# Towards Open-World Person Re-Identification by One-Shot Group-based Verification

Wei-Shi Zheng, *Member, IEEE,* Shaogang Gong, and Tao Xiang

***Abstract*— Solving the problem of matching people across non-overlapping multi-camera views, known as person re-identification (re-id), has received increasing interests in computer vision. In a real-world application scenario, a watch-list (gallery set) of a handful of known target people are provided with very few (in many cases only a single) image(s) (shots) per target. Existing re-id methods are largely unsuitable to address this open-world re-id challenge because they are designed for (1) a closed-world scenario where the gallery and probe sets are assumed to contain exactly the same people, (2) person-wise identification whereby the model attempts to verify exhaustively against each individual in the gallery set, and (3) learning a matching model using multi-shots. In this paper, a novel transfer local relative distance comparison (t-LRDC) model is formulated to address the open-world person re-identification problem by one-shot group-based verification. The model is designed to mine and transfer useful information from a labelled open-world non-target dataset. Extensive experiments demonstrate that the proposed approach outperforms both non-transfer learning and existing transfer learning based re-id methods.**

***Index Terms*— Group-based verification, open-world re-identification, transfer relative distance comparison**

## I. Introduction

Person re-identification [12], which addresses the problem of matching people across disjoint camera views in a multi-camera system, has gained increasing interests in recent years [28], [14], [40], [16], [8], [48], [50], [22], [46], [29]. Person re-identification (re-id) is useful for a number of public safety and security applications. In a typical real-world application, a watch-list of a handful of known people is provided as the gallery/target set for searching through a large volume of video surveillance footages where the people on the watch-list are likely to re-appear. This is an extremely challenging task because the video footages typically contain other people not on the watch-list. In addition, a target person may look similar to any of the other people whilst dissimilar to the target gallery image(s) due to significant changes in view angle and lighting conditions across camera views [12]. To further compound the problem, there may only be one gallery image (one-shot) available for each target person which prevents effective learning of the target's appearance variations. For example, the gallery image could be captured by an eye-witness using his/her mobile phone; or the suspect is captured by video but at a very low frame rate (typical in most of the existing recorded public space CCTV video footages) and/or in a crowded environment with many occlusions so that he/she is only clearly visible (reliably detectable) in one frame. Furthermore, solving the re-id problem becomes significantly harder in busy public spaces because the probe set would contain mostly irrelevant (non-target) people. For each target person, it is thus more likely to have someone looking similar in this large pool of irrelevant people. We call this *open-world re-identification* (Fig. 1). For such a challenging problem, relying on a fully automated system to provide accurate verification exhaustively against each target individual on the watch-list is not scalable nor tractable. However, it is reasonable to expect an automated system to provide some screening for human operators by solving an easier problem of verifying whether a probe person is on the watch-list as in a set, which is termed as *group-based person verification*, whilst leaving the more challenging task of individual identification within the set to a human operator. Since the watch-list is typically small, the latter task of human verification can be carried out quickly and more robustly.

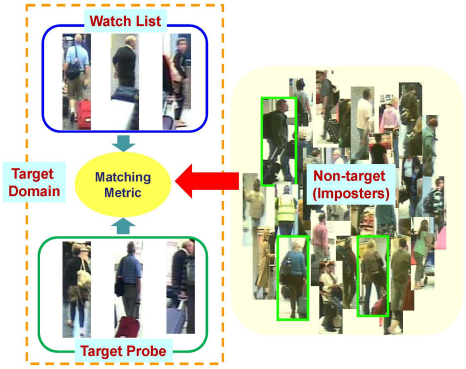



Fig. 1. The One-Shot Open-World Group-based Person Re-Identification Problem: It is assumed that only one image is available for each person on a small watch list. The orange dash line denotes the conventional closed-world person re-identification setting, where the probe set only contains the target people. Under the open-world person re-identification setting, there are a large amount of non-target imposters captured along with the target people on the watch list. Their images will also appear in the probe set and some of them will look visually similar to the target people (see those highlighted by green boxes). In general, the number of non-target imposters is unknown. In practice, it can be a known large number (as compared to the watch-list) but not a constant.

Person re-identification has quickly become an expansive field. Most existing approaches seek either the best feature representation [28], [14], [5], [40], [8], [46], [18], [24] or the best matching metrics [14], [17], [31], [50], [25], [36], [22], [29] for re-identifying people under often drastic appearance changes across camera views. However, none of them is suitable for solving the open-world one-shot group-based person verification problem because: (1) They assume a closed-world setting where the probe set comprises exactly the same people contained in the gallery set. When the probe set contains mostly non-target people (many more than those in the gallery set), the re-identification problem becomes significantly more difficult. (2) They are designed for person-wise exhaustive verification rather than for group-wise

Wei-Shi Zheng is with School of Information Science and Technology, Sun Yat-sen University, China, and is also with the Guangdong Province Key Laboratory of Computational Science, Guangzhou,China, wszheng@ieee.org

Shaogang Gong and Tao Xiang are with School of Electronic Engineering and Computer Science, Queen Mary University of London, {s.gong,t.xiang}@qmul.ac.uk

verification, i.e. a probe image is matched against every individual in the gallery set to find a winner-takes-all match. This approach may be intractable/unrobust to an open-world one-shot re-id problem due to the fact that the probe image may not belong to anyone on the watch-list resulting in a forced mismatch. (3) Most existing learning models for person re-identification cannot be readily applied for a direct verification modelling on target people, because they require multiple images (multi-shots) of each target person in the gallery in order to model appearance variations under viewpoint changes, to infer invariant features, or to learn a matching distance metric. Much of the strength of the existing methods diminishes as the number of samples per person decreases, and most of them stop working when there is only one-shot available in the gallery set for a target.

This paper presents the first attempt to solve the problem of open-world person re-identification with a sparsely sampled gallery set. Our method is based on transfer distance learning: a labelled non-target data pool (source data) is exploited and useful information is transferred to the target data in order to overcome the data sparsity problem. In our case, the target data is a small gallery set (watch-list) and the non-target data pool consists of a large quantity of non-target people labelled into matching pairs across camera views. In general, a transfer learning approach offers a natural solution to the data sparsity problem. However, conventional transfer learning techniques do not address the open-world and group-based person verification problems. To overcome this problem, we propose a novel transfer distance learning method that mines useful information from labelled non-target data to explicitly learn how to separate non-target people from a small group of target people, tackling both the open-world and group-based verification problems in a single framework.

More specifically, in line with relative distance comparison [50], we formulate a transfer relative distance comparison approach for exploring useful relative comparison information from non-target people to assist person verification on target people. We explore three types of selective relative comparisons from non-target data in order to better capture intra-class variations and refine class boundaries for the target people, as well as separating the target set as a whole from the non-target people pool. This starts with mining non-target people images visually similar to those of target people, followed by: (1) transferring intra-class variation, i.e. simulating each target person's intra-class variation by utilising the intra-class variation of the corresponding visually similar non-target people and thus the relative distance comparison can be formulated between target people for one-shot learning; (2) transferring inter-class variation, i.e. enriching the inter-class comparison by adding related relative distance comparisons from similar non-target people; and (3) enforcing group separation by introducing relative distance comparision constraints between the target people group and any non-visually similar non-target people.

Apart from exploiting new relative comparisons tailor-made for our new re-id problem, a key feature distinguishing our model from the existing relative distance comparison model [50] is that we perform local comparisons, that is, comparisons are restricted to the neighbourhood of a positive difference vector set. Therefore, we call our proposed model the transfer local relative distance comparison (t-LRDC). This is mainly motivated by the computational demand of a relative distance comparison model – in a conventional model such as RDC [50] or RankSVM [31], the number of relative comparison constraints is quadratic to the number of data points in the training set and is thus unscalable given a large non-target source dataset. With our t-LRDC model, the computational cost particularly the memory usage will be greatly reduced. Furthermore, we show both theoretically and experimentally that in our formulation, dimension reduction techniques such as PCA can be applied before model learning without sacrificing the performance of the proposed model, resulting in further reduction in computational cost. In addition, by limiting the comparisons to only the local ones, we avoid the model bias introduced by exhaustive relative comparisons, leading to better matching performance.

Extensive experiments are conducted on four benchmark datasets under an open-world experimental setting to validate the effectiveness and efficiency of the proposed method. Since this is a new setting, the existing evaluation metrics designed for the closed-world re-id setting cannot be used; new evaluation metrics are thus proposed. Our results show the proposed transfer local relative distance comparison model outperforms existing related methods for open-world one-shot re-identification.

## II. Related Work

Recent work on person re-identification mainly focuses on two aspects: finding distinctive feature representation and learning discriminant models, both of which aim to compute an optimal matching score/distance between a gallery image and a probe image. The proposed feature representations of people's appearance include color histogram [28], [14], [18], principal axis histogram [16], rectangle region histogram [5], graph representation [10], spatial co-occurrence representation [40], multiple feature based representation [14], [8], and finding saliency features [46]. Due to the large intra-class and inter-class variations of appearance [50], feature representation that is invariant to appearance changes across camera views may not exist. Consequently, there have been concerted recent efforts at learning the best matching metrics given any feature representation. The models adopted include Adaboost [14], learning to rank [31], and distance/subspace learning [17], [50], [15], [25], [36], [29], [22]. Our method is also based on distance metric learning. It is closely related to the relative distance comparison work in [50]. However, a key difference between the proposed work and the existing learning based methods is that our method is designed explicitly for solving the more realistic open-world one-shot group-based person verification problem. Specifically, given one shot per target person, previous methods, designed for multi-shot learning, can only apply a model learned from the non-target people without any model adaptation. In contrast, our model is able to select the most relevant information to transfer/adapt to the target data even with a single shot. Our experiments (see Sec. IV) demonstrate that our method outperforms the existing learning based methods for one-shot group-based open-world re-identification. Note that our notion of group-based person verification is very different from and orthogonal to that of association of group of people in [47] which uses the people walking together as contextual information.

Although there is no previous attempt on the open-world and group-based verification challenges for person re-identification, the third challenge identified in this paper, namely the extremely sparsely sample gallery/target set, has been tackled recently. Loy et al exploited unlabelled data in a manifold ranking framework to enrich the labelled target data set for label propagation [23].

However, this method does not give a solution to learning distance given one shot per target person. Similar to our method, Li et al [21] search for visually similar people from a large pool of labelled non-target data to enrich the target dataset which can be as sparse as one-shot. However, a different distance metric has to be learned for each probe image with different gallery set (identified by temporal information). This method is thus computationally very expensive and not scalable to time-critical applications. Moreover, it does not consider the effects of imposters during the re-identification and thus does not address the open-world group-based person verification problem. Recently, cross domain/dataset transfer learning in a multi-task learning framework has been employed which utilises labelled non-target data captured in different visual scenes (domains) [19], [43]. In these methods, group-based person verification is not considered, and their methods are orthogonal to ours, i.e. can be combined when labelled data from other datasets are available.

Transfer learning is a long-established topic which is studied extensively beyond person re-identification [2], [6], [20], [38], [26], [1], [45]. Here only a few most relevant general purpose transfer learning models are reviewed. The approach in [9] follows the setting in [26], that is, it assumes that labels of target data are not available but sufficient target data are used for training. Differently, our open-world person verification setting assumes that only one image of each target class is available for training (i.e. one-shot learning) and the rest unseen target data are not available for training. Due to this difference, the concept of maximum mean discrepancy (MMD) [26], which is a statistical measure relying on sufficient samples, is not applicable to our problem. This also makes some recent multi-task metric learning method, e.g. multi-task large margin metric learning method (M-LMNN) [27] impossible to implement in this case. Although there is some related one-shot learning methods discussed for object categorisation, they are either designed for specific vision model (e.g. constellation model [20], which uses prior knowledge about the hierarchical structure of categories [32]) or restricted to binary classification transfer [37]. They are thus unsuitable to our open-world person verification problem.

We incorporate local modelling in formulating the proposed t-LRDC method to make our model more scalable. There are some existing local distance/subsapce learning methods, such as LMNN [42], LDM [44] and neighbourhood component analysis (NCA) [11]. More recently, two local models, Locally-Adaptive Decision Functions (LADF) [22] and Local Fisher Discriminant Analysis (LFDA) [29], have been exploited for person re-identification. Note that despite being local data specific, LADF does not perform local data selection and focus its learning on local data. In contrast, our model does. In addition, our model computes the local neighbourhood relative to a difference vector set, instead of relative to each target data point as in LMNN, LDM, NCA and LFDA. That is because t-LRDC is a method for local feature difference modelling, rather than a local feature point modelling. Moreover, t-LRDC differs from existing local distance learning methods in that it computes the neighbourhoods adaptively by keeping them updated at each iteration during the stochastic gradient descent based optimisation steps, rather than fixing the neighbourhood using Euclidean distance as in other methods. We show in our experiments that the novel local modelling approach leads to superior re-id performance (see Sec. IV). In addition, those mentioned local distance/subspace methods cannot address the need for one-shot learning when only one gallery image is available for each target person.

The proposed t-LRDC is most closely related to our previous work RDC [50] and RankSVM [31], both of which exploit relative comparisons for model learning. However, neither RDC nor RankSVM addresses any of the following three new challenges tackled in this work: one-shot, open world and group-based verification, for person re-identification. In addition, t-LRDC differs from RDC and RankSVM in that: 1) t-LRDC performs relative comparisons locally corresponding to similar distances only so as to avert bias caused by hard/global relative comparisons; 2) Apart from having much less comparisons thus lower computational cost, t-LRDC can perform relative comparisons after feature dimensionality reduction without loss in matching performance, whilst the same cannot be said for RDC and RankSVM due to the use of absolute difference vectors, as validated in our experiments.

