Camera On-boarding for Person Re-identification using Hypothesis Transfer Learning

Sk Miraj Ahmed[1,*], Aske R Lejbølle[2,*,†], Rameswar Panda[3], Amit K. Roy-Chowdhury[1]

[1] University of California, Riverside, [2] Aalborg University, Denmark, [3] IBM Research AI, Cambridge

{sahme047@, alejboel@, rpand002@, amitrc@ece.}ucr.edu

Abstract

Most of the existing approaches for person re-identification consider a static setting where the number of cameras in the network is fixed. An interesting direction, which has received little attention, is to explore the dynamic nature of a camera network, where one tries to adapt the existing re-identification models after on-boarding new cameras, with little additional effort. There have been a few recent methods proposed in person re-identification that attempt to address this problem by assuming the labeled data in the existing network is still available while adding new cameras. This is a strong assumption since there may exist some privacy issues for which one may not have access to those data. Rather, based on the fact that it is easy to store the learned re-identifications models, which mitigates any data privacy concern, we develop an efficient model adaptation approach using hypothesis transfer learning that aims to transfer the knowledge using only source models and limited labeled data, but without using any source camera data from the existing network. Our approach minimizes the effect of negative transfer by finding an optimal weighted combination of multiple source models for transferring the knowledge. Extensive experiments on four challenging benchmark datasets with a variable number of cameras well demonstrate the efficacy of our proposed approach over state-of-the-art methods.

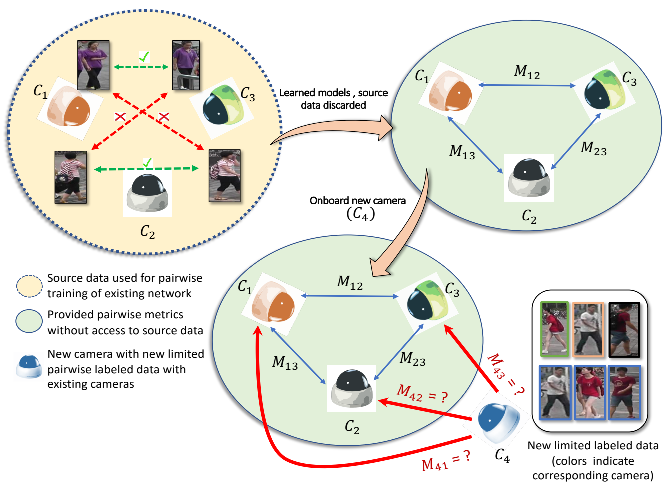


Figure 1: Consider a three camera (C_1, C_2 and C_3) network, where we have only three pairwise distance metrics (M_{12}, M_{23} and M_{13}) available for matching persons, and no access to the labeled data due to privacy concerns. A new camera, C_4, needs to be added into the system quickly, thus, allowing us to have only very limited labeled data across the new camera and the existing ones. Our goal in this paper is to learn the pairwise distance metrics (M_{41}, M_{42} and M_{43}) between the newly inserted camera(s) and the existing cameras, using the learned source metrics from the existing network and a small amount of labeled data available after installing the new camera(s).

1. Introduction

Person re-identification (re-id), which addresses the problem of matching people across different cameras, has attracted intense attention in recent years [7, 29, 51]. Much progress has been made in developing a variety of methods to learn features [16, 21, 22] or distance metrics by exploiting unlabeled and/or manually labeled data. Recently, deep learning methods have also shown significant performance improvement on person re-id [1, 15, 31, 32, 44, 52]. However, with the notable exception of [25, 26], most of these works have not yet considered the dynamic nature of a camera network, where new cameras can be introduced at any time to cover a certain related area that is not well-covered by the existing network of cameras. To build a more scalable person re-identification system, it is very essential to consider the problem of how to on-board new cameras into an existing network with little additional effort.

Let us consider K number of cameras in a network for which we have learned $\binom{K}{2}$ number of optimal pairwise

*Equal Contribution

†This work was done while AL was a visiting student at UC Riverside.

matching metrics, one for each camera pair (see Figure 1 for an illustrative example). However, during an operational phase of the system, new camera(s) may be temporarily introduced to collect additional information, which ideally should be integrated with minimal effort. Given newly introduced camera(s), the traditional re-id methods aim to re-learn the pairwise matching metrics using a costly training phase. This is impractical in many situations where the newly added camera(s) need to be operational soon after they are added. In this case, we cannot afford to wait a long time to obtain significant amount of labeled data for learning pairwise metrics, thus, we only have limited labeled data of persons that appear in the entire camera network after addition of the new camera(s).

Recently published works [25, 26] attempt to address the problem of on-boarding new cameras to a network by utilizing old data that were collected in the original camera network, combined with newly collected data in the expanded network, and source metrics to learn new pairwise metrics. They also assume the same set of people in all camera views, including the new camera (i.e., before and after on-boarding new cameras) for measuring the view similarity. However, this is unrealistic in many surveillance scenarios as source camera data may have been lost or not accessible due to privacy concerns. Additionally, new people may appear after the target camera is installed who may or may not have appeared in existing cameras. Motivated by this observation, we pose an important question: *How can we swiftly on-board new camera(s) in an existing re-id framework (i) without having access to the source camera data that the original network was trained on, and (ii) relying upon only a small amount of labeled data during the transient phase, i.e., after adding the new camera(s).*

Transfer learning, which focuses on transferring knowledge from a source to a target domain, has recently been very successful in various computer vision problems [18, 23, 30, 46, 49]. However, knowledge transfer in our system is challenging, because of limited labeled data and absence of source camera data while on-boarding new cameras. To solve these problems, we develop an efficient model adaptation approach using *hypothesis transfer learning* that aims to transfer the knowledge using only source models (i.e., learned metrics) and limited labeled data, but without using any original source camera data. *Only a few labeled identities that are seen by the target camera, and one or more of the source cameras, are needed for effective transfer of source knowledge to the newly introduced target cameras.* Henceforth, we will refer to this as *target data*. Furthermore, unlike [25, 26], which identify only one best source camera that aligns maximally with the target camera, our approach focuses on identifying an optimal weighted combination of multiple source models for transferring the knowledge.

