Consistent-Aware Deep Learning for Person Re-identification in a Camera Network

Ji Lin[1,4], Liangliang Ren[1,2,3], Jiwen Lu[1,2,3]*, Jianjiang Feng[1,2,3], Jie Zhou[1,2,3]
[1]Department of Automation, Tsinghua University, Beijing, China
[2]State Key Lab of Intelligent Technologies and Systems, Beijing, China
[3]Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, China
[4]Department of Electronic Engineering, Tsinghua University, Beijing, China
lin-j14@mails.tsinghua.edu.cn; renll16@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn;
jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn

Abstract

In this paper, we propose a consistent-aware deep learning (CADL) approach for person re-identification in a camera network. Unlike most existing person re-identification methods which identify whether two pedestrian images are from the same person or not, our approach aims to obtain the maximal correct matches for the whole camera network. Different from recently proposed camera network based re-identification methods which only consider the consistent information in the matching stage to obtain a globally optimal association, we exploit such consistent-aware information under a deep learning framework where both feature representation and image matching are automatically learned. Specifically, we reach the globally optimal solution and balance the performance between different cameras by optimizing the similarity and data association iteratively with certain consistent constraints. Experimental results show that our method obtains significant performance improvement and outperforms the state-of-the-art methods by large margins.

1. Introduction

Person re-identification aims to match pedestrians across multiple camera views. It is a challenging problem due to the changes of scale, illumination, viewing angle, pose, *etc*., and has numerous application including visual surveillance, robotics, multimedia and forensics.

Most current approaches focus on pairwise re-identification, which distinguish whether two images captured from different cameras are from the same person or not. Existing person re-identification methods can be roughly divided into two categories: image-based and video-based. Image-based methods [19, 24, 36, 40, 41, 45–47] focus on seeking effective feature descriptors which are robust to the changes of light, pose, viewing angle, *etc*., or designing discriminative similarity metrics for person matching. Video-based methods [13, 23] focus on promising video modeling and matching techniques to reduce the influences of occlusion and illumination changes.

One key application of pairwise re-identification is visu-

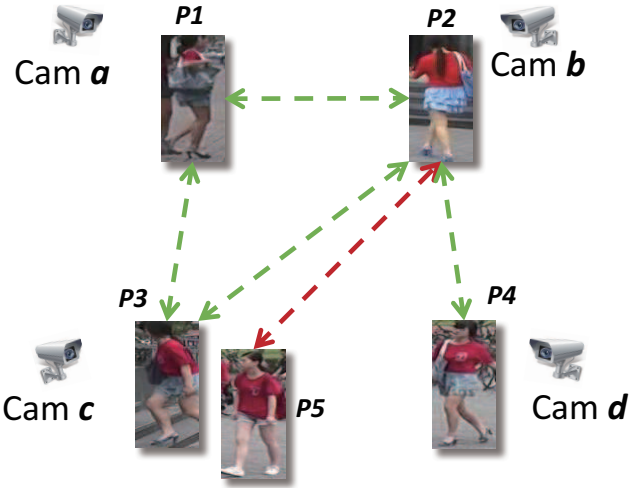

Figure 1. An illustration of person re-identification in a camera network. The green lines refers to correct matches and the red lines indicates wrong matches. If P1 is considered to be the same person as both P2 and P3, then P2 and P3 must be the same person. Otherwise, the inconsistency will arise. This constraints results in the upper left green triangle. Similarly, P2 and P5 cannot be considered as the same person, as P3 and P5 are different persons in the same camera view (best viewed in the color pdf file).

*Corresponding author.