In summary, the main contributions of this work are: (1) For the first time, the problem of one-shot open-world group-based person verification for person re-identification is tackled. (2) A novel transfer relative distance comparison model t-LRDC is proposed which explicitly learns transferable local relative comparison information for enriching target relative comparison and separating non-target people from target people from a large non-target data pool. (3) A novel optimisation algorithm is formulated to learn the t-LRDC model which is efficient and scalable. (4) We propose novel evaluation metrics for measuring the verification performance under an open-world setting and carry out extensive experiments to compare the proposed methods with state-of-the-art alternatives. We first introduced the open-world group-based verification in a related early and preliminary version of the work published in [49]. Apart from having major differences in the model formulation, particularly the introduction of local modelling, this work differs significantly from [49] in formulating a verification model for the one-shot scenario.

## III. Transfer Learning for Group-based Person Verification

### A. Problem Statement

We consider the group-based person verification problem, that is, given a target data set consisting of as few as one shot per target person, we aim to verify whether a probe image matches anyone on the list. We take a transfer learning approach to learn a shared verification model by exploiting non-target data captured in the same environment in order to achieve a more robust group-based verification under an open-world setting.

Formally, suppose $N_T$ limited target training data are available from $m_t$ different target people $C^t_1, \cdots, C^t_{m_t}$ denoted by $\{\mathbf{x}_i, y_i\}_{i=1}^{N_T}$, where the person ID is $y_i \in \{C^t_1, \cdots, C^t_{m_t}\}$ and $\mathbf{x}_i$ denotes the $i^{th}$ target sample represented in a feature space. In this work, we assume that only one image is available for each target person, i.e. $N_T = m_t$. In addition, we are also given a larger source training data set comprising $m_s$ non-target people denoted by $\{\mathbf{x}_i, y_i\}_{i=N_T+1}^{N}$, where $y_i \in \{C^s_1, \cdots, C^s_{m_s}\}$ for $i = N_T+1, \cdots, N$ and $N - N_T >> N_T$. The problem is how to learn a more robust matching model by using these non-target data for group-based person re-identification on the small set of target people against any non-target people (open world).

We take a relative distance learning approach and aim to learn a distance function $d(\mathbf{x}, \mathbf{x}')$ between two data points $\mathbf{x}$ and $\mathbf{x}'$,

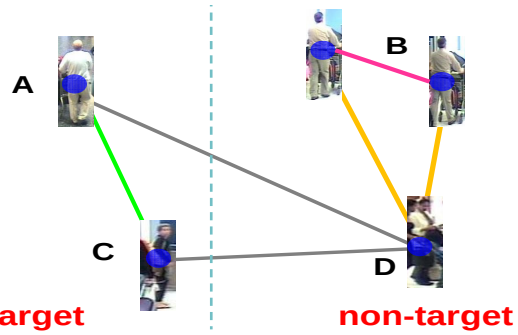



Fig. 2. An illustration of the three types of knowledge transfer. There are four different variations among target and non-target data: 1) the target inter-class variations (green lines); 2) the *selected* inter-class variation between target and non-similar non-target images (grey lines); 3) the *selected* non-target intra-class variations (magenta lines); 4) the *selected* non-target inter-class variations (yellow lines). A magenta line and a green line denote an approximate target intra-inter class pair and is used in the knowledge transfer to enrich intra-class variation (Eq. (7)); a magenta line and a yellow line denote a target specific non-target intra-inter class pair to transfer knowledge to enrich inter-class variation (Eq. (10)); a green line and a grey line denote a group separation intra-inter class pair to transfer knowledge to enrich group separation (Eq. (13)).

which is modelled as

$$d(\mathbf{x},\mathbf{x}') = (\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}') \tag{1}$$

for some semi-positive matrix $\mathbf{M}$, which is always a low-rank matrix and implies computing the Euclidean distance between data after they are projected into a lower-rank subspace. Such a distance metric function is learned subject to a number of relative distance comparison constraints. In particular, we consider the relative distance comparison between two types of distances $d(\mathbf{x},\mathbf{x}')$ and $d(\mathbf{x},\mathbf{x}'')$, where $\mathbf{x}$ and $\mathbf{x}'$ are from the same class (person) and $\mathbf{x}$ and $\mathbf{x}''$ are from different classes. This comparison is to enforce that the inter-class distance is greater than that of inter-class distance in the learned feature space defined by $M$. However, under an one-shot setting, no such comparisons can be formed from the target data – there is no intra-class distance. We therefore aim to exploit the source non-target data to form these distance comparison constraints, that is, to transfer knowledge from the source data in the form of selected relative distance comparisons.

### B. Framework Formulation

Our main idea is that since different people may have similar appearance, including similar dressing, body shape and objects associated with, it is possible to use the target people images to select those similar non-target people images to form relative distance comparisons. Specifically, we assume that if a non-target person image $\mathbf{x}_s$ is similar to one of the target images $\mathbf{x}_t$ in a feature space, the appearance variation of $\mathbf{x}_s$ should also be similar, i.e. $P(\Delta_s|\mathbf{x}_s,\Omega) \approx P(\Delta_t|\mathbf{x}_t,\Omega)$, where $\Delta_s$ and $\Delta_t$ indicate the intra-class appearance variation with respect to $\mathbf{x}_s$ and $\mathbf{x}_t$, respectively, and $\Omega$ is a feature space to be learned by our model. Based on this assumption, three types of knowledge transfer are performed with corresponding relative distance comparisons formed.

Before describing them, let us first look at how similar looking non-target people can be selected. We measure the appearance similarity between a target person image and a non-target person image by using the cosine similarity below:

$$s(\mathbf{x}_s,\mathbf{x}_t) = \frac{|\mathbf{x}_s^T\mathbf{x}_t|}{||\mathbf{x}_s||||\mathbf{x}_t||}. \tag{2}$$

We say the two appearances are visually similar when $s(\mathbf{x}_s,\mathbf{x}_t)$ is larger than a given threshold $h$, indicated by the following function $g$

$$g(\mathbf{x}_s,\mathbf{x}_t) = \begin{cases} 1, & s(\mathbf{x}_s,\mathbf{x}_t) \geq h; \\ 0, & s(\mathbf{x}_s,\mathbf{x}_t) < h. \end{cases} \tag{3}$$

The two appearances are thus not visually similar if $g(\mathbf{x}_s,\mathbf{x}_t)=0$. Note that we measure the visual similarity based on cosine similarity because this metric is well suited to texture and colour features, typical for representing appearance for person re-id, without any need for further feature transform. In addition, the cosine similarity naturally has a value ranging between 0 and 1, which is suitable for deriving a semantic threshold. The threshold $h$ is a parameter that controls how much information can be transferred from the source data to the target data. Based on this selection criterion, the following three types of knowledge transfer are carried out.

**Knowledge transfer to enrich intra-class variation –** In this transfer, given a target person, similar non-target people are selected to substitute for the missing intra-class variation for the target person/class. An example is shown in Fig. 2 where the two images linked by the magenta line are from a non-target person B whose appearance is similar to that of the target person A. These two images are then used to represent the intra-class variation of A and a distance comparison is formed to require that the distance between these two intra-class images should be shorter than the related target inter-class distance, i.e. the magenta line should be shorter than the green line with the learned distance metric/function.

Formally, suppose $\mathbf{x}_t$ is the only available image for a target person labelled with $y_t$, $t \leq N_T$. Then, we find a set of similar non-target person images that are similar to $\mathbf{x}_t$, denoted by

$$\mathcal{L}(\mathbf{x}_t) = \{\mathbf{x}_{t_j}|g(\mathbf{x}_t,\mathbf{x}_{t_j})=1,\ N_T+1 \leq t_j \leq N\}. \tag{4}$$

Next we can generate a set of pairwise samples to build the intra-class variation from non-target data by

$$\mathcal{P}(\mathbf{x}_t) = \{(\mathbf{x}_{t_j},\mathbf{x}_s)|\mathbf{x}_{t_j} \in \mathcal{L}(\mathbf{x}_t),\ y_s = y_{t_j},\ N_T+1 \leq t_j, s \leq N\},$$

where $\mathbf{x}_{t_j}$ and $\mathbf{x}_s$ are from the same non-target class in $\mathcal{P}(\mathbf{x}_t)$. Subsequently, we use the distance $d(\mathbf{x}_{t_j},\mathbf{x}_s)$ to simulate the target intra-class distance of the target image $\mathbf{x}_t$, and introduce the constraint that the target inter-class distance $d(\mathbf{x}_t,\mathbf{x}_{t'})$ should be greater than the intra-class distance $d(\mathbf{x}_{t_j},\mathbf{x}_s)$. We call these intra-inter class pairs the "approximate target intra-inter class pairs". Hence, relative comparison modelling based on this transfer is to learn the distance function that minimises the following cost

$$\min_d \sum_{t=1}^{N_T}\sum_{t'=1}^{N_T}\sum_{(\mathbf{x}_{t_j},\mathbf{x}_s)\in\mathcal{P}(\mathbf{x}_t)} g(\mathbf{x}_t,\mathbf{x}_{t_j})\ell(d(\mathbf{x}_{t_j},\mathbf{x}_s) < d(\mathbf{x}_t,\mathbf{x}_{t'})) \tag{5}$$

where $\ell(\cdot)$ is a loss function that penalises violations of the expected distance order (inter-class distance should be greater than intra-class distance). To simplify the notations in the above equation, we introduce $\mathbb{O}_g(\mathbf{x}_t)$ as

$$\begin{aligned}\mathbb{O}_g(\mathbf{x}_t) = \{&(\mathbf{x}_{t_j},\mathbf{x}_s,\mathbf{x}_{t'})| \\ &g(\mathbf{x}_t,\mathbf{x}_{t_j})=1,\ y_{t_j}=y_s, y_t \neq y_{t'}, \\ &1 \leq t' \leq N_T,\ N_T+1 \leq t_j, s \leq N\}\end{aligned} \tag{6}$$

Then, Eq. (5) can be written as

$$\min_d \sum_{t=1}^{N_T}\sum_{(\mathbf{x}_{t_j},\mathbf{x}_s,\mathbf{x}_{t'})\in\mathbb{O}_g(\mathbf{x}_t)} \ell(d(\mathbf{x}_{t_j},\mathbf{x}_s) < d(\mathbf{x}_t,\mathbf{x}_{t'})) \tag{7}$$

**Knowledge transfer to enrich inter-class variation –** In this transfer, given a target person image, similar non-target people images are again selected as before. But this time the objective is to use these selected non-target people to enrich the inter-class variations. An example of such enrichment can be seen in Fig. 2, where for target person A, we select a pair of images of non-target person B to form an intra-class distance. Instead of requiring that the magenta line being shorter than the green line, this time we require that the magenta line is shorter than the two yellow lines. Note that since there are many more yellow lines than the green lines due to the much larger size of the source non-target data set, this relative comparison greatly enriches the inter-class variations. It is also worth mentioning that although this type of comparisons seemingly only involve the non-target data, this transfer is not blindly done without adaptation towards the target data. This is because the pair of images for person B are *selected* by measuring their similarity to the target person A.

Formally, this new set of relative distance comparisons are between the intra-class distance $d(\mathbf{x}_s, \mathbf{x}_{s'})$ and related inter-class distance $d(\mathbf{x}_s, \mathbf{x}_{s''})$:

$$\min_d \sum_{t=1}^{N_T} \sum_{s=N_T+1}^{N} \sum_{s'=N_T+1, y_{s'}=y_s}^{N} \sum_{s''=N_T+1, s''\neq s', y_{s''}\neq y_s}^{N} g(\mathbf{x}_t, \mathbf{x}_s)\ell(d(\mathbf{x}_s, \mathbf{x}_{s'}) < d(\mathbf{x}_s, \mathbf{x}_{s''})) \tag{8}$$

We call the above intra-inter class pairs for comparison the "target specific non-target intra-inter class pairs". Let $\mathbb{O}_a(\mathbf{x}_t)$ denote the following:

$$\begin{aligned}\mathbb{O}_a(\mathbf{x}_t) =\{(\mathbf{x}_s, \mathbf{x}_{s'}, \mathbf{x}_{s''}) | g(\mathbf{x}_t, \mathbf{x}_s) = 1,\ y_{s'} = y_s,\\ y_{s''} \neq y_s,\ N_T + 1 \leq s, s', s'' \leq N\}\end{aligned} \tag{9}$$

Then Eq. (8) can be written as

$$\min_d \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_s, \mathbf{x}_{s'}, \mathbf{x}_{s''}) \in \mathbb{O}_a(\mathbf{x}_t)} \ell(d(\mathbf{x}_s, \mathbf{x}_{s'}) < d(\mathbf{x}_s, \mathbf{x}_{s''})) \tag{10}$$

**Knowledge transfer to enforce group separation –** The first two types of transfer are designed to address the data sparsity problem (one-shot for intra and small watch list for inter). In the third type, we tackle the open-world group-based verification. Specifically, as stated before, we aim to learn a distance function that can match a probe image against the whole target/gallery data set. It is thus intuitive that such a distance will make all the images from the target set close to each other whilst pushing the non-target people away. However, since the similarly looking non-target people images have been used in the previous two types of transfer to enrich the target data intra and inter-class variation, the pairs of target person image and any corresponding visually similar non-target person image should not be included for modelling. Otherwise, the third type of knowledge transfer could contradict the first two types[1]. Therefore, for each target person image, we use only those non-target people images that are not visually similar to it for modelling the constraint. In the illustrative example in Fig. 2, a set of relative distance comparisons are formed based on this knowledge transfer which corresponds to constraining the green line to be shorter than the grey lines in the learned subspace defined by $\mathbf{M}$.