Our approach works as follows. Given a set of pairwise source metrics and limited labeled target data after adding the new camera(s), we develop an efficient convex optimization formulation based on hypothesis transfer learning [4, 13] that minimizes the effect of negative transfer from any outlier source metric while transferring knowledge from source to the target cameras. More specifically, we learn the weights of different source metrics and the optimal matching metric jointly by alternating minimization, where the weighted source metric is used as a biased regularizer that aids in learning the optimal target metric only using limited labeled data. The proposed method, essentially, learns which camera pairs in the existing source network best describe the environment that is covered by the new camera and one of the existing cameras. Note that our proposed approach can be easily extended to multiple additional cameras being introduced at a time in the network or added sequentially one after another.

1.1. Contributions

We address the problem of swiftly on-boarding new camera(s) into an existing person re-identification network without having access to the source camera data, and relying upon only a small amount of labeled target data in the transient phase, i.e., after adding the new cameras. Towards solving the problem, we make the following contributions.

- We propose a robust and efficient multiple metric hypothesis transfer learning algorithm to efficiently adapt a newly introduced camera to an existing person re-id framework without having access to the source data.
- We theoretically analyse the properties of our algorithm and show that it minimizes the risk of negative transfer and performs closely to fully supervised case even when a small amount of labeled data is available.
- We perform rigorous experiments on multiple benchmark datasets to show the effectiveness of our proposed approach over existing alternatives.

2. Related Work

Person Re-identification. Most of the methods in person re-id are based on supervised learning. These methods apply extensive training using lots of manually labeled training data, and can be broadly classified in two categories: (i) *Distance metric learning based* [9, 12, 16, 37, 45, 47] (ii) *Deep learning based* [1, 28, 33, 40, 44, 52, 53]. *Distance metric learning based* methods tend to learn distance metrics for camera pairs using pairwise labeled data between those cameras, whereas end-to-end *Deep learning based* methods tend to learn robust feature representations of the persons, taking into consideration all the labeled data across all the cameras at once. To overcome the problem of manual labeling, several unsupervised [17, 18, 34, 43, 47, 48] and semi-supervised [5, 38, 39, 41] methods have been developed over the past decade. However, these methods do not consider the case where new cameras are added to an existing network. The most recent approach in this direction [25, 26] has considered unsupervised domain adaptation of the target camera by making a strong assumption of accessibility of the source data. None of these methods have considered the fact of not having access to the source data in the dynamic camera network setting. This is relevant, as source camera data might have been deleted after a while due to privacy concerns.

Hypothesis Transfer Learning. Hypothesis transfer learning [4, 13, 19, 24, 42] is a type of transfer learning that uses only the learned classifiers from a source domain to efficiently learn a classifier in the target domain, which contains only limited labeled data. This approach is practically appealing as it does not assume any relationship between source and target distribution, nor the availability of source data, which may be non accessible [13]. Most of the literature has dealt with simple linear classifiers for transferring knowledge [13, 35]. One recent work [27] has addressed the problem of transferring the knowledge of a source metric, which is a positive semi-definite matrix, with some provable guarantees. However, it has been analyzed for only a single source metric and the weight of the metric is calculated by minimizing a cost function using sub-gradient descent from the generalization bound separately, which is a highly non-convex non-differential function. In [35], the method has addressed transfer of multiple linear classifiers in an SVM framework, where the corresponding weights are calculated jointly with the target classifiers in a single optimization. Unlike these approaches, our approach addresses the case of transfer from multiple source metrics by jointly optimizing for target metric, as well as the source weights to reduce the risk of negative transfer.

3. Methodology

Let us consider a camera network with K cameras for which we have learned a total $N = \binom{K}{2}$ pairwise metrics using extensive labeled data. We wish to install some new camera(s) in the system that need to be operational soon after they are added, i.e., without collecting and labeling lots of new training data. We do not have access to the old source camera data, rather, we only have the pairwise source distance metrics. Moreover, we also have access to only a limited amount of labeled data across the target and different source cameras, which is collected after installing the new cameras. Using the source metrics and the limited pairwise source-target labeled data, we propose to solve a constrained convex optimization problem (Eq. 1) that aims to transfer knowledge from the source metrics to the target efficiently while minimizing the risk of negative transfer.

Formulation. Suppose we have access to the optimal distance metric $M_{ab} \in \mathbb{R}^{d \times d}$ for the a and b-th camera pair of an existing re-id network, where d is the dimension of the feature representation of the person images and $a, b \in \{1, 2 \dots K\}$. We also have limited pairwise labeled data $\{(x_{ij}, y_{ij})\}_{i=1}^{C}$ between the target camera τ and the source camera s, where $x_{ij} = (x_i - x_j)$ is the feature difference between image i in camera τ and image j in camera s, $C = \binom{n_{\tau s}}{2}$, where $n_{\tau s}$ is the total number of ordered pair images across cameras τ and s, and $y_{ij} \in \{-1, 1\}$. $y_{ij} = 1$ if the persons i and j are the same person across the cameras, and -1 otherwise. Note that our approach does not need the presence of every person seen in the new target camera across all the source cameras; rather, it is enough for some people in the target camera to be seen in at least one of the source cameras, in order to compute the new distance metric across source-target pairs. Let S and D be defined as $S = \{(i, j) \mid y_{ij} = 1\}$ and $D = \{(i, j) \mid y_{ij} = -1\}$. Our main goal is to learn the optimal metric between target and each of the source cameras by using the information from all the pairwise source metrics $\{M_j\}_{j=1}^{N}$ and limited labeled data $\{(x_{ij}, y_{ij})\}_{i=1}^{C}$. In standard metric learning context, the distance between two feature vectors $x_i \in \mathbb{R}^d$ and $x_j \in \mathbb{R}^d$ with respect to a metric $M \in \mathbb{R}^{d \times d}$ is calculated by $\sqrt{(x_i - x_j)^\top M (x_i - x_j)}$.

