Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009. 2

[16] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *ICCV*, 2015. 1

[17] P. Geladi and B. R. Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 1986. 5

[18] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 2, 4, 6

[19] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 2, 4

[20] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 5

[21] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009. 3, 7, 8

[22] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, 2011. 5

[23] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. 2

[24] L. Jie, T. Tommasi, and B. Caputo. Multiclass transfer learning from unconstrained priors. In *ICCV*, 2011. 2

[25] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2016. 1

[26] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015. 2, 5

[27] Y. Kawanishi, Y. Wu, M. Mukunoki, and M. Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, 2014. 5, 7

[28] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Person re-identification by unsupervised ℓ_1 graph learning. In *ECCV*, 2016. 3, 5

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

[30] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 3, 5, 7, 8

[31] G. Kou, D. Ergu, and J. Shang. Enhancing data consistency in decision matrix: Adapting hadamard model to mitigate judgment contradiction. *European Journal of Operational Research*, 2014. 5

[32] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011. 2

[33] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 5

[34] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 3

[35] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 3

[36] 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. 2

[37] 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

[38] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015. 2, 8

[39] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *TPAMI*, 2015. 1, 2, 5, 8

[40] C. Liu, C. Change Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *ICCV*, 2013. 3

[41] C. Liu, S. Gong, and C. C. Loy. On-the-fly feature importance mining for person re-identification. *PR*, 2014. 3

[42] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV*, 2012. 1

[43] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet cnn for person re-identification. In *MM*, 2016. 2

[44] 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

[45] A. J. Ma, J. Li, P. C. Yuen, and P. Li. Cross-domain person reidentification using domain adaptation ranking svms. *TIP*, 2015. 3

[46] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012. 3

[47] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012. 3

[48] Z. Ma, Y. Yang, F. Nie, N. Sebe, S. Yan, and A. G. Hauptmann. Harnessing lab knowledge for real-world action recognition. *IJCV*, 2014. 2

[49] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Re-identification in the function space of feature warps. *TPAMI*, 2015. 2, 8

[50] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *ECCV*, 2016. 3, 5

[51] N. Martinel and C. Micheloni. Re-identify people in wide area camera network. In *CVPRW*, 2012. 1, 5, 8

[52] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 1, 5

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

[54] J. M. Phillips and S. Venkatasubramanian. A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625*, 2011. 4, 5

[55] A. K. Roy-Chowdhury and B. Song. Camera networks: The acquisition and analysis of videos over wide areas. *Synthesis Lectures on Computer Vision*, 2012. 2

[56] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[57] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *ICCV*, 2009. 5

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

[59] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 2013. 1

[60] H. Wang, S. Gong, X. Zhu, and T. Xiang. Human-in-the-loop person re-identification. In *ECCV*, 2016. 3

[61] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *TPAMI*, 2016. 5

[62] X. Wang, W.-S. Zheng, X. Li, and J. Zhang. Cross-scenario transfer person re-identification. *TCSVT*, 2015. 3

[63] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016. 2

[64] Z. Wu, Y. Li, and R. J. RRadke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *TPAMI*, 2016. 1, 2

[65] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016. 2

[66] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016. 3

[67] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 1

[68] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016. 2

[69] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. G. Hauptmann. How related exemplars help complex event detection in web videos? In *ICCV*, 2013. 2

[70] D. Yi, Z. Lei, S. Liao, S. Z. Li, et al. Deep metric learning for person re-identification. In *ICPR*, 2014. 2

[71] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 3, 5

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

[73] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1

[74] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1, 2

[75] W.-S. Zheng, S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012. 3

[76] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *TPAMI*, 2013. 2

[77] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *TPAMI*, 2016. 3