Formally, this is to minimise the following cost function:

$$\min_d \sum_{t=1}^{N_T} \sum_{t'=1, y_{t'}\neq y_t}^{N_T} \sum_{s=N_T+1}^{N} (1 - g(\mathbf{x}_t, \mathbf{x}_s))\ell(d(\mathbf{x}_t, \mathbf{x}_{t'}) < d(\mathbf{x}_t, \mathbf{x}_s)) \tag{11}$$

Let $\mathbb{O}_b(\mathbf{x}_t)$ denote the following:

$$\begin{aligned}\mathbb{O}_b(\mathbf{x}_t) =\{(\mathbf{x}_{t'}, \mathbf{x}_s) | g(\mathbf{x}_t, \mathbf{x}_s) = 0,\ y_t \neq y_{t'},\ y_s \neq y_t,\\ 1 \leq t' \leq N_T,\ N_T + 1 \leq s \leq N\}\end{aligned} \tag{12}$$

We call the above intra-inter class pairs for comparison the "group separation intra-inter class pairs". Then we can express Eq. (11) in a more simplified way by

$$\min_d \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_{t'}, \mathbf{x}_s) \in \mathbb{O}_b(\mathbf{x}_t)} \ell(d(\mathbf{x}_t, \mathbf{x}_{t'}) < d(\mathbf{x}_t, \mathbf{x}_s)) \tag{13}$$

It is important to point out that the three types of knowledge transfer are formulated based on intuitive yet reasoned principles and each type plays a critical role in addressing the challenges posed by one-shot open-world group verification: (1) In the first type of transfer, since there is only one image available for each target person, it is not possible to model the intra-class variation. Therefore the only way is to exploit the visually similar non-target person's images to enrich the intra-class variation for the target people. (2) In the second type of transfer, since the watch-list is small, the inter-class variation is also limited. Again, exploiting the visually similar non-target person's images in the source data is both plausible and attractive to enrich the inter-class variation; (3) In the third type of transfer, since we aim to perform group-based verification, i.e. making sure people on the watch list separable from those who are not, the separation between target and non-target data is enforced.

Finally, by integrating the above three constraints/cost functions (Eqs. (7, 10, 13)), our transfer relative distance comparison model learns the distance function parameterised by $\mathbf{M}$ by:

$$\begin{aligned}&\min_{\mathbf{M}\succeq 0} f(\mathbf{M})\\ &f(\mathbf{M}) = \frac{1-\alpha}{\#\mathbb{O}_g} \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_{t_j}, \mathbf{x}_s, \mathbf{x}_{t'}) \in \mathbb{O}_g(\mathbf{x}_t)} \ell(d(\mathbf{x}_{t_j}, \mathbf{x}_s) < d(\mathbf{x}_t, \mathbf{x}_{t'}))\\ &+ \frac{\alpha}{\#\mathbb{O}_a + \#\mathbb{O}_b}\Big( \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_s, \mathbf{x}_{s'}, \mathbf{x}_{s''}) \in \mathbb{O}_a(\mathbf{x}_t)} \ell(d(\mathbf{x}_s, \mathbf{x}_{s'}) < d(\mathbf{x}_s, \mathbf{x}_{s''}))\\ &+ \beta \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_{t'}, \mathbf{x}_s) \in \mathbb{O}_b(\mathbf{x}_t)} \ell(d(\mathbf{x}_t, \mathbf{x}_{t'}) < d(\mathbf{x}_t, \mathbf{x}_s))\Big).\end{aligned} \tag{14}$$

where $\#\mathbb{O}_g = \sum_{t=1}^{N_T} \#\mathbb{O}_g(\mathbf{x}_t)$, $\#\mathbb{O}_a = \sum_{t=1}^{N_T} \#\mathbb{O}_a(\mathbf{x}_t)$, $\#\mathbb{O}_b = \sum_{t=1}^{N_T} \#\mathbb{O}_b(\mathbf{x}_t)$, $\alpha \in [0, 1]$ and $\beta \geq 0$. Depending on the choice of the loss function $\ell(\cdot)$, this optimisation problem can be solved differently. In this work, we measure the relative comparison loss in Eq. (14) using the hinge-loss function as follows:

$$\ell_h(d(\mathbf{x}_i, \mathbf{x}_k) < d(\mathbf{x}_i, \mathbf{x}_j)) = \max\left\{0, d(\mathbf{x}_i, \mathbf{x}_k) + \rho - d(\mathbf{x}_i, \mathbf{x}_j)\right\}^2. \tag{15}$$

where we set $\rho = 1$ in this work.

### C. Local Modelling

Solving the optimisation problem in Eq. (14) is expensive due to the sheer number of comparisons/constraints from the three types of knowledge transfer (quadratic to the number of training

[1]For a more in-depth discussion on this, please refer to the supplementary material.

images). Our approach to making this problem more tractable is to perform local comparison instead of comparing all related distance pairs exhaustively. More specifically, comparisons are only formed when the two difference vectors are in a local neighbourhood. This concept of local distance comparison is illustrated in Fig. 3. In this example, A and B are of the same person and the other four images are from four other people. With a slight abuse of notation, we denote $d(A,B)$ as the distance to be learned and $\vec{A,B}$ the difference vector. Exhaustive/global comparison requires that $d(A,B)$ is smaller than $d(A,C)$ , $d(A,D)$, $d(A,E)$, and $d(A,F)$ respectively. Now we introduce a different vector neighbourhood for $\vec{A,B}$, within which we will have $\vec{A,C}$ and $\vec{A,D}$. With our local relative distance comparison we only require that $d(A,B)$ is smaller than $d(A,C)$ and $d(A,D)$. This not only reduces the number of comparisons by half, but also leads to a more relaxed constraint alleviating the risk of overfitting. In particular, among the two removed constraints/pairs, $d(A,B) > d(A,E)$ is hard to meet. However, as shown in Eq. (27) those pairs if violated have a relatively small effect on learning the model due to the small magnitude of $\vec{A,E}$. In contrast, although $d(A,B) < d(A,F)$ is easier to satisfy, when violated, those pairs would have a larger effect on model learning due to the large magnitude of $\vec{A,F}$. Our experiments in Sec. IV-D validate this analysis that by introducing the local modelling, we not only gain computational efficiency but also achieve overall better re-id performance[2].

Formally, since relative distance comparison is concerned with the comparison between feature difference vectors, we consider a local relative distance comparison modelling between two nearby difference vectors $\mathbf{x}_j - \mathbf{x}_i$ and $\mathbf{x}_m - \mathbf{x}_i$ as follows

$$\begin{aligned} &d(\mathbf{x}_i,\mathbf{x}_j) < d(\mathbf{x}_i,\mathbf{x}_m) - \rho,\ \rho > 0 \\ &\text{when } (\mathbf{x}_m - \mathbf{x}_i) \in \mathcal{N}_k^{\mathbf{p}}(\mathbf{x}_i,\mathcal{D}),\ (\mathbf{x}_j - \mathbf{x}_i) \in \mathcal{D}, \end{aligned} \tag{16}$$

where $\mathbf{p} \in \{T,S\}$ indicates whether non-target data are used to form the neighbourhood set $\mathcal{N}_k^{\mathbf{p}}(\mathbf{x}_i,\mathcal{D})$. Eq. (16) means that for the expected smaller distance $d(\mathbf{x}_i,\mathbf{x}_j)$, we find the neighbouring difference vector $(\mathbf{x}_m - \mathbf{x}_i)$ by searching the $k$ nearest difference vectors to the set $\mathcal{D}$ that contains $(\mathbf{x}_j - \mathbf{x}_i)$. Those $k$ nearest difference vectors constitute the set $\mathcal{N}_k^{\mathbf{p}}(\mathbf{x}_i,\mathcal{D})$. As shown in the next paragraph, $\mathcal{D}$ can be a set of intra-class difference vectors for a certain class. Here we define the distance between a vector and a set as the minimum distance between that vector and each vector in that set in a low-rank subspace induced by the metric $d$.

Next, we shall explain what the set $\mathcal{D}$ and $\mathcal{N}_k^{\mathbf{p}}(\mathbf{x}_i,\mathcal{D})$ are. First, we have two settings of $\mathcal{D}$ below:

1) $\mathcal{D}_{y_i}^{+}(\mathbf{x}_i)$ denotes all the intra-class difference vectors related to $\mathbf{x}_i$ within class $y_i$, i.e. $\mathcal{D}_{y_i}^{+}(\mathbf{x}_i) = \{(\mathbf{x}_q - \mathbf{x}_i) \mid y_q = y_i\}$;
2) $\mathcal{D}_{y_i}^{-}(\mathbf{x}_i)$ denotes all the inter-class difference vectors between $\mathbf{x}_i$ and any other image out of class $y_i$ but still from one of the target classes, i.e. $\mathcal{D}_{y_i}^{-}(\mathbf{x}_i) = \{(\mathbf{x}_q - \mathbf{x}_i) \mid y_q \neq y_i \ \&\ 1 \leq q \leq N_T\}$;

Herein the notation $\mathbf{p}$ in $\mathcal{N}_k^{\mathbf{p}}(\mathbf{x}_i,\mathcal{D})$ is explained as follows:

1) When $\mathbf{p}$ is denoted as "$T$", $\mathcal{N}_k^{T}(\mathbf{x}_i,\mathcal{D})$ consists of the $k$ nearest difference vectors to the set $\mathcal{D}$, where these $k$ nearest difference vectors $(\mathbf{x}_m - \mathbf{x}_i)$ are found in the subset $\mathbf{Q}^T(\mathbf{x}_i)$ formed by all the difference vectors between any different-class target image and $\mathbf{x}_i$, i.e. $\mathbf{Q}^T(\mathbf{x}_i) = \{(\mathbf{x}_m - \mathbf{x}_i) \mid y_m \neq y_i \ \&\ 1 \leq m \leq N_T\}$.
2) When $\mathbf{p}$ is denoted as "$S$", $\mathcal{N}_k^{S}(\mathbf{x}_i,\mathcal{D})$ consists of $k$ nearest difference vectors $(\mathbf{x}_m - \mathbf{x}_i)$ found in the subset $\mathbf{Q}^S(\mathbf{x}_i)$ formed by all the difference vectors between any non-target image and $\mathbf{x}_i$, i.e. $\mathbf{Q}^S(\mathbf{x}_i) = \{(\mathbf{x}_m - \mathbf{x}_i) \mid y_m \neq y_i \ \&\ N_T + 1 \leq m \leq N\}$.

[2]More detailed analysis and evaluations on the effect from excluding the two types of constraints can be found in the supplementary material.

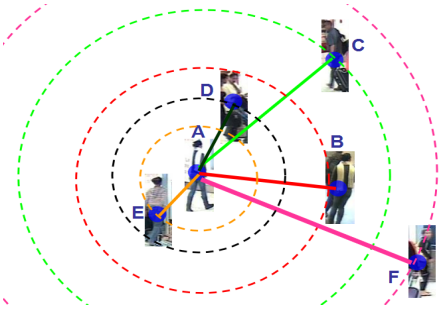



Fig. 3. Illustration of our local relative comparison. Among the six images, A and B belong to the same person whilst the other four are of four other people. See text for more details.