Thus, we formulate the following optimization problem for calculating the optimal metric $M_{\tau s}$ between target camera τ and the s-th source camera, with n_s and n_d number of similar and dissimilar pairs, as follows:

$$\begin{aligned}
\underset{M_{\tau s}, \beta}{\text{minimize}} \quad & \frac{1}{n_s} \sum_{(i,j) \in S} x_{ij}^\top M_{\tau s} x_{ij} + \lambda \| M_{\tau s} - \sum_{j=1}^{N} \beta_j M_j \|_F^2 \\
\text{subject to} \quad & \frac{1}{n_d} \sum_{(i,j) \in D} (x_{ij}^\top M_{\tau s} x_{ij}) - b \geq 0, \, M_{\tau s} \succeq 0, \\
& \beta \geq 0, \, \|\beta\|_2 \leq 1
\end{aligned} \tag{1}$$

The above objective consists of two main terms. The first term is the normalized sum of distances of all similar pair of features between camera τ and s with respect to the Mahalanobis metric $M_{\tau s}$, and the second term represents the Frobenius norm of the difference of $M_{\tau s}$ and weighted combination of source metrics squared. λ is a regularization parameter to balance the two terms. Note that the second term in Eq. 1 is essentially related to hypothesis transfer learning [4, 13] where the hypotheses are the source metrics. The first constraint represents that the normalized sum of distances of all dissimilar pairs of features with respect to $M_{\tau s}$ is greater than a user defined threshold b, and the second constraints the distance metrics to always lie in the positive semi-definite cone. While the third constraint keeps all the elements of the source weight vector non-negative, the last constraint ensures that the weights should not deviate much from zero (through upper-bounding the ℓ-2 norm by 1).

Notation. We use the following notations in the optimization steps.

(a) $\mathcal{C}_1 = \{M \in \mathbb{R}^{d \times d} \mid \frac{1}{n_d} \sum_{(i,j) \in D} (x_{ij}^\top M x_{ij}) - b \geq 0\}$

(b) $\mathcal{C}_2 = \{M \in \mathbb{R}^{d \times d} \mid M \succeq 0\}$

(c) $\mathcal{C}_3 = \{\beta \in \mathbb{R}^N \mid \beta \geq 0 \cap \|\beta\|_2 \leq 1\}$

Optimization. The proposed optimization problem (1) is not jointly convex over $M_{\tau s}$ and β. To solve this nonconvex optimization over large size matrices, we devise an iterative algorithm to efficiently solve (1) by alternatively solving for two sub-problems. For the sake of brevity, we denote $M_{\tau s}$ as M in the subsequent steps. Specifically, in the first step, we fix the weight β and take a gradient step with respect to M in the descent direction with step size α (Eq. 2). Then, we project the updated M onto $\mathcal{C}_1$ and $\mathcal{C}_2$ in an alternating fashion until convergence (Eq. 3 and Eq. 4). In the next step, we fix the the updated M and take a step with size γ towards the direction of negative gradient with respect to β (Eq. 6). In the last step, we simply project β onto the set $\mathcal{C}_3$ (Eq. 7). Algorithm 1 summarizes the alternating minimization procedure to optimize (1). We briefly describe these steps below and refer the reader to the supplementary material for more mathematical details.

Algorithm 1: Algorithm to Solve Eq. 1

Input: Source metric $\{M_j\}_{j=1}^N$, $\{(x_{ij}, y_{ij})\}_{i=1}^C$
Output: Optimal metric $M^\star$
Initialization: $M^k, \beta^k, k = 0$;
while *convergence* **do**
    $M^{k+1} = M^k - \alpha \nabla_M f(M, \beta^k)|_{M=M^k}$ (Eq. 2);
    while *convergence* **do**
        $M^{k+1} = \Pi_{\mathcal{C}_1}(M^{k+1})$ (Eq. 3);
        $M^{k+1} = \Pi_{\mathcal{C}_2}(M^{k+1})$ (Eq. 4);
    end
    $\beta^{k+1} = \beta^k - \gamma \nabla_\beta (f(M^{k+1}, \beta)|_{\beta=\beta^k}$ (Eq. 6);
    $\beta^{k+1} = \Pi_{\mathcal{C}_3}(\beta^{k+1})$ (Eq. 7);
    $k = k + 1$;
end

Step 1: Gradient w.r.t M with fixed β.
With k being the iteration number and M^k, β^k being M and β in the k-th iteration, we compute the gradient of the objective function (1) with respect to M by fixing $\beta = \beta^k$ at the k-th iteration as follows:

$$\nabla_M f(M, \beta^k)|_{M=M^k} = \Sigma_S + 2\lambda(M^k - \sum_{j=1}^{N} \beta_j^k M_j), \quad (2)$$

where $\Sigma_S = \frac{1}{n_s} \sum_{(i,j) \in S} x_{ij} x_{ij}^\top$ and $f(M, \beta^k) = \frac{1}{n_s} \sum_{(i,j) \in S} x_{ij}^\top M x_{ij} + \lambda \|M - \sum_{j=1}^{N} \beta_j^k M_j\|_F^2$.

Step 2: Projection of M onto $\mathcal{C}_1$ and $\mathcal{C}_2$. The projection of M onto $\mathcal{C}_1$ (denoted as $\Pi_{\mathcal{C}_1}(M)$) can be computed by solving a constrained optimization as follows:

$$\Pi_{\mathcal{C}_1}(M) = \arg\min_{\hat{M}} \quad \frac{1}{2}\|\hat{M} - M\|_F^2$$
$$\text{Subject to} \frac{1}{n_d} \sum_{(i,j) \in D} (x_{ij}^\top \hat{M} x_{ij}) - b \geq 0$$

By writing the Lagrange for the above constrained optimization and using KKT conditions with strong duality, the projection of M onto $\mathcal{C}_1$ can be written as

$$\Pi_{\mathcal{C}_1}(M) = M + \max\left\{0, \frac{\left(b - \frac{1}{n_d} \sum_{(i,j) \in D} x_{ij}^\top M x_{ij}\right)}{\|\Sigma_D\|_F^2}\right\} \Sigma_D, \quad (3)$$

where $\Sigma_D = \frac{1}{n_d} \sum_{(i,j) \in D} x_{ij} x_{ij}^\top$. Similarly, using spectral value decomposition, the projection of M onto $\mathcal{C}_2$ can be written as

$$\Pi_{\mathcal{C}_2}(M) = V diag([\hat{\lambda}_1 \quad \hat{\lambda}_2 \ldots \hat{\lambda}_n]) V^\top, \quad (4)$$

where V is the eigenvector matrix of M, λ_i is the i-th eigenvalue of M and $\hat{\lambda}_j = \max\{\lambda_j, 0\} \quad \forall \quad j \in [1 \ldots d]$.

Step 3: Gradient w.r.t β with fixed M. By fixing $M = M^{k+1}$ in the objective function, differentiating it w.r.t β_i, the i-th element of β at the point $\beta = \beta^k$, we get

$$\nabla_{\beta_i}(f(M^{k+1}, \beta))|_{\beta_i=\beta_i^k} = 2\lambda\beta_i^k \text{trace}(M_i^\top M_i) - 2\lambda \text{trace}(M_i^\top (M^{k+1} - \sum_{j=1, j\neq i}^{N} \beta_j^k M_j)) \quad (5)$$

By denoting $\nabla_{\beta_i}(f(M^{k+1}, \beta))|_{\beta_i=\beta_i^k}$ as a_i^k, we get

$$\nabla_\beta (f(M^{k+1}, \beta))|_{\beta=\beta^k} = \begin{bmatrix} a_1^k & a_2^k & \ldots & a_N^k \end{bmatrix}^\top \quad (6)$$

Step 4: Projection of β onto $\mathcal{C}_3$. This step essentially projects a vector to the first quadrant of an N-dimensional unit norm hyper-sphere. The closed form expression of the projection onto $\mathcal{C}_3$ is as follows:

$$\Pi_{\mathcal{C}_3}(\beta^{k+1}) = \max\left\{0, \frac{\beta^{k+1}}{\max\{1, \|\beta^{k+1}\|_2\}}\right\} \quad (7)$$

4. Discussion and Analysis

One of the key differences between our approach and existing methods is that the nature of our problem deals with the multiple metric setting within the hypothesis transfer learning framework. In this section, following [27], we theoretically analyze the properties of our Algorithm 1 for transferring knowledge from multiple metrics.

Let $\mathcal{T}$ be a domain defined over the set $(\mathcal{X} \times \mathcal{Y})$ where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \in \{-1, 1\}$ denote the feature and label set, respectively, and has a probability distribution denoted by $\mathcal{D}_\mathcal{T}$. Let T be the target domain defined by $\{(x_i, y_i)\}_{i=1}^n$ consisting of n i.i.d samples, each drawn from the distribution $\mathcal{D}_\mathcal{T}$. The optimization proposed in Eq.1 of [27] (page. 2) is defined as:

$$\underset{M \succeq 0}{\text{minimize}} \;\; L_T(M) + \lambda \|M - M_S\|_F^2 \tag{8}$$

Fixing the value of β in our proposed optimization (1), we have an optimization problem equivalent to (8), where $M_S = \sum_{j=1}^N \beta_j M_j$ and

$$L_T(M) = \frac{1}{n_s} \sum_{(i,j) \in S} x_{ij}^\top M x_{ij} + \mu^\star \big(b - \frac{1}{n_d} \sum_{(i,j) \in D} x_{ij}^\top M x_{ij}\big) \tag{9}$$

Note that $\mu^\star$ in Eq. 9 is the optimal dual variable for the inequality constraint optimization (1) with the weight vector fixed. Clearly, the expression is linear, hence convex in M, and has a finite lipschitz constant k.

Theorem 1. *For the convex and k-Lipschitz loss (shown in supp) defined in (9) the average bound can be expressed as*

$$\mathbb{E}_{T \sim \mathcal{D}_{\mathcal{T}^n}} [L_{\mathcal{D}_\mathcal{T}}(M^\star)] \leq L_{\mathcal{D}_\mathcal{T}}(\widehat{M_S}) + \frac{8k^2}{\lambda n}, \tag{10}$$

where n is the number of target labeled examples, $M^\star$ is the optimal metric computed from Algorithm 1, $\widehat{M_S}$ is the average of all source metrics defined as $\frac{\sum_{j=1}^N M_j}{N}$, $\mathbb{E}_{T \sim \mathcal{D}_{\mathcal{T}^n}} [L_{\mathcal{D}_\mathcal{T}}(M^\star)]$ is the expected loss by $M^\star$ computed over distribution $\mathcal{D}_\mathcal{T}$ and $L_{\mathcal{D}_\mathcal{T}}(\widehat{M_S})$ is the loss of average of source metrics computed over $\mathcal{D}_\mathcal{T}$.