Based on the relative comparison defined in Eq. (16), we can now develop a relaxed transfer relative distance comparison modelling by constraining all the relative distance comparisons around the neighbourhood of a difference dataset (either $\mathcal{D}_{y_i}^{+}(\mathbf{x}_i)$ or $\mathcal{D}_{y_i}^{-}(\mathbf{x}_i)$). Specifically, we introduce the following three sets of local relative comparisons, which are the local versions of sets $\mathbb{O}_g(\mathbf{x}_t)$, $\mathbb{O}_a(\mathbf{x}_t)$ and $\mathbb{O}_b(\mathbf{x}_t)$ by following the idea in Eq. (16) to limit the relative comparisons to local ones, respectively:

$$\begin{aligned} \mathbb{O}_g^{\ell}(\mathbf{x}_t) = \{&(\mathbf{x}_{t_j},\mathbf{x}_s,\mathbf{x}_{t'}) \\ &|(\mathbf{x}_{t'} - \mathbf{x}_t) \in \mathcal{N}_k^{T}(\mathbf{x}_t,\mathcal{D}_{y_{t_j}}^{+}(\mathbf{x}_{t_j})) \ \&\ (\mathbf{x}_s - \mathbf{x}_{t_j}) \in \mathcal{D}_{y_{t_j}}^{+}(\mathbf{x}_{t_j}), \\ &g(\mathbf{x}_t,\mathbf{x}_{t_j}) = 1,\ y_{t_j} = y_s, y_t \neq y_{t'},\ 1 \leq t' \leq N_T, \\ &N_T + 1 \leq t_j, s \leq N\} \end{aligned} \tag{17}$$

$$\begin{aligned} \mathbb{O}_a^{\ell}(\mathbf{x}_t) = \{&(\mathbf{x}_s,\mathbf{x}_{s'},\mathbf{x}_{s''}) \\ &|(\mathbf{x}_{s''} - \mathbf{x}_s) \in \mathcal{N}_k^{S}(\mathbf{x}_s,\mathcal{D}_{y_s}^{+}(\mathbf{x}_s)) \ \&\ (\mathbf{x}_{s'} - \mathbf{x}_s) \in \mathcal{D}_{y_s}^{+}(\mathbf{x}_s), \\ &g(\mathbf{x}_t,\mathbf{x}_s) = 1,\ y_{s'} = y_s,\ y_{s''} \neq y_s, \\ &N_T + 1 \leq s, s', s'' \leq N\} \end{aligned} \tag{18}$$

$$\begin{aligned} \mathbb{O}_b^{\ell}(\mathbf{x}_t) = \{&(\mathbf{x}_{t'},\mathbf{x}_s) \\ &|(\mathbf{x}_s - \mathbf{x}_t) \in \mathcal{N}_k^{S}(\mathbf{x}_t,\mathcal{D}_{y_t}^{-}(\mathbf{x}_t)) \ \&\ (\mathbf{x}_{t'} - \mathbf{x}_t) \in \mathcal{D}_{y_t}^{-}(\mathbf{x}_t), \\ &g(\mathbf{x}_t,\mathbf{x}_s) = 0,\ y_t \neq y_{t'},\ y_s \neq y_t, \\ &1 \leq t' \leq N_T,\ N_T + 1 \leq s \leq N\} \end{aligned} \tag{19}$$

In addition, let $\#\mathbb{O}_g^{\ell} = \sum_{t=1}^{N_T} \#\mathbb{O}_g^{\ell}(\mathbf{x}_t)$, $\#\mathbb{O}_a^{\ell} = \sum_{t=1}^{N_T} \#\mathbb{O}_a^{\ell}(\mathbf{x}_t)$, $\#\mathbb{O}_b^{\ell} = \sum_{t=1}^{N_T} \#\mathbb{O}_b^{\ell}(\mathbf{x}_t)$. Following Eq. (14) and Eq. (15), we propose the following criterion for optimisation

$$\begin{aligned} &\min_{\mathbf{M}\succeq 0} f(\mathbf{M}) + \frac{\lambda}{2(1-\lambda)}||\mathbf{M}||_F^2 \\ &f(\mathbf{M}) = \frac{1-\alpha}{\#\mathbb{O}_g^{\ell}} \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_{t_j},\mathbf{x}_s,\mathbf{x}_{t'}) \in \mathbb{O}_g^{\ell}(\mathbf{x}_t)} \ell_h(d(\mathbf{x}_{t_j},\mathbf{x}_s) < d(\mathbf{x}_t,\mathbf{x}_{t'})) \\ &+ \frac{\alpha}{\#\mathbb{O}_a^{\ell} + \#\mathbb{O}_b^{\ell}} \Big\{ \\ &\quad \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_s,\mathbf{x}_{s'},\mathbf{x}_{s''}) \in \mathbb{O}_a^{\ell}(\mathbf{x}_t)} \ell_h(d(\mathbf{x}_s,\mathbf{x}_{s'}) < d(\mathbf{x}_s,\mathbf{x}_{s''})) \\ &\quad + \beta \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_{t'},\mathbf{x}_s) \in \mathbb{O}_b^{\ell}(\mathbf{x}_t)} \ell_h(d(\mathbf{x}_t,\mathbf{x}_{t'}) < d(\mathbf{x}_t,\mathbf{x}_s)) \Big\} \end{aligned} \tag{20}$$

**Algorithm 1:** Learning Procedure for t-LRDC model

**Data**: Dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$, Maximum Iteration $P$, $\varepsilon > 0$
**begin**
  $\mathbf{M}_0 \longleftarrow D^{-1} \cdot \mathbf{I}$ ;
  $n \longleftarrow 0$;
  **while** $n \leq P$ **do**
    **Active Set**:
      Compute active sets $\mathbb{O}_g^\ell(\mathbf{x}_t, n), \mathbb{O}_a^\ell(\mathbf{x}_t, n), \mathbb{O}_b^\ell(\mathbf{x}_t, n)$;
      by Eqs. (21)-(23);
    **Gradient Descent**:
      Compute the gradient matrix $\Delta_\mathbf{M}$ by Eq. (27);
      $\overline{\mathbf{M}_{n+1}} \longleftarrow \mathbf{M}_n - \eta_n \cdot \Delta_\mathbf{M}$, $\eta_n = \frac{1}{n+1}$;
    **Projection**:
      Project $\overline{\mathbf{M}_{n+1}}$ onto $\mathbb{O}^+$ by Eq. (29)
      and obtain $\mathbf{M}_{n+1}$;
    **if** $||\mathbf{M}_n - \mathbf{M}_{n+1}||_F^2 < \varepsilon$ **then**
      break;
    **end**
    $n \longleftarrow n + 1$;
  **end**
**end**
**Output**: $\mathbf{M} = \mathbf{M}_{n+1}$

Note that in the above formulation, a ridge regularisation term is introduced on the matrix $\mathbf{M}$ parameterised by some non-negative $\lambda(\in [0, 1))$ in order to gain better generalisation ability. We call the above model (Eq. (20)) the *transfer local relative distance comparison* (t-LRDC). Note that since less relative comparison pairs are used in t-LRDC at each step for optimisation, solving the cost function in Eq. (20) is less costly than in Eq. (14).

One would like to know what price if any this reduction in the number of comparisons will pay in terms of model performance. To answer that, we show that under a set of general conditions, using local relative comparison would not lead to loss in relative comparison as compared to performing all relative comparisons (i.e. global relative comparison), by the following theorem, where the proof is obvious and thus omitted due to lack of space.

*Theorem 1:* Local relative comparison is equivalent to global relative comparison if all relative comparisons in Eq. (20) hold and the following statements are true

1) Given $\mathbf{x}_t$, $t \leq N_T$, for any $(\mathbf{x}_{t_j}, \mathbf{x}_s, \mathbf{x}_{t'}) \in \mathbb{O}_g(\mathbf{x}_t)$, $d(\mathbf{x}_t, \mathbf{x}_{t'}) \geq \min_{(\mathbf{x}_{t_j}, \mathbf{x}_{j''}) \in \mathbb{O}_g^\ell(\mathbf{x}_t)} d(\mathbf{x}_t, \mathbf{x}_{j''})$;
2) Given $\mathbf{x}_t$, $t \leq N_T$, for any $(\mathbf{x}_s, \mathbf{x}_{s'}, \mathbf{x}_{s''}) \in \mathbb{O}_a(\mathbf{x}_t)$, $d(\mathbf{x}_s, \mathbf{x}_{s''}) \geq \min_{(\mathbf{x}_s, \mathbf{x}_{j'}, \mathbf{x}_{j''}) \in \mathbb{O}_a^\ell(\mathbf{x}_t)} d(\mathbf{x}_s, \mathbf{x}_{j''})$;
3) Given $\mathbf{x}_t$, $t \leq N_T$, for any $(\mathbf{x}_{t'}, \mathbf{x}_s) \in \mathbb{O}_b(\mathbf{x}_t)$, $d(\mathbf{x}_t, \mathbf{x}_s) \geq \min_{(\mathbf{x}_{t'}, \mathbf{x}_{j'}) \in \mathbb{O}_b^\ell(\mathbf{x}_t)} d(\mathbf{x}_t, \mathbf{x}_{j'})$;

In practice, as shown in Sec. IV, due to the selection of constraints, better verification performance is obtained overall.

Note that the formations of $\mathcal{N}_k^T(\cdot,\cdot)$ and $\mathcal{N}_k^S(\cdot,\cdot)$ in Eq. (16) are dependent of the Mahalanobis metric (Eq. (1)) parameterised by $\mathbf{M}$, which is to be learned. It is important to point out that although it may appear to be a conundrum that $\mathbf{M}$ is used to describe locality *before* it is yet to be learned, we shall introduce a stochastic method to overcome this problem in the next section, so that the neighbourhood is updated *simultaneously* when $\mathbf{M}$ is updated during the optimisation process.

### D. Optimisation Algorithm

We solve the optimisation problem in (20) using stochastic gradient [3], [35], [42]. The stochastic gradient method can be used for sum-minimisation, that is, rather than performing the batch gradient for all terms in a sum, it selects parts of the sum at each iteration to compute the gradient. It is thus suitable for large scale machine learning tasks and has better generalisation capability compared to alternatives [3]. In particular, for our problem it focuses its computation on the sample pairs that violate the constraints in the Sets $\mathbb{O}_g^\ell(\mathbf{x}_t)$, $\mathbb{O}_b^\ell(\mathbf{x}_t)$ and $\mathbb{O}_a^\ell(\mathbf{x}_t)$ leading to an efficient algorithm.

More specifically, at the $n^{th}$ step during the iterative optimisation process, we first construct three active sets which consist of distance comparison pairs that violate the relative comparison constraints of $\mathbb{O}_g^\ell$, $\mathbb{O}_a^\ell$ and $\mathbb{O}_b^\ell$ respectively:

$$\begin{aligned}\mathbb{O}_g^\ell(\mathbf{x}_t, n) = \{(\mathbf{x}_{t_j}, \mathbf{x}_s, \mathbf{x}_{t'})|\ & d_n(\mathbf{x}_{t_j}, \mathbf{x}_s) \geq d_n(\mathbf{x}_t, \mathbf{x}_{t'}),\\ & (\mathbf{x}_{t'} - \mathbf{x}_t) \in \mathcal{N}_k^T(\mathbf{x}_t, \mathcal{D}_{y_{t_j}}^+(\mathbf{x}_{t_j}), n) \& \ (\mathbf{x}_s - \mathbf{x}_{t_j}) \in \mathcal{D}_{y_{t_j}}^+(\mathbf{x}_{t_j}),\\ & g(\mathbf{x}_t, \mathbf{x}_{t_j}) = 1,\ y_{t_j} = y_s, y_t \neq y_{t'},\\ & 1 \leq t' \leq N_T,\ N_T + 1 \leq t_j, s \leq N\}\end{aligned} \tag{21}$$

$$\begin{aligned}\mathbb{O}_a^\ell(\mathbf{x}_t, n) = \{(\mathbf{x}_s, \mathbf{x}_{s'}, \mathbf{x}_{s''})|\ & d_n(\mathbf{x}_s, \mathbf{x}_{s'}) \geq d_n(\mathbf{x}_s, \mathbf{x}_{s''}),\\ & (\mathbf{x}_{s''} - \mathbf{x}_s) \in \mathcal{N}_k^S(\mathbf{x}_s, \mathcal{D}_{y_s}^+(\mathbf{x}_s), n) \ \& \ (\mathbf{x}_{s'} - \mathbf{x}_s) \in \mathcal{D}_{y_s}^+(\mathbf{x}_s),\\ & g(\mathbf{x}_t, \mathbf{x}_s) = 1,\ y_{s'} = y_s,\ y_{s''} \neq y_s,\\ & N_T + 1 \leq s, s', s'' \leq N\}\end{aligned} \tag{22}$$

$$\begin{aligned}\mathbb{O}_b^\ell(\mathbf{x}_t, n) = \{(\mathbf{x}_{t'}, \mathbf{x}_s)|\ & d_n(\mathbf{x}_t, \mathbf{x}_{t'}) \geq d_n(\mathbf{x}_t, \mathbf{x}_s),\\ & (\mathbf{x}_s - \mathbf{x}_t) \in \mathcal{N}_k^S(\mathbf{x}_t, \mathcal{D}_{y_t}^-(\mathbf{x}_t), n) \ \& \ (\mathbf{x}_{t'} - \mathbf{x}_t) \in \mathcal{D}_{y_t}^-(\mathbf{x}_t),\\ & g(\mathbf{x}_t, \mathbf{x}_s) = 0,\ y_t \neq y_{t'},\ y_s \neq y_t,\\ & 1 \leq t' \leq N_T,\ N_T + 1 \leq s \leq N\}\end{aligned} \tag{23}$$

where the distance function $d_n$ is computed as

$$d_n(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T \mathbf{M}_n (\mathbf{x} - \mathbf{x}'), \tag{24}$$

and $\mathbf{M}_n$ is the metric learned at the last iterative step. $\mathcal{N}_k^T(\mathbf{x}_t, \mathcal{D}_{y_{t_j}}^+(\mathbf{x}_{t_j}), n)$, $\mathcal{N}_k^S(\mathbf{x}_s, \mathcal{D}_{y_s}^+(\mathbf{x}_s), n)$, and $\mathcal{N}_k^S(\mathbf{x}_t, \mathcal{D}_{y_t}^-(\mathbf{x}_t), n)$ mean forming the nearest neighbouring sets $\mathcal{N}_k^T(\mathbf{x}_t, \mathcal{D}_{y_{t_j}}^+(\mathbf{x}_{t_j}))$, $\mathcal{N}_k^S(\mathbf{x}_s, \mathcal{D}_{y_s}^+(\mathbf{x}_s))$, and $\mathcal{N}_k^S(\mathbf{x}_t, \mathcal{D}_{y_t}^-(\mathbf{x}_t))$ by first projecting the difference vectors into the low-rank subspace induced by the $\mathbf{M}_n$ learned in Eq. (24) at the $n^{th}$ step, respectively. Using the three active sets $\mathbb{O}_g^\ell(\mathbf{x}_t, n)$, $\mathbb{O}_b^\ell(\mathbf{x}_t, n)$ and $\mathbb{O}_a^\ell(\mathbf{x}_t, n)$, we compute the gradient for the following function:

$$f(\mathbf{M}, n) + \frac{\lambda}{2(1-\lambda)} ||\mathbf{M}||_F^2, \tag{25}$$

where

$$\begin{aligned}f(\mathbf{M}, n) = & \\ & \frac{1-\alpha}{\#\mathbb{O}_g^\ell(n)} \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_{t_j}, \mathbf{x}_s, \mathbf{x}_{t'}) \in \mathbb{O}_g^\ell(\mathbf{x}_t, n)} \ell_h(d(\mathbf{x}_{t_j}, \mathbf{x}_s) < d(\mathbf{x}_t, \mathbf{x}_{t'}))\\ & + \frac{\alpha}{\#\mathbb{O}_a^\ell(n) + \#\mathbb{O}_b^\ell(n)} \Big\{ \\ & \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_s, \mathbf{x}_{s'}, \mathbf{x}_{s''}) \in \mathbb{O}_a^\ell(\mathbf{x}_t, n)} \ell_h(d(\mathbf{x}_s, \mathbf{x}_{s'}) < d(\mathbf{x}_s, \mathbf{x}_{s''}))\\ & + \beta \sum_{t=1}^{N_T} \sum_{(\mathbf{x}_{t'}, \mathbf{x}_s) \in \mathbb{O}_b^\ell(\mathbf{x}_t, n)} \ell_h(d(\mathbf{x}_t, \mathbf{x}_{t'}) < d(\mathbf{x}_t, \mathbf{x}_s)) \Big\}\end{aligned} \tag{26}$$

and $\#\mathbb{O}_g^\ell(n) = \sum_{t=1}^{N_T} \#\mathbb{O}_g^\ell(\mathbf{x}_t, n)$, $\#\mathbb{O}_a^\ell(n) = \sum_{t=1}^{N_T} \#\mathbb{O}_a^\ell(\mathbf{x}_t, n)$, and $\#\mathbb{O}_b^\ell(n) = \sum_{t=1}^{N_T} \#\mathbb{O}_b^\ell(\mathbf{x}_t, n)$.