Proof. The proof is given in supplementary material. □

Implication of Theorem 1: Since we transfer knowledge from multiple source metrics, and do not know which is the most generalizable over the target distribution (i.e., the best source metric), the most sensible thing is to check for the average performance of using each of the source metrics directly over the target test data. It is equivalently giving all the source metrics equal weights and not using any of the target data for training purpose. The bound in Theorem (9) shows that, on average, the metric learned form Algorithm 1 tends to do better than, or in worst case, at least equivalent to the average of source metrics with a fast convergence rate of $\mathcal{O}(\frac{1}{n})$ with limited number of target samples [27].

Theorem 2. *With probability $(1 - \delta)$, for any metric M learned from Algorithm 1 we have,*

$$\begin{aligned} L_{\mathcal{D}_\mathcal{T}}(M) \leq & L_T(M) + \mathcal{O}(\frac{1}{n}) + \\ & \Bigg(\sqrt{\frac{L_T(\sum_{j=1}^N \beta_j M_j)}{\lambda}} + \|\sum_{j=1}^N \beta_j M_j\|_F\Bigg) \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}, \end{aligned} \tag{11}$$

where $L_{\mathcal{D}_\mathcal{T}}(M)$ is the loss over the original target distribution (true risk), $L_T(M)$ is the loss over the existing target data (empirical risk), and n is the number of target samples.

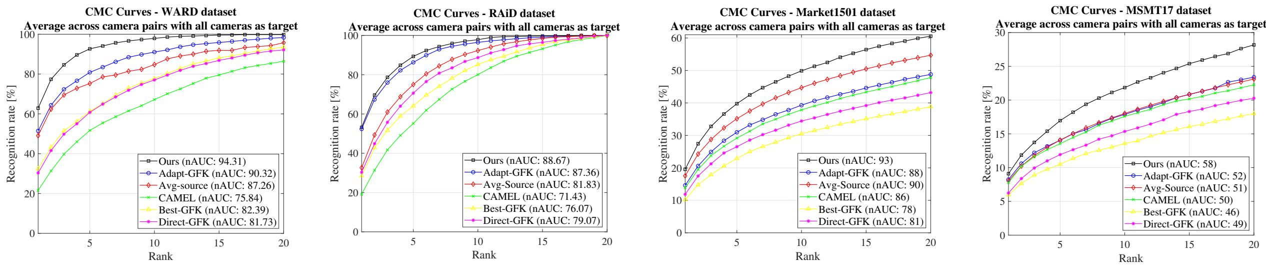
Proof. See the supplementary material for the proof. □

Implication of Theorem 2: This bound shows that given only a small amount of labeled target data, our method performs closely to the fully supervised case. The right hand side of the inequality (11) consists of the term $\mathcal{O}(\frac{1}{n}) + \Phi(\beta)\mathcal{O}(\frac{1}{\sqrt{n}})$. Since the optimal weight $\beta^\star$ from optimization (1) will be sparse due to the way β is constrained, zero weights will automatically be assigned to the outlier metrics, i.e., outlier M_js, resulting in zero values for the terms $\beta_k^\star L_T(M_j)$ corresponding to those indices j and hence smaller value of $\Phi(\beta)$. As a result, the $\mathcal{O}(\frac{1}{\sqrt{n}})$ term will be less dominant in (11) than $\mathcal{O}(\frac{1}{n})$, due to smaller associated coeffiecient $\Phi(\beta^\star)$ and, hence, can be ignored. Thus, due to the faster decay rate of $\mathcal{O}(\frac{1}{n})$, this implies that with very limited target data, the empirical risk will converge to the true risk. Furthermore, when n is very large (the fully supervised case), $\mathcal{O}(\frac{1}{\sqrt{n}})$ will be close to zero and cannot be altered by multiplication with any coefficient. This implies that the source metrics will not have any effect on learning when there is enough labeled target data available and are only useful in the presence of limited data as in our application domain.

Negative Transfer: In optimization (1), we jointly estimate the optimal metric, as well as the weight vector, which determines which source to transfer from and with how much weight. If a source metric is not a good representative of the target distribution, for an optimal λ, the weight associated to this metric will automatically be set to zero or close to zero by optimization (1), due to the sparsity constraint of β. Hence, our approach minimizes the risk of negative transfer.

5. Experiments

Datasets. We test the effectiveness of our method by experimenting on four publicly available person re-id datasets such as WARD [20], RAiD [2], Market1501 [50], and



(a) (b) (c) (d)

Figure 2: CMC curves averaged over all target camera combinations, introduced one at a time. (a) WARD with 3 cameras, (b) RAiD with 4 cameras, (c) Market1501 with 6 cameras and (d) MSMT17 with 15 cameras. Best viewed in color.

MSMT17 [36]. There are several other re-id datasets like ViPeR [8], PRID2011 [11] and CUHK01 [14]; however, those do not apply in our case due to availability of only two cameras. RAiD and WARD are smaller datasets with 43 and 70 persons captured in 4 and 3 cameras, respectively, whereas Market1501 and MSMT17 are more recent and large datasets with 1,501 and 4,101 persons captured across 6 and 15 cameras, respectively.