Denote the gradient of Eq. (25) with respect to $\mathbf{M}$ by $\Delta_\mathbf{M}$. Then, $\Delta_\mathbf{M}$ is computed as follows:

$$\begin{aligned}\Delta_{\mathbf{M}} = &\frac{\lambda}{(1-\lambda)}\mathbf{M}+\\ &\frac{2(1-\alpha)}{\#\mathbb{O}_g^\ell(n)}\sum_{t=1}^{N_T}\sum_{(\mathbf{x}_{t_j},\mathbf{x}_s,\mathbf{x}_{t'})\in\mathbb{O}_g^\ell(\mathbf{x}_t,n)} C_{t,t_j,s,t'}\Big(\\ &(\mathbf{x}_{t_j}-\mathbf{x}_s)(\mathbf{x}_{t_j}-\mathbf{x}_s)^T-(\mathbf{x}_t-\mathbf{x}_{t'})(\mathbf{x}_t-\mathbf{x}_{t'})^T\Big)\\ &+\frac{2\alpha}{\#\mathbb{O}_a^\ell(n)+\#\mathbb{O}_b^\ell(n)}\Big\{\\ &\sum_{t=1}^{N_T}\sum_{(\mathbf{x}_s,\mathbf{x}_{s'},\mathbf{x}_{s''})\in\mathbb{O}_a^\ell(\mathbf{x}_t,n)} C_{s,s',s''}\Big(\\ &(\mathbf{x}_s-\mathbf{x}_{s'})(\mathbf{x}_s-\mathbf{x}_{s'})^T-(\mathbf{x}_s-\mathbf{x}_{s''})(\mathbf{x}_s-\mathbf{x}_{s''})^T\Big)\\ &+\sum_{t=1}^{N_T}\sum_{(\mathbf{x}_{t'},\mathbf{x}_s)\in\mathbb{O}_b^\ell(\mathbf{x}_t,n)} C_{t,t',s}\Big(\\ &(\mathbf{x}_t-\mathbf{x}_{t'})(\mathbf{x}_t-\mathbf{x}_{t'})^T-(\mathbf{x}_t-\mathbf{x}_s)(\mathbf{x}_t-\mathbf{x}_s)^T\Big)\Big\}\end{aligned} \tag{27}$$

where $C_{t,t_j,s,t'} = d(\mathbf{x}_{t_j},\mathbf{x}_s)+\rho-d(\mathbf{x}_t,\mathbf{x}_{t'}), C_{s,s',s''} = d(\mathbf{x}_s,\mathbf{x}_{s'})+\rho-d(\mathbf{x}_s,\mathbf{x}_{s''}), C_{t,t',s} = d(\mathbf{x}_t,\mathbf{x}_{t'})+\rho-d(\mathbf{x}_t,\mathbf{x}_s)$. It shows that for each update, the gradient matrix is generated using weighted covariance matrices associated to the three active sets, where the weights are $C_{t,t_j,s,t'}, C_{s,s',s''}, C_{t,t',s}$ respectively. Note that these weights must be positive, as the stochastic gradient update here only selects those constraint-violated relative comparisons for update. So, the larger the weight is, the more constraint-violated a relative comparison is.

Finally, the updated $\mathbf{M}_{n+1}$ is obtained by

$$\overline{\mathbf{M}_{n+1}} = \mathbf{M}_n - \eta_n \cdot \Delta_{\mathbf{M}}, \tag{28}$$

where the learning rate $\eta_n$ should decrease as more steps are taken with an appropriate rate. In our work, we simply let $\eta_n = (n+1)^{-1}$.

Note that the computed core matrix $\overline{\mathbf{M}_{n+1}}$ in Eq. (28) is symmetric but not always semi-positive definite. Thus, in the projected stochastic gradient method, to make sure a semi-positive core matrix is learned at each step, the core matrix $\overline{\mathbf{M}_{n+1}}$ has to be projected onto the solution set denoted by $\mathbb{O}^+$, which is the set of semi-positive matrices of the same size as $\overline{\mathbf{M}_{n+1}}$. This is achieved by finding a semi-positive matrix $\mathbf{M}_{n+1}$ by

$$\mathbf{M}_{n+1} = \arg\min_{\mathbf{A}\in\mathbb{O}^+} ||\mathbf{A}-\overline{\mathbf{M}_{n+1}}||_F^2, \tag{29}$$

It can be verified that $\mathbf{M}_{n+1}$ in Criterion (29) is computed by

$$\mathbf{M}_{n+1} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^T, \tag{30}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with each diagonal term being the positive eigenvalue of matrix $\overline{\mathbf{M}_{n+1}}$ and each column of $\mathbf{L}$ is the corresponding eigenvector.

The above steps are repeated and will terminate when a stopping criterion is met. The whole algorithm is summarised in Algorithm 1.

### E. Linear Dimensionality Reduction

In person re-identification, the feature dimension is typically high (e.g. larger than 1000); it is thus computationally expensive to perform eigen-value decomposition in the above learning algorithm for our t-LRDC model. In this section, we show that this problem can be alleviated by adding dimensionality reduction pre-processing step without sacrifice of model learning performance.

In particular, using eigen-value decomposition, we can factorise matrix $\mathbf{M}$ into $\mathbf{M} = \mathbf{P}\mathbf{P}^T$, where $\mathbf{P} = [\mathbf{p}_1,\cdots,\mathbf{p}_m]$. Let $\mathcal{U} = span\{(\mathbf{x}_i-\mathbf{x}_j)|1\le i,j\le N\}$. Then we can have the following theorem.

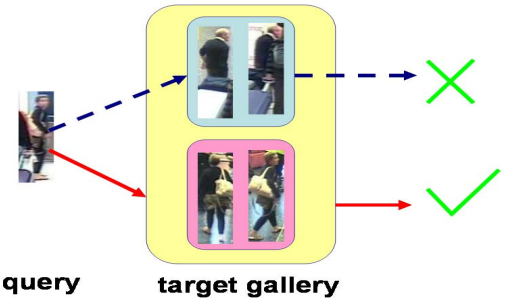



Fig. 4. For individual verification, if the query image is matched to a wrong target person (as shown by the dashed blue line), the match is incorrect. In contrast, for set verification, the match is correct as long as the query image is matched to one of the target people (as shown by the solid red line).

*Theorem 2:* Given training data $\mathbf{x}_1,\cdots,\mathbf{x}_N$, it is sufficient to learn $\mathbf{p}_i$ in $\mathcal{U}$

*Proof:* If $\mathbf{p}_i$ is not in $\mathcal{U}$, then there exists $\mathbf{q}_i\in\mathcal{U}$ and $\mathbf{v}_i\in\mathcal{U}^\perp$ such that $\mathbf{p}_i = \mathbf{q}_i+\mathbf{v}_i$. That is $\mathbf{P} = \mathbf{Q}+\mathbf{V}$, where $\mathbf{Q} = [\mathbf{q}_1,\cdots,\mathbf{q}_m]$ and $\mathbf{V} = [\mathbf{v}_1,\cdots,\mathbf{v}_m]$. Hence, for any pairwise $(\mathbf{x}_i,\mathbf{x}_j)$ the distance between them is

$$\begin{aligned}&(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i-\mathbf{x}_j)\\ =&(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{P}\mathbf{P}^T(\mathbf{x}_i-\mathbf{x}_j) = (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{Q}\mathbf{Q}^T(\mathbf{x}_i-\mathbf{x}_j).\end{aligned} \tag{31}$$

This is because for any $(\mathbf{x}_i-\mathbf{x}_j)$, we have $\mathbf{V}^T(\mathbf{x}_i-\mathbf{x}_j) = \mathbf{0}$. So the above equation suggests that the information out of the range space of $\mathbf{X}$ is not useful to construct matrix $\mathbf{M}$ for optimising the criterion based on training data $\mathbf{x}_1,\cdots,\mathbf{x}_N$. ■

The above theorem suggests we can first reduce the data dimensionality by projecting the data onto $\mathcal{U}$ and then perform t-LRDC learning. One can verify that the projection for dimension reduction is equivalent to learning by principal component analysis (PCA) [41]. Based on this theorem we thus propose to learn an approximate t-LRDC model in the PCA space by retaining the main dimensions corresponding to large eigenvalues, resulting in a much smaller data dimension and more efficient model learning. In our experiments, the numbers of basis vectors in $\mathcal{U}$ are 207, 203, 424 and 631 on i-LIDS, CAVIAR, ETHZ and VIPeR, respectively, reduced from its original 2784 dimensional feature space.

### F. Group-based Person Verification

Now with the learned distance function we can measure the distance between a probe/query image, which may or may not belong to one of the target people, and a set of gallery images containing the target people. We call this process of verifying target people on the watch list as *group-based person verification*. As motivated in Sec. I, in practice, it is often more desirable to perform verification against the whole set, which we term as *'Set Verification'*. In our implementation, the distance between a query and the set is the minimal distance between it and any person image of all people on the watch list. Alternatively, the learned distance can also be used to perform a more conventional task, that is, to verify whether this query image comes from a target person $C_k^t$ and not from any of the others (including the other target people). This is termed as *'Individual Verification'*.

The difference between set verification and individual verification is illustrated in Figure 4. In particular, the individual verifi-

cation can be considered as a special case of set verification. The difference between these two verifications is that set verification tells whether the detected person is within our interest but does not perform verification on the person identity of any query image. The individual verification performs the latter but would not be able to measure explicitly the probability that the person of the query image is on the watch-list. Their relation is similar to the relation between joint probability density function and marginal probability density function. When there is only one person in the group, set verification is the same as individual verification.

### G. Discussions

**Relations to existing models.** As mentioned in Sec. II, there are two main features that distinguish our t-LRDC model from the existing learning methods used for person re-id: (1) Our model is learned using three types of knowledge transfer/constraints (see Sec. III-B); the first two are designed to solve the data sparsity problems (one-shot and small watch list) and the third for the open-world group-based person verification task first identified in this paper. Most existing learning methods would not work with one-shot per target person. The best they could do under our setting is to learn their models using the source/non-target training data and then apply the models to the target people without adaptation. This is similar to using our second type of constraints except that we use the target person to select a subset of non-target people to form those constraints. None of the existing models utilise the third types of knowledge transfer. They are thus intrinsically not designed for tackling the group-based verification task. (2) Our model is a local relative distance comparison method focusing on comparison pairs in a local neighbourhood between similar distances. This has two benefits: lower computational cost and memory usage; and avoiding model over-fitting by removing those hard/global relative comparison constraints. All existing alternative relative comparison models enforce global constraints thus do not have these two benefits.

**Generalisation of existing models.** It is however possible to generalise some of the existing models for the group-based verification task. Specifically, using the risk functions and distance functions in relative distance comparison (RDC) [50] to replace the loss function $\ell_h$ and distance function $d$ in Eq. (14), respectively, we are able to develop a *transfer RDC* (t-RDC) model which uses exactly the same three types of constraints as our model does. In particular, we re-define the loss function $\ell$ and distance function $d$ as

$$d(\mathbf{x},\mathbf{x}') = |\mathbf{x}-\mathbf{x}'|^T\mathbf{M}|\mathbf{x}-\mathbf{x}'|,$$
$$\ell(d(\mathbf{x}',\mathbf{x}'') < d(\mathbf{x}',\mathbf{x}''')) = \log\left(1+\exp\left\{d(\mathbf{x}',\mathbf{x}'')-d(\mathbf{x}',\mathbf{x}''')\right\}\right),$$

where $|\mathbf{x}-\mathbf{x}'|$ is the absolute data difference vector [50]. Similarly we can generalise RankSVM [31] by first redefining the distance $d$ as a projection response function below

$$d(\mathbf{x},\mathbf{x}') = -\mathbf{M}^T|\mathbf{x}-\mathbf{x}'|, \tag{32}$$

where the $\mathbf{M}$ is defined as a projection vector. Then, by using the above function $d$, the hinge-loss function and further inserting a regularisation term $||\mathbf{M}||^2$ in Eq. (14), we can develop the *transfer RankSVM* (t-RankSVM).

**Limitations of the generalisations of existing models.** Although the generalised models above can now tackle the open-world group-based verification problem, there are still two unsolved problems limiting their capabilities. (1) Both t-RDC and t-RankSVM still perform global relative comparison. It is not straightforward and may not even be possible to introduce the proposed local modelling. In particular, for t-RDC this is due to the sequential learning procedure used which is quite different from the stochastic gradient decent method used in our model; whilst t-RankSVM in essence optimises based on a margin rather than a real distance; thus the concept of local modelling does not apply. (2) Both t-RDC and t-RankSVM rely on using absolute data difference to measure the difference between two data points. Using absolute data difference would lead to notably larger cost of memory use. This is because the absolute operation prevents dimension reduction directly on high-dimensional data without loss of discriminant ability (see Sec. III-E). This is verified by our experiments which show that t-RDC and t-RankSVM with the same PCA pre-processing step fail dramatically. In comparison, the proposed model is not based on the absolute data difference and dimension reduction can be used before learning without loss in performance as proved in Sec. III-E.


Fig. 5. Examples of Images for the four datasets. Images of each column are from the same person.

## IV. Experiments

### A. Datasets and Settings

**Datasets**. Four widely used benchmark datasets are used in our experiments. They include the i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset [47], [39], [48], the ETHZ dataset [34], [7], the CAVIAR4REID (CAVIAR) dataset [4] and the VIPeR dataset [13]. The i-LIDS MCTS dataset consists of 119 people with a total 476 person images with an average of 4 images, which were captured by multiple non-overlapping cameras indoor at a busy airport arrival hall. Many of these images undergo large illumination change and are subject to occlusions. The ETHZ dataset consists of 146 people and 8555 images in total, which were captured using a moving camera in a busy street scene. The CAVIAR dataset contains images from 72 people, where 10 images were randomly selected for each person. The VIPeR dataset consists of 632 people captured outdoor with two images for each person with a normalised size of 128 $\times$ 64 pixels. View angle change is the most significant cause of appearance

change with most of the matched image pairs containing one front/back view and one side-view. Overall, these four datasets cover different condition changes across camera views and are representative of the real-world person re-id challenges. Figure 5 shows examples of images for each dataset.

**Feature Representation**. The popular histogram based feature representation for person re-identification [14], [30], [47], [48] is adopted, which is a mixture of colour features (including RGB, YCbCr, HSV color) and texture feature pattern (extracted by Schmid and Gabor filters). Each image is represented by a feature vector in a 2784 dimensional feature space.

**Compared methods**. We compare the proposed t-LRDC model with the following alternative models. (1) The baseline is the L1-Norm distance metric, which does not rely on any learning. (2) *Naive transfer learning models*. These models are not designed for extracting knowledge from the non-target training set and adapt it towards the target data. The knowledge transfer is thus 'naive'.These include recent distance/subspace learning methods designed for person re-id, such as KISSME [17], LADF [22], salience modelling [46] and LFDA [29], One-class SVM (OCSVM) [33], two local distance learning methods LMNN [42] and local distance learning method LDM [44], and two models based on relative comparison: RDC [50] and RankSVM [31]. Among them, the salience modelling in [46] is unsupervised. (3) *Generalisation of existing models for group-based verification*. These include two alternative transfer models developed in this work, namely the t-RDC and t-RankSVM described in Sec. III-G. (4) A variant of our t-LRDC model without the local modelling, t-LRDC(global) which instead of using local relative distance comparison constraints, uses global relative comparison. This is to evaluate the effect of local modelling.

**Experimental settings**. Our open-world group-based person verification experiments are designed to verify whether a query person image comes from the people on a watch-list with the presence of non-target person images during the verification. More specifically, for each dataset, we randomly selected all images of $p$ people (classes) to set up the target data set and the rest to constitute the non-target data set. The target data set was further divided into a training set and a testing set, where one image of each person was randomly selected for training (one-shot). We also randomly divided the non-target data set into training and testing sets, where six images at maximum were randomly selected for each non-target person. Such a random division was done by person; that is, the images of half of the non-target people in the data set were used as training non-target person images and the rest as testing non-target images so that there is no overlap of non-target people between the training and testing sets. The experiment was conducted 10 times and the average verification performance was then computed. For verification, on each dataset, both individual verification and set verification (see Sec. III-F for definition) are reported. In our experiments, the number of target people (i.e. $p$) was fixed to be 6 for each round. The effect of gallery size will be discussed later.

For all iterative methods, the maximum iteration was set to 100. PCA dimension reduction (by preserving 100% data energy) was used for all distance models except RDC and RankSVM, making them tractable on a PC platform. There are five free parameters in our t-LRDC model: $\lambda$, $\alpha$ and $\beta$ in Eq. (20), $h$ in Eq. (3) and $k$ in Eq. (16). By default, we set $\lambda = 0.3$, $\alpha = 0.8$, $\beta = 0.6$, $h = 0.72$ and $k = 3$ in all experiments. We found that the result of our model is less sensitive to the value of $\lambda$; the effect of other parameters will be discussed in details in Sec. IV-D. The parameters of the compared methods were set to the same as described in the original work where they were first introduced.

**Evaluation metrics**. There is no metrics in person re-identification that can be used readily for an open-world group-based verification task. We thus have to define a set of new ones. In particular, since a lot of images of non-target people were mixed with the target ones as query images, we need to quantify the performance on how well a true target has been verified and how bad a false target has passed through the verification and their relations. Therefore, we introduce the true target rate (TTR) and false target rate (FTR) as follows:

$$\text{True Target Recognition(TTR)} = \frac{\#\mathcal{TTQ}}{\#\mathcal{TQ}}, \quad (33)$$

$$\text{False Target Recognition(FTR)} = \frac{\#\mathcal{FNTQ}}{\#\mathcal{NTQ}}. \quad (34)$$

where

$\mathcal{TQ} = \{\text{query target images from target people}\}$;
$\mathcal{NTQ} = \{\text{query non-target images from non-target people}\}$;
$\mathcal{TTQ} = \{\text{query target images that are verified as one of the target people}\}$;
$\mathcal{FNTQ} = \{\text{query non-target images that are verified as one of the target people}\}$.

Note that for performing individual verification for each target person (see Sec. III-F), the above metrics can still be used, and in this case the non-target people mean any other person except that target person.

In our experiments, the similarity function $sim(\mathbf{x}, \mathbf{x}')$ is specified as the inverse of computed distance to determine the rank of matching. A value $r$ is used to threshold these scores and therefore results of the TTR value against FTR value are reported for each method by changing the threshold value $r$. This is similar to the ROC performance in face verification, but it differs in that we also care about the verification on whether the query image belongs to one of the target people (i.e. set verification).

### *B. Comparison with Naive Transfer Models*

The compared naive transfer models have two types depending on whether they are based on relative comparison constraints. The TTR vs. FTR performance of our t-LRDC against those not based on relative comparisons (KISSME [17], OCSVM [33], LMNN [42], LDM [44], LADF [22] and LFDA [29]) as well as the L1-norm baseline and salience modelling [46] are shown in Tables I and II. Note that due to its unsupervised nature, salience modelling can naturally be used for the one-shot open world person re-identification. The results show clearly that our model significantly outperforms the six alternative supervised transfer learning models and the unsupervised salience model. The performance gap is specially significant on the two more challenging i-LIDS and VIPeR datasets where the view angle changes across camera views cause different people look alike. It can be seen that under this open-world setting, many learning based models (e.g. LDM on i-LIDS) yield even poorer performance than the non-learning based L1-norm baseline. This suggests that blindly transferring knowledge extracted from non-target data without adaptation can have a negative impact, i.e. negative transfer. This negative transfer problem does not exist in conventional setting where the gallery and probe set always contain the same set of

| Database | i-LIDS | | | | | | ETHZ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FTR | 0.1% | 1% | 5% | 10% | 20% | 30% | 0.1% | 1% | 5% | 10% | 20% | 30% |
| t-LRDC | 14.58 | **32.03** | **48.36** | **61.64** | 74.57 | 81.65 | 45.13 | 65.62 | 83.82 | 89.86 | 95.20 | 97.67 |
| t-LRDC(Global) | 13.45 | 30.94 | 47.35 | 61.07 | **76.66** | **87.30** | 42.22 | 62.72 | 79.95 | 86.69 | 92.45 | 96.89 |
| t-RDC | **16.78** | 30.98 | 45.31 | 57.12 | 72.07 | 81.91 | **54.14** | **76.29** | **88.07** | 91.91 | 96.02 | 98.38 |
| t-RankSVM | 14.31 | 27.12 | 42.06 | 55.10 | 70.86 | 77.31 | 50.49 | 74.70 | 87.82 | **92.72** | **96.60** | **99.09** |
| t-RDC-PCA | 10.85 | 24.49 | 39.39 | 49.64 | 63.57 | 70.92 | 42.33 | 61.57 | 76.23 | 82.76 | 89.54 | 92.94 |
| t-RankSVM-PCA | 7.44 | 17.06 | 36.76 | 46.76 | 60.31 | 70.05 | 35.13 | 55.75 | 74.98 | 81.72 | 87.54 | 91.36 |
| RDC [50] | 15.16 | 28.04 | 44.89 | 57.53 | 70.89 | 79.99 | 53.16 | 75.07 | 87.30 | 91.67 | 95.16 | 97.63 |
| RankSVM [31] | 12.09 | 23.66 | 40.97 | 56.07 | 69.26 | 77.76 | 47.87 | 72.40 | 86.62 | 91.56 | 95.96 | 98.82 |
| OCSVM [33] | 6.00 | 6.34 | 11.78 | 17.87 | 28.59 | 36.25 | 0.56 | 2.23 | 11.62 | 18.36 | 28.11 | 35.12 |
| KISSME [17] | 11.77 | 25.46 | 36.74 | 44.92 | 61.00 | 67.79 | 46.49 | 61.21 | 76.31 | 85.33 | 93.06 | 96.94 |
| LMNN [42] | 8.61 | 20.81 | 41.43 | 49.92 | 58.00 | 68.85 | 41.80 | 58.65 | 75.43 | 82.59 | 90.74 | 93.43 |
| LDM [44] | 8.51 | 18.24 | 39.08 | 48.80 | 61.65 | 72.96 | 29.76 | 49.80 | 69.37 | 78.05 | 86.01 | 90.83 |
| LADF [22] | 7.86 | 20.72 | 39.88 | 53.80 | 69.29 | 79.89 | 20.23 | 53.14 | 76.67 | 85.86 | 93.67 | 96.42 |
| LFDA [29] | 7.22 | 13.43 | 24.72 | 35.47 | 50.11 | 63.74 | 27.49 | 43.98 | 60.96 | 73.52 | 84.83 | 89.23 |
| Salience [46] | 6.00 | 6.10 | 8.07 | 11.81 | 20.40 | 29.48 | 26.87 | 44.76 | 55.85 | 63.09 | 71.80 | 79.92 |
| L1-norm | 8.42 | 19.90 | 43.50 | 53.22 | 60.53 | 69.29 | 42.39 | 60.47 | 77.45 | 84.45 | 89.52 | 92.97 |
| Database | CAVIAR | | | | | | VIPeR | | | | | |
| FTR | 0.1% | 1% | 5% | 10% | 20% | 30% | 0.1% | 1% | 5% | 10% | 20% | 30% |
| t-LRDC | **15.45** | **28.13** | **40.76** | **50.78** | 60.80 | 69.99 | **23.47** | **47.27** | 75.41 | 86.88 | 98.04 | 99.17 |
| t-LRDC(Global) | 13.78 | 25.85 | 39.87 | 49.01 | 63.18 | 71.58 | 19.63 | 39.04 | 69.25 | 84.13 | 96.17 | 98.13 |
| t-RDC | 14.08 | 24.40 | 39.30 | 49.13 | **63.59** | **71.82** | 19.38 | 40.72 | 73.18 | 88.42 | 97.85 | 98.71 |
| t-RankSVM | 10.64 | 20.90 | 35.67 | 45.50 | 57.83 | 68.23 | 22.73 | 45.95 | 76.12 | **89.44** | 97.86 | 98.99 |
| t-RDC-PCA | 12.48 | 23.38 | 36.55 | 45.43 | 57.65 | 66.49 | 18.79 | 24.73 | 40.03 | 54.54 | 76.71 | 85.61 |
| t-RankSVM-PCA | 12.41 | 20.00 | 33.37 | 42.23 | 55.73 | 63.98 | 17.53 | 21.60 | 34.38 | 44.12 | 68.77 | 79.58 |
| RDC [50] | 14.61 | 23.40 | 37.32 | 47.08 | 59.40 | 69.15 | 19.27 | 43.98 | **77.95** | 88.62 | 96.00 | 99.89 |
| RankSVM [31] | 6.33 | 16.64 | 31.43 | 42.04 | 56.25 | 64.13 | 20.27 | 44.97 | 77.41 | 89.16 | 96.70 | **100** |
| OCSVM [33] | 1.85 | 2.56 | 5.75 | 11.04 | 22.99 | 33.12 | 16.66 | 16.69 | 17.12 | 20.68 | 26.03 | 36.62 |
| KISSME [17] | 13.40 | 23.60 | 33.96 | 43.45 | 54.47 | 64.25 | 16.93 | 29.97 | 68.92 | 79.80 | 93.50 | 98.73 |
| LMNN [42] | 13.78 | 23.01 | 36.50 | 43.65 | 54.69 | 63.22 | 17.11 | 21.98 | 41.73 | 55.23 | 75.51 | 84.26 |
| LDM [44] | 9.48 | 17.65 | 29.86 | 39.72 | 53.96 | 62.74 | 16.76 | 18.54 | 33.18 | 50.81 | 68.42 | 81.82 |
| LADF [22] | 6.04 | 17.63 | 38.19 | 49.72 | 63.46 | 74.8 | 18.59 | 27.43 | 68.40 | 84.37 | **99.41** | **100** |
| LFDA [29] | 9.39 | 16.15 | 27.20 | 36.37 | 47.15 | 56.54 | 16.66 | 20.54 | 28.65 | 41.42 | 56.89 | 67.06 |
| Salience [46] | 13.45 | 24.05 | 34.99 | 43.53 | 52.63 | 59.45 | 16.67 | 16.84 | 17.81 | 19.03 | 25.46 | 35.32 |
| L1-norm | 13.57 | 24.27 | 35.95 | 44.93 | 53.54 | 62.83 | 16.96 | 21.13 | 38.11 | 46.94 | 63.45 | 76.55 |

TABLE I

ONE-SHOT INDIVIDUAL VERIFICATION RESULTS: TRUE TARGET RATE (TTR) IN % AGAINST FALSE TARGET RATE (FTR).

people and the gallery set has multiple images per target person. Consequently these models are typically shown to have a big improvement over non-learning based approach in previous work.