Feature Extraction and Matching. We use Local Maximal Occurrence (LOMO) feature [16] of length $29,960$ in RAiD and WARD datasets. However, since LOMO usually performs poorly on large datasets [7], for Market1501 and MSMT17 we extract features from the last layer of an Imagenet [3] pre-trained ResNet50 network [10] (denoted as IDE features in our work). We follow standard PCA technique to reduce the feature dimension to 100, as in [12, 25].

Performance Measures. We provide standard Cumulative Matching Curves (CMC) and normalized Area Under Curve (nAUC), as is common in person re-id [2, 12, 16, 26]. While the former shows accumulated accuracy by considering the k-most similar matches within a ranked list, the latter is a measure of re-id accuracy, independent on the number of test samples. Due to the space constraint, we only report average CMC curves for most experiments and leave the full CMC curves in the supplementary material.

Experimental Settings. For RAiD we follow the protocol in [16] and randomly split the persons into a training set of 21 persons and a test set of 20 persons, whereas for WARD, we randomly split the 70 persons into a set of 35 persons for training and rest 35 persons for testing. For both datasets, we perform 10 train/test splits and average accuracy across all splits. We use the standard training and testing splits for both Market1501 and MSMT17 datasets. During testing, we follow a multi-query approach by averaging all query features of each id in the target camera and compare with all features in the source camera [50].

Compared Methods. We compare our approach with the following methods. (1) Two variants of Geodesic Flow Kernel (GFK) [6] such as Direct-GFK where the kernel between a source-target camera pair is directly used to evaluate the accuracy and Best-GFK where GFK between the best source camera and the target camera is used to evaluate accuracy between all source-target camera pairs as in [25, 26]. Both methods use the supervised dimensionality reduction method, Partial Least Squares (PLS), to project features into a low dimensional subspace [25, 26]. (2) State-of-the-art method for on-boarding new cameras [25, 26] that uses transitive inference over the learned GFK across the best source and target camera (Adapt-GFK). (3) Clustering-based Asymmetric MEtric Learning (CAMEL) method of [47], which projects features from source and target camera to a shared space using a learned projection matrix. For all compared methods, we use their publicly available code and perform evaluation in our setting.

5.1. On-boarding a Single New Camera

We consider one camera as newly introduced target camera and all the other as source cameras. We consider all the possible combinations for conducting experiments. In addition to the baselines described above, we compare against the accuracy of average of the source metrics (Avg-Source) by applying it directly over the target test set to prove the validity of Theorem 1. We also compute the GFK kernels in two settings; by considering only target data available after introducing the new cameras (Figure 2) and by considering the presence of both old source data and the new labeled data after camera installation as in [25, 26] (Figure 3).

Implementation details. We split training data into disjoint source and target data considering the fact that the persons that appear in the new camera after installation may or may not be seen before in the source cameras. That is, for Market1501 and MSMT17, we split the training data into 90% of persons that are only seen by the source cameras and 10% that are seen in both source cameras and the new target camera after the installation. Since there are much fewer persons in RAiD and WARD training set, we split the persons into 80% source and 20% target for those two datasets. For each dataset, we evaluate every source-target pair and average accuracy across all pairs. Furthermore, we average accuracy across all cameras as target. Note that the train

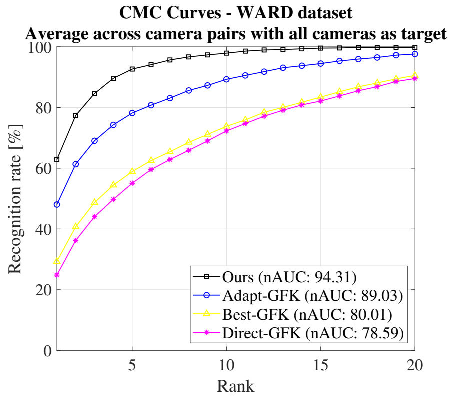


Figure 3: CMC curves averaged over all the target camera combinations, introduced one at a time, on the WARD dataset. Note that both old and new source data are used for calculation of GFK. Best viewed in color.

and test set are kept disjoint in all our experiments.

Results. Figure 2 and 3 show the results. In all cases, our method outperforms all the compared methods. The most competitive methods are those of Adapt-GFK and Avg-Source that also use source metrics. For the remaining methods, we see the limitation of only using limited target data to compute the new metrics. For Market1501, we see that Avg-Source outperforms the Adapt-GFK baseline indicating the advantage of knowledge transfer from multiple source metric compared to one single best source metric as in [25, 26]. However, our approach still outperforms the Avg-Source baseline by a margin of 20.60%, 13.81%, 2.01% and 1.07% in Rank-1 accuracy on RAiD, WARD, Market1501 and MSMT17, respectively, validating our implications of Theorem 1. Furthermore, we observe that even without accessing the source training data that was used for training the network before adding a new camera, our method outperforms the GFK based methods that use all the source data in their computations (see Figure 3). To summarize, the experimental results show that our method performs better on both small and large camera networks with limited supervision, as it is able to adapt multiple source metrics through reducing negative transfer by dynamically weighting the source metrics.

5.2. On-boarding Multiple New Cameras

We perform this experiment on Market1501 dataset using the same strategy as in Section 5.1 and compare our results with other methods while adding multiple target cameras to the network, either continuously or in parallel.

Parallel On-boarding of Cameras: We randomly select two or three cameras as target while keeping the remaining as source. All the new target cameras are tested against both source cameras and other target cameras. The results of adding two and three cameras in parallel (at the same time) are shown in Figure 4 (a) and (b), respectively. In both cases, our method outperforms all the compared methods with an increasing margin as rank increases. We outperform the most competitive CAMEL in Rank-1 accuracy by 5.45% and 3.73%, while adding two and three cameras respectively. Furthermore, our method better adapts source metrics since it has the capability of assigning zero weights to the metrics that do not generalize well over target data. Meanwhile, Adapt-GFK has a high probability of using the outlier source metrics in the presence of fewer available source metrics, which causes negative transfer. This has been shown in Figure 4 where GFK based methods are performing worse than CAMEL, which is computed just with limited supervision without using any source metrics.

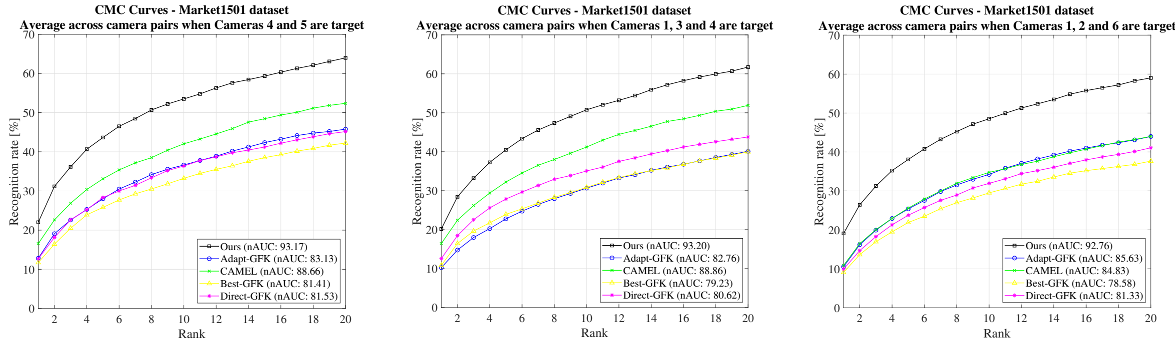
Sequential On-boarding of Cameras: For this experiment, we randomly select three target cameras that are added sequentially. A target camera is tested against all source cameras and previously added target cameras. The results are shown in Figure 4 (c). Similar to parallel on-boarding, our methods outperforms compared methods by a large margin. In this setting, we outperform CAMEL by 8.22% in Rank-1 accuracy. Additionally, compared to all GFK-based methods, the Rank-1 margin is kept constant at 10% for both parallel and sequential on-boarding. These results show the scalability of our proposed method while adding multiple cameras to a network, irrespective of whether they are added in parallel or sequentially.

5.3. Different Labeled Data in New Cameras

We perform this experiment to show the implications of Theorem 2 by using different percentages of labeled target data (10%, 20%, 30%, 50%, 75% and 100%) in our method. We compare with a widely used KISS metric learning (KISSME) [12] algorithm and show the difference in Rank-1 accuracy as a function of labeled target data. Figure 5 (a) shows the results. At only 10% labeled data, the difference between our method and KISSME [12] is almost 30%; however, as we add more labeled data, the Rank-1 accuracy becomes equivalent for the two methods at 100% labeled data. This confirms the implications of Theorem 2, where we showed that with increasing labeled target data, the effect of source metrics in learning becomes negligible.

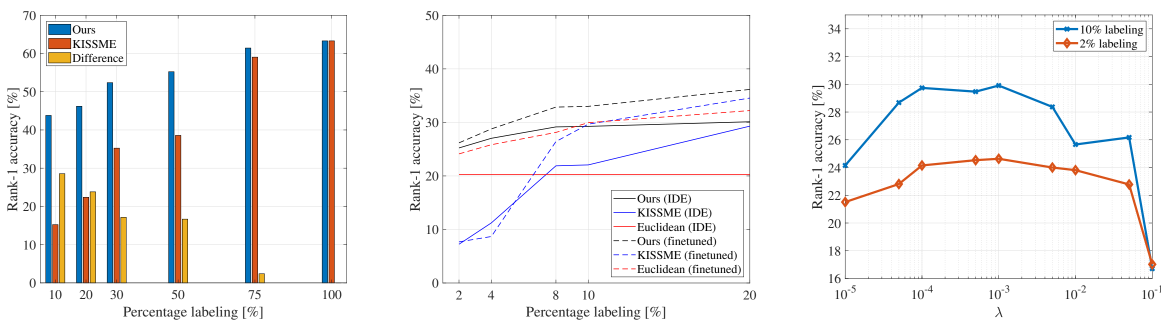
5.4. Finetuning with Deep Features

This section shows the strength of our method while comparing with CNN features extracted from a network trained on the source data (we train a ResNet50 model [10], pretrained on the Imagenet dataset). Without transfer learning, we have two options: (a) directly use the source model to extract features in the target and do matching based on Euclidean/KISSME metric (IDE), (b) finetune the source model using limited target data and then extract features to do matching using Euclidean/KISSME (finetuned). We



(a) (b) (c)

Figure 4: CMC curves averaged across target cameras on Market1501 dataset. (a) and (b) show results while adding two and three cameras in parallel, (c) show result while adding three cameras sequentially one after another. Best viewed in color.



(a) (b) (c)

Figure 5: (a) Effect of different percentage of target labelling on WARD dataset for justifying Theorem 2, (b) Analysis of our method with deep features trained on source camera data in Market1501 dataset with 6th camera as target, (c) Sensitivity of λ on the Rank-1 performance tested using deep features in Market1501 with 6th camera as target. Best viewed in color.

compared these baselines with our method with different percentage of labeling on Market1501 dataset, where the pairwise metrics are computed using the source features extracted from the model without any finetuning. We use those source metrics along with the target features, extracted before (Ours(IDE)) and after finetuning the source model (Ours(finetuned)). Please see supplementary material for more details. Figure 5 (b) shows the results. Ours(IDE) outperforms Euclidean(IDE) by a margin of 10% on Market with 20% of labeled target data. The difference between Ours(finetuned) and Euclidean/KISSME (finetuned) is more noticeable with less labeled data and it becomes smaller with increase in labeled target data (Theorem 2). However Ours(finetuned) consistently outperforms all the other baselines for up to 20% labeling.