Note that in general much inferior results are obtained using the compared local distance/subspace learning models including LMNN, LDM, LADF and LFDA, especially when the FTR is low. The main reason is that LMNN, LDM and LFDA all compute the neighbourhood of each data point using the Euclidean metric and keep a constant neighbourhood fixed throughout the whole training process. Although LADF has a local decision boundary, no neighbourhood is defined to focus the learning on local data variations. In comparison, t-LRDC updates the neighbourhood at each iteration to enable the neighbourhood to be computed much more accurately. In addition, the local modelling of t-LRDC is very different from that in LMNN, LDM and LFDA, as it can be viewed that t-LRDC defines the locality between similar distances rather than between similar data points. It is interesting to notice that LADF is competitive when FTR is high. However, its performance is weak for low FTR values, sometimes weaker than the non-learning based L1-norm.

The relative comparison based naive transfer models, RDC [50] and RankSVM [31] are more competitive. Their results are thus tabulated in Tables I and II to show more details. It is evident that on most datasets and for both the individual and set verification tasks, our t-LRDC model's performance is overall superior to that of RDC and RankSVM. In addition, t-LRDC also spends much less memory cost as shown in Table III.

### C. Comparison with Generalisations of Naive Transfer Models

In Sec. III-G, the two relative comparison based naive transfer models RDC [50] and RankSVM [31] are generalised to utilise the same three types of knowledge transfer as our t-LRDC in order to tackle the group-based verification problem. The resultant t-RDC and t-RankSVM models are compared against our t-LRDC in Tables I and II. As mentioned in Sec. III-G, the main differences between t-LRDC and these generalisations are (1) local modelling for computational efficiency and the removal of hard-to-satisfy and overfitting constraints outside a local neighbourhood, and (2) the use of absolute difference in these methods. The results show that following the same transfer learning formulation, overall both RDC and RankSVM benefit from the knowledge transfer leading to improvement in performance. In particular, among all methods compared, t-RDC and t-RankSVM together with t-LRDC almost always achieve the best verification results as shown in Tables I and II. But among them, overall t-LRDC has the best performance. The reason why occasionally t-LRDC is sometimes inferior to t-RDC (e.g. on ETHZ) is mainly because t-LRDC needs to select a subset of distance comparison pairs. This selection could sometimes introduce errors leading to inferior performance compared to the global approach. The proposed t-LRDC is also computationally more efficient. Specifically, local modelling is hard to perform for t-RDC and t-RankSVM as analysed in Sec. III-G; importantly, no PCA-based dimensionality reduction can be used as pre-processing. As a result, they have much greater computational cost compared to t-LRDC, especially

| Database | i-LIDS | | | | | | ETHZ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FTR | 0.1% | 1% | 5% | 10% | 20% | 30% | 0.1% | 1% | 5% | 10% | 20% | 30% |
| t-LRDC | 9.65 | 18.75 | **33.55** | **43.18** | **51.75** | **63.36** | 36.63 | 47.90 | 62.81 | 71.38 | 80.42 | 85.43 |
| t-LRDC(Global) | 8.10 | 16.97 | 32.62 | 39.43 | 48.26 | 58.69 | 31.34 | 46.34 | 60.84 | 67.28 | 76.17 | 81.40 |
| t-RDC | **10.82** | **20.73** | 32.24 | 37.70 | 48.73 | 58.08 | **38.97** | **59.66** | **74.86** | **81.48** | **86.84** | 90.25 |
| t-RankSVM | 8.82 | 16.29 | 26.73 | 34.18 | 46.86 | 56.92 | 33.24 | 57.10 | 72.82 | 80.10 | 86.54 | 90.19 |
| t-RDC-PCA | 6.98 | 13.55 | 25.37 | 34.18 | 44.49 | 53.78 | 32.38 | 46.48 | 59.61 | 68.21 | 76.04 | 80.81 |
| t-RankSVM-PCA | 7.19 | 10.33 | 18.63 | 24.98 | 42.94 | 54.04 | 22.59 | 38.75 | 54.60 | 61.99 | 72.16 | 78.94 |
| RDC [50] | 7.72 | 17.32 | 28.63 | 38.13 | 47.73 | 58.63 | 38.76 | 57.64 | 73.79 | 80.76 | 86.67 | 90.18 |
| RankSVM [31] | 7.20 | 14.48 | 23.40 | 31.99 | 47.57 | 59.40 | 31.09 | 53.63 | 70.81 | 78.88 | 85.13 | **90.65** |
| OCSVM [33] | 6.02 | 7.05 | 13.27 | 18.22 | 29.28 | 36.44 | 1.01 | 3.34 | 12.89 | 18.95 | 28.48 | 35.56 |
| KISSME [17] | 8.88 | 14.68 | 26.83 | 35.23 | 41.36 | 48.91 | 34.83 | 49.35 | 59.94 | 68.16 | 76.87 | 84.63 |
| LMNN [42] | 6.84 | 10.03 | 21.88 | 32.83 | 46.00 | 54.15 | 33.61 | 43.91 | 58.51 | 66.56 | 76.21 | 82.81 |
| LDM [44] | 7.13 | 10.12 | 19.87 | 25.30 | 41.92 | 56.19 | 21.58 | 32.47 | 49.26 | 58.39 | 69.34 | 77.26 |
| LADF [22] | 7.25 | 11.66 | 23.24 | 31.35 | 44.68 | 56.88 | 10.23 | 26.88 | 51.49 | 61.75 | 73.12 | 81.26 |
| LFDA [29] | 6.59 | 8.51 | 15.73 | 21.28 | 30.28 | 42.29 | 19.80 | 31.42 | 44.64 | 52.37 | 63.01 | 69.96 |
| Salience [46] | 6.00 | 6.08 | 7.11 | 10.52 | 17.55 | 26.41 | 16.51 | 35.92 | 53.25 | 61.25 | 73.00 | 83.20 |
| L1-norm | 7.19 | 9.58 | 18.92 | 30.54 | 48.17 | 57.71 | 31.39 | 46.06 | 60.13 | 66.88 | 77.15 | 83.31 |
| Database | CAVIAR | | | | | | VIPeR | | | | | |
| FTR | 0.1% | 1% | 5% | 10% | 20% | 30% | 0.1% | 1% | 5% | 10% | 20% | 30% |
| t-LRDC | **11.65** | 16.74 | **29.92** | **36.72** | **47.53** | 58.04 | 16.66 | **27.43** | **45.99** | **63.27** | **75.11** | **88.62** |
| t-LRDC(Global) | 11.09 | 15.40 | 26.37 | 34.24 | 45.82 | 56.67 | 18.48 | 20.63 | 37.20 | 54.26 | 67.42 | 78.49 |
| t-RDC | 9.44 | **16.89** | 26.50 | 34.51 | 46.22 | 56.44 | 17.11 | 21.20 | 38.08 | 50.97 | 72.28 | 79.22 |
| t-RankSVM | 5.45 | 13.94 | 23.08 | 29.48 | 41.89 | 53.57 | **18.96** | 23.83 | 42.49 | 58.38 | 72.82 | 83.62 |
| t-RDC-PCA | 9.00 | 13.87 | 24.65 | 33.90 | 46.28 | 56.70 | 16.66 | 19.21 | 24.55 | 32.75 | 49.96 | 58.70 |
| t-RankSVM-PCA | 8.31 | 15.20 | 22.69 | 29.48 | 39.72 | 51.48 | 16.66 | 18.60 | 22.20 | 27.08 | 39.67 | 53.08 |
| RDC [50] | 10.18 | 16.66 | 26.89 | 34.83 | 47.22 | **58.07** | 16.79 | 22.57 | 41.24 | 56.02 | 69.29 | 83.11 |
| RankSVM [31] | 3.35 | 10.14 | 20.20 | 28.16 | 42.42 | 53.56 | 16.92 | 23.17 | 42.45 | 57.36 | 72.87 | 82.21 |
| OCSVM [33] | 2.17 | 2.75 | 6.06 | 11.31 | 23.60 | 33.42 | 16.66 | 16.70 | 17.13 | 20.85 | 26.07 | 36.72 |
| KISSME [17] | 9.36 | 16.39 | 25.41 | 32.35 | 41.59 | 50.70 | 16.68 | 20.24 | 30.37 | 52.22 | 74.25 | 83.83 |
| LMNN [42] | 9.50 | 15.15 | 25.49 | 34.00 | 46.61 | 55.62 | 16.76 | 17.62 | 21.93 | 31.96 | 52.70 | 62.87 |
| LDM [44] | 6.39 | 11.28 | 19.12 | 27.56 | 39.55 | 49.93 | 16.66 | 17.53 | 23.17 | 30.31 | 46.06 | 62.19 |
| LADF [22] | 4.0 | 8.75 | 19.33 | 28.68 | 43.52 | 51.99 | 17.21 | 18.92 | 26.25 | 44.35 | 65.93 | 82.37 |
| LFDA [29] | 7.73 | 11.72 | 20.26 | 26.51 | 36.91 | 48.48 | 16.66 | 16.77 | 23.00 | 31.09 | 44.12 | 51.28 |
| Salience [46] | 10.13 | 15.15 | 25.58 | 32.74 | 44.89 | 52.73 | 16.67 | 16.73 | 17.44 | 18.54 | 21.60 | 25.88 |
| L1-norm | 10.48 | 15.58 | 26.38 | 34.55 | 45.12 | 54.87 | 16.72 | 17.24 | 20.81 | 33.80 | 48.20 | 61.58 |

TABLE II

ONE-SHOT SET VERIFICATION RESULTS: TRUE TARGET RATE (TTR) IN % AGAINST FALSE TARGET RATE (FTR).

| | t-LRDC | t-RDC | t-RankSVM | RDC | RankSVM |
|---|---|---|---|---|---|
| Sensitive to PCA | × | √ | √ | √ | √ |
| Max Memory Cost | ∼0.7 G | ∼16.9G | ∼16.9G | ∼16.1G | ∼16.1G |

TABLE III

COST COMPARISON: RELATIVE COMPARISON LEARNING ON VIPER

in terms of memory usage. Table III shows that taking VIPeR for example, t-LRDC consumes 0.7G of memory space at maximum, while t-RDC needs about 16.9G at maximum. Tables I and II also show that if the same dimension reduction is applied (t-RDC-PCA and t-RankSVM-PCA) their performance dropped by big margins. This is because with absolute difference, Theorem 2 does not apply anymore and there is no guarantee that the performance will not degrade as in the case of t-LRDC[3].

### *D. Further Evaluations on t-LRDC*

**Effectiveness of the three types of knowledge transfer**. In t-LRDC, three different types of knowledge transfer are conducted, resulting in three different types of constraints (see Sec. III-B). The relative weights of the different constraints are controlled by parameters $\alpha$ and $\beta$ in Eq. (20). In order to evaluate how much different types of constraints contribute to the final performance, we vary the value of $\alpha$ and $\beta$ and report the results in Tables IV and V. The results show that all three types of constraints are useful. For example, if we set $\alpha = 0$, only the first type is used; the results are clearly inferior to those when all three are used. Overall, the performance of t-LRDC is not sensitive to the weightings when they are not set to the extreme values.

**Sensitivity to the non-target selection parameter** $h$. The parameter $h$ in Eq. (3) controls how many non-target people are considered to be similar to one of the target person therefore used for enriching both the intra and inter-class variations. Its effect is shown in Table VI. It can be seen that the performance is not sensitive as long as the value is not extreme. Note that when $h = 1$, it means no similar non-target people are selected, and all non-target people are used for the third type of knowledge transfer. This clearly leads to poor performance since no knowledge can be transferred to enrich the target intra and inter-class variations [4].