5.5. Parameter Sensitivity

We perform this experiment to study the effect of λ in optimization (1) for a given percentage of labeled target data. Figure 5 (c) shows the Rank-1 accuracy of our proposed method for different values of λ. From optimization 1, when $\lambda \rightarrow \infty$ the left term can be neglected resulting in optimal M and β to be zero. However, when $\lambda \rightarrow 0$, the regularization term is neglected resulting in no transfer. We can see from Figure 5 (c) that there is an operating zone of λ (e.g., in the range of 10^{-4} to 10^{-2}), that is neither too high nor too low for useful transfer from source metrics.

6. Conclusions

We addressed a critically important problem in person re-identification which has received little attention thus far - how to quickly on-board new cameras into an existing camera network. We showed this can be addressed effectively using hypothesis transfer learning using only learned source metrics and a limited amount of labeled data collected after installing the new camera(s). We provided theoretical analysis to show that our approach minimizes the effect of negative transfer through finding an optimal weighted combination of multiple source metrics. We showed the effectiveness of our approach on four standard datasets, significantly outperforming several baseline methods.

Acknowledgements. This work was partially supported by ONR grant N00014-19-1-2264 and NSF grant 1724341.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015. 1, 2

[2] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*, pages 330–345. Springer, 2014. 5, 6

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6

[4] Simon S Du, Jayanth Koushik, Aarti Singh, and Barnabás Póczos. Hypothesis transfer learning via transformation functions. In *Advances in Neural Information Processing Systems*, pages 574–584, 2017. 2, 3

[5] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):83, 2018. 3

[6] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012. 6

[7] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):523–536, 2018. 1, 6

[8] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008. 6

[9] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *CVPR*, pages 498–505. IEEE, 2009. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 7

[11] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 6

[12] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012. 2, 6, 7

[13] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *ICML*, pages 942–950, 2013. 2, 3

[14] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, pages 31–44. Springer, 2012. 6

[15] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 1

[16] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, June 2015. 1, 2, 6

[17] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, volume 33, pages 8738–8745, 2019. 3

[18] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *CVPR*, pages 7948–7956, 2018. 2, 3

[19] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in neural information processing systems*, pages 1041–1048, 2009. 3

[20] Niki Martinel and Christian Micheloni. Re-identify people in wide area camera network. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 31–36. IEEE, 2012. 5

[21] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016. 1

[22] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptors with application to person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[23] Hyeonwoo Noh, Taehoon Kim, Jonghwan Mun, and Bohyung Han. Transfer learning via unsupervised task discovery for visual question answering. In *CVPR*, pages 8385–8394, 2019. 2

[24] Francesco Orabona, Claudio Castellini, Barbara Caputo, Angelo Emanuele Fiorilla, and Giulio Sandini. Model adaptation with least-squares svm for adaptive hand prosthetics. In *2009 IEEE International Conference on Robotics and Automation*, pages 2897–2903. IEEE, 2009. 3

[25] Rameswar Panda, Amran Bhuiyan, Vittorio Murino, and Amit K Roy-Chowdhury. Unsupervised adaptive re-identification in open world dynamic camera networks. In *CVPR*, pages 7054–7063, 2017. 1, 2, 3, 6, 7

[26] Rameswar Panda, Amran Bhuiyan, Vittorio Murino, and Amit K Roy-Chowdhury. Adaptation of person re-identification models for on-boarding new camera (s). *Pattern Recognition*, 96:106991, 2019. 1, 2, 3, 6, 7

[27] Michaël Perrot and Amaury Habrard. A theoretical analysis of metric hypothesis transfer learning. In *ICML*, pages 1708–1717, 2015. 3, 5

[28] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, pages 5399–5408, 2017. 2

[29] Amit K Roy-Chowdhury and Bi Song. Camera networks: The acquisition and analysis of videos over wide areas. *Synthesis Lectures on Computer Vision*, 3(1):1–133, 2012. 1

[30] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *CVPR*, pages 4360–4369, 2019. 2

[31] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *CVPR*, pages 393–402, 2019. 1

[32] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, pages 7134–7143, 2019. 1

[33] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiao-

hua Xie. Spatial-temporal person re-identification. In *AAAI*, volume 33, pages 8933–8940, 2019. 2

[34] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, pages 2275–2284, 2018. 3

[35] Yu-Xiong Wang and Martial Hebert. Learning by transferring from unsupervised universal sources. In *AAAI*, pages 2187–2193. 3

[36] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 6

[37] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009. 2

[38] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *CVPR*, pages 1187–1196, 2019. 3

[39] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, pages 5177–5186, 2018. 3

[40] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, pages 1249–1258, 2016. 2

[41] Xiaomeng Xin, Jinjun Wang, Ruji Xie, Sanping Zhou, Wenli Huang, and Nanning Zheng. Semi-supervised person re-identification using multi-view clustering. *Pattern Recognition*, 88:285–297, 2019. 3

[42] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM MM*, pages 188–197. ACM, 2007. 3

[43] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR*, pages 3633–3642, 2019. 3

[44] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, pages 1389–1398, 2019. 1, 2

[45] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017. 2

[46] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, pages 5704–5713, 2019. 2

[47] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, pages 994–1002, 2017. 2, 3, 6

[48] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, pages 2148–2157, 2019. 3

[49] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, pages 3712–3722, 2018. 2

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 5, 6

[51] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[52] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 1, 2

[53] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, pages 3741–3750, 2017. 2