**Effectiveness of local modelling: neighbourhood size** $k$. We also vary the neighbourhood parameter $k$ (Eq. (16)) which determines how big the neighbourhood is for local modelling. It can be observed in Table VII that, for local relative comparison, good results are obtained when we set $k = 2$ or 3. In particular, when all neighbour are used (i.e. $k = Inf$), it is the t-LRDC(Global) shown in Tables I and II. For both individual and set person verification, t-LRDC always performs better than t-LRDC(Global). The superiority is much clearer when FTR is low, for example 0.1% and 1%. The experimental results justify the use of local relative distance comparison as discussed in Sec. III-C. In addition, with local modelling (when $k = 3$) the training time is reduced by around 50% as compared to using all constraints exhaustively.

[3]The results of t-LRDC without PCA are the same as t-LRDC with PCA, since the solution of t-LRDC is defined in the PCA space by Theorem 2. The results are thus not included.

[4]For more evaluation on the performance of pairwise parameters among $h$, $\alpha$ and $\beta$, please refer to the supplementary material.

| | Individual Verification | | | | | | | | | | | Set Verification | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| i-LIDS | 20.71 | 32.99 | 29.96 | 31.71 | 32.10 | **33.24** | 31.69 | 33.08 | 32.03 | 32.97 | 23.57 | 12.93 | 17.39 | 17.42 | 16.76 | 18.24 | 17.33 | 18.12 | 17.67 | **18.75** | 18.21 | 12.39 |
| ETHZ | 49.30 | 56.53 | 57.54 | 58.87 | 61.56 | 59.09 | 61.34 | **66.72** | 65.62 | 66.08 | 63.42 | 31.76 | 39.12 | 41.09 | 42.90 | 43.93 | 43.61 | 44.96 | **49.37** | 47.90 | 48.10 | 47.86 |
| CAVIAR | 21.72 | 24.84 | 26.65 | 24.19 | 24.41 | 27.76 | 28.02 | 27.22 | 28.13 | **28.69** | 24.22 | 13.42 | 15.50 | 15.37 | 14.60 | 15.64 | 16.03 | 15.66 | 15.94 | 16.74 | **17.36** | 15.41 |
| VIPeR | 22.80 | 22.94 | 29.19 | 30.71 | 32.57 | 38.55 | 43.26 | **47.48** | 47.27 | 45.62 | 46.46 | 16.70 | 16.88 | 16.88 | 17.32 | 17.44 | 17.81 | 23.31 | 23.43 | **27.43** | 23.40 | 22.93 |

TABLE IV

EFFECTS OF $\alpha$ ON T-LRDC (EQ. (20)): TRUE TARGET RATE (%) WHEN FTR = 1%.

| | Individual Verification | | | | | | | | | | | Set Verification | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| i-LIDS | 29.66 | 31.86 | 32.72 | 30.75 | 30.20 | 32.28 | 32.03 | **33.76** | 33.45 | 32.26 | 31.92 | 19.35 | 18.95 | **19.97** | 19.17 | 19.70 | 18.63 | 18.75 | 19.57 | 19.02 | 18.00 | 18.80 |
| ETHZ | 64.60 | 67.25 | 67.33 | 67.31 | **67.40** | 66.12 | 65.62 | 64.33 | 63.78 | 65.20 | 64.91 | 49.19 | 50.19 | **50.63** | 50.23 | 50.43 | 48.46 | 47.90 | 47.17 | 47.02 | 47.31 | 47.20 |
| CAVIAR | 27.86 | 28.43 | 28.29 | **29.45** | 29.20 | 28.15 | 28.13 | 27.60 | 27.99 | 26.77 | 27.09 | 16.35 | 15.99 | 15.43 | 16.11 | 16.35 | 15.69 | 16.74 | 16.77 | 17.24 | 16.96 | **17.38** |
| VIPeR | 43.66 | 45.56 | 46.02 | 46.88 | 46.81 | **47.32** | 47.27 | 46.61 | 47.01 | 47.03 | 46.39 | 22.24 | 23.48 | 23.74 | 24.55 | 26.35 | 26.46 | **27.43** | 27.06 | 27.43 | 26.70 | 26.21 |

TABLE V

EFFECTS OF $\beta$ ON T-LRDC (EQ. (20)): TRUE TARGET RATE (%) WHEN FTR = 1%.

| | Individual Verification | | | | | | | | | Set Verification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | 0 | 0.3 | 0.5 | 0.7 | 0.72 | 0.75 | 0.77 | 0.9 | 1 | 0 | 0.3 | 0.5 | 0.7 | 0.72 | 0.75 | 0.77 | 0.9 | 1 |
| i-LIDS | 29.08 | 29.08 | 28.03 | 32.01 | 32.03 | **33.34** | 27.38 | 15.83 | 11.23 | 16.58 | 16.58 | 16.05 | **19.45** | 18.75 | 17.42 | 17.48 | 10.35 | 9.66 |
| ETHZ | 64.46 | 64.46 | **65.63** | 66.53 | 65.62 | 64.29 | 65.44 | 52.56 | 37.49 | 49.21 | 49.21 | **49.55** | 48.45 | 47.90 | 46.92 | 46.49 | 39.67 | 36.93 |
| CAVIAR | 27.42 | 27.59 | 26.64 | 28.02 | **28.13** | 26.00 | 25.13 | 16.75 | 14.54 | 15.59 | 15.59 | 14.56 | 15.69 | **16.74** | 16.08 | 16.52 | 14.24 | 14.01 |
| VIPeR | 42.46 | 42.46 | 45.67 | 46.79 | 47.27 | **49.47** | 47.82 | 20.26 | 17.91 | 21.79 | 21.79 | 22.54 | 26.94 | 27.43 | **29.01** | 28.68 | 17.71 | 17.76 |

TABLE VI

EFFECTS OF THE SIMILARITY THRESHOLD $h$ (EQ. (3)) IN T-LRDC: TRUE TARGET RATE (%) WHEN FTR = 1%.

| | Individual Verification | | | | | | Set Verification | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | 1 | 2 | 3 | 4 | 5 | Inf | 1 | 2 | 3 | 4 | 5 | Inf |
| i-LIDS | 32.98 | **34.53** | 32.03 | 31.39 | 30.71 | 30.94 | **20.25** | 18.18 | 18.75 | 20.08 | 18.08 | 16.97 |
| ETHZ | **65.73** | 65.70 | 65.62 | 64.73 | 64.80 | 62.72 | 48.25 | 48.27 | 47.90 | 47.91 | **48.48** | 46.34 |
| CAVIAR | 28.08 | 27.52 | **28.13** | 27.74 | 25.10 | 25.85 | 15.80 | 16.55 | **16.74** | 15.87 | 15.48 | 15.40 |
| VIPeR | 45.94 | 48.29 | 47.27 | 46.31 | **49.18** | 39.04 | 22.59 | 21.68 | **27.43** | 24.55 | 24.54 | 20.63 |

TABLE VII

EFFECTS OF THE LOCAL MODELLING NEIGHBOURHOOD SIZE $k$ (EQ. (16)) IN T-LRDC: TRUE TARGET RATE (%) WHEN FTR = 1%.

**Set verification vs. individual verification**. It is clear from all the results reported that individual verification always achieved higher values as compared to the one for set verification. However, as analysed in Sec. III-F, individual verification is an one-to-one verification and cannot make a joint verification explicitly to say whether the person of a query image is one of the several people on the watch-list. It is similar to the case where the value of a joint probability density function is always lower than that of each marginal probability density function.

## V. CONCLUSIONS

We have re-formulated the person re-identification problem as an one-shot group-based verification problem to meet the requirement of more realistic real-world applications. To the best of our knowledge, it is the first attempt on addressing the person re-identification problem under this challenging setting. Since there are always limited images for the people on the watch-list, a transfer relative comparison framework is proposed to utilise non-target person images to assist the verification of target people. Three different types of knowledge transfer are exploited resulting in three different types of relative comparison constraints. Both global and local relative comparison models are proposed. In particular, the local relative comparison approach has been verified as a more efficient approach, as well as having the potential to avoid model over-fitting for better verification performance. A gradient decent based efficient optimisation algorithm is formulated which can further improve the model tractability by utilising dimensionality reduction as a pre-processing step without loss of performance.

In this work, the proposed transfer learning method is based on relative distance comparison. We believe that the model can be further improved by incorporating more advanced features and techniques such as the one based on finding salient patches [46]. Also, currently we assume that both target and non-target data are collected from the same or similar environment. It would be an interesting and also more challenging problem to perform transfer learning across datasets captured from different camera networks. Since target and source data are captured in distinctly different environments, the cross-camera view illumination and pose variations on the target environment cannot be learned from source data and thus it becomes much more challenging to extract transferrable transformation from source data. Our ongoing work includes exploring some commonly used features across datasets and considering extending our model by combining it with some recently proposed cross-domain/dataset transfer learning methods [19], [43].

## ACKNOWLEDGMENTS

This work was supported partially by Guangdong Provincial Government of China through the Computational Science Innovative Research Team Program, the National Natural Science of Foundation of China (Nos. 61472456, U1135001), the 12th Five-year Plan China S&T Supporting Programme (No. 2012BAK16B06), Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, and in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265. S. Gong and T. Xiang are supported by the UK Home Office CONTEST Programme and Vision Semantics Limited.

## REFERENCES

[1] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *IEEE Computer Vision and Pattern Recognition*, 2005.

[3] L. Bottou, O. Bousquet, and G. Zrich. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, 2008.

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*, 2011. http://dx.doi.org/10.5244/C.25.68.

[5] P. Dollar, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[6] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[7] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *IEEE International Conference on Computer Vision*, 2007.

[8] M. Farenzena, L. Bazzani, A. Perina, M. Cristani, and V. Murino. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[9] B. Geng, D. Tao, and C. Xu. Daml: Domain adaptation metric learning. *IEEE Transactions on Image Processing*, 20(10):2980–2989, 2011.

[10] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[11] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, 2005.

[12] S. Gong, M. Cristani, C. C. Loy, and T. Hospedales. The re-identification challenge. In Gong, Cristani, Yan, and Loy, editors, *Person Re-Identification*, pages 1–21. Springer, 2014.

[13] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International workshop on performance evaluation of tracking and surveillance*, 2007.

[14] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008.

[15] M. Hirzer, P. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, 2012.

[16] W. Hu, M. Hu, X. Zhou, J. Lou, T. Tan, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):663–671, 2006.

[17] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[18] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.

[19] R. Layne, T. M. Hospedales, and S. Gong. Domain transfer for person re-identification. In *ACM International Conference on Multimedia, Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams, Barcelona*, 2013.

[20] F. F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligent*, 28(4):594–611, 2006.

[21] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, 2013.

[22] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[23] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *International Conference on Image Processing*, 2013.

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

[25] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[26] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *AAAI Conference on Artificial Intelligence*, 2008.

[27] S. Parameswaran and K. Q. Weinberger. Large margin multi-task metric learning. *Advances in neural information processing systems*, 23, 2010.

[28] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. In *International Conference on Pattern Recognition*, 2006.

[29] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[30] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *British Machine Vision Conference*, 2008.

[31] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010.

[32] R. Salakhutdinov, J. Tenenbaum, and A. Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *Workshop on Unsupervised and Transfer Learning*, 2012.

[33] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[34] W. Schwartz and L. Davis. Learning discriminative appearance based models using partial least squares. In *Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[35] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, 2007.

[36] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li. Person re-identification by regularized smoothing kiss metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10):1675–1685, 2013.

[37] T. Tommasi and B. Caputo. The more you know, the less you learn: From knowledge transfer to one-shot learning of object categories. In *British Machine Vision Conference*, 2009.

[38] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *IEEE Computer Vision and Pattern Recognition*, 2004.

[39] UK. Home Office i-LIDS multiple camera tracking scenario definition. 2008.

[40] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *IEEE International Conference on Computer Vision*, 2007.

[41] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1222–1228, 2004.

[42] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, 2006.

[43] Y. Wu, W. Li, M. Minoh, and M. Mukunoki. Can feature-based inductive transfer learning help person re-identification? In *International Conference on Image Processing*, 2013.

[44] L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *AAAI Conference on Artificial Intelligence*, 2006.

[45] Y. Zhang and D. Yeung. A convex formulation for learning task relationships in multi-task learning. In *Conference on Uncertainty in AI*, 2010.

[46] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[47] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *British Machine Vision Conference*, 2009.

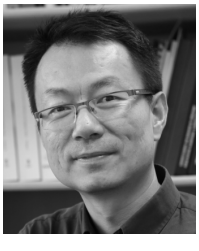
[48] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[49] W.-S. Zheng, S. Gong, and T. Xiang. Transfer re-identification: From person to set-based verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[50] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.


**Wei-Shi Zheng** is an Associate Professor of Sun Yat-sen University. Prior to that, he received his Ph.D. degree in applied mathematics from Sun Yat-Sen University in 2008, and has been a Postdoctoral Researcher on the EU FP7 SAMURAI Project at Queen Mary University of London. His research interests include object association and categorization in visual surveillance.

**Shaogang Gong** Shaogang Gong is Professor of Visual Computation at Queen Mary University of London, a Fellow of the Institution of Electrical Engineers and a Fellow of the British Computer Society. He received his D.Phil in computer vision from Keble College, Oxford University in 1989. His research interests include computer vision, machine learning and video analysis.



**Tao Xiang** received the PhD degree in electrical and computer engineering from the National University of Singapore in 2002. He is currently a Reader (Associate Professor) in the School of Electronic Engineering and Computer Science, Queen Mary University of London. His research interests include computer vision, machine learning, and data mining. He has published over 100 papers in international journals and conferences and co-authored a book, Visual Analysis of Behaviour: From Pixels to Semantics.