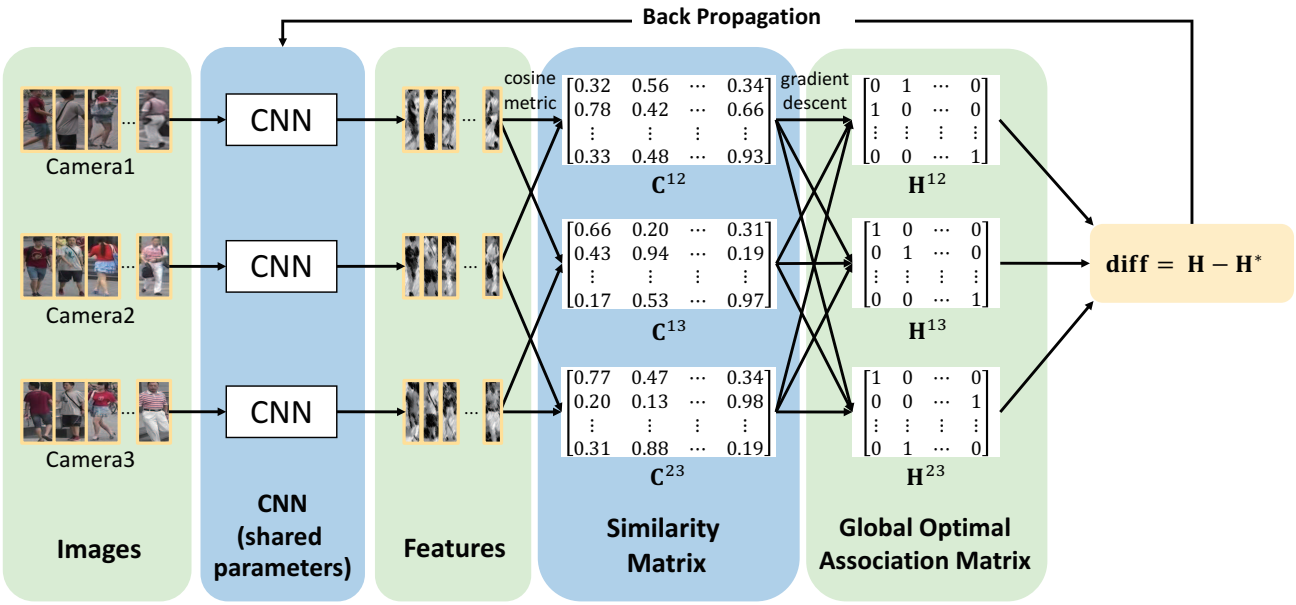

Figure 2. The overall framework of our approach. We consider the simplest situation where there are 3 cameras in the camera network. First, we take a batch of persons as the input and feed them into the CNN deep network. Then, we use the cosine similarity to obtain the similarity matrix $\mathbf{C}$. By using a specially designed gradient descent method, we obtain the globally optimal association matrices $\mathbf{H}$. We further compute the difference as $\mathbf{H} - \mathbf{H}^*$, and propagate it back to deep model to update the CNN features.

al surveillance in a camera network, which aims to identify persons across multiple different cameras. While the re-identification performance is encouraging for a pair of cameras, the overall performance across the whole camera network is still far from satisfactory and the inconsistency information usually occurs. Figure 1 illustrates the difference between the pairwise re-identification and the camera network re-identification. Assume there are three cameras (a, b and c) in the network, person $P1$ in camera a is matched to person $P2$ in camera b and person $P3$ in camera c, then $P2$ and $P3$ must be considered as the same person. However, the re-identification system may recognize that $P2$ and $P3$ are not the same person according to their appearance information from camera b and c directly, which bring an inconsistency for this camera network. In other words, pairwise re-identification methods cannot obtain the globally optimal matching results for a whole camera network. To address this, Das *et al.* [3,5] proposed a camera network based re-identification approach which exploit such consistent information in camera network. However, the consistent information was only exploited in their matching part and not utilized in the training stage.

In this paper, we proposed a consistent-aware deep learning (CADL) approach for person re-identification in a camera network, where the whole overall framework is shown in Figure 2. Unlike conventional deep learning based person re-identification methods which employ a large number of labeled samples to train the deep model for feature extraction, our approach aims to seek the globally optimal matching for the whole network. Specifically, we used a gradient descent algorithm to seek the globally optimal matching by maximizing the sum of all matching similarity for all camera pairs, while satisfying all the consistent constraints simultaneously. After the globally optimal solution was obtained, we propagated the difference to adjust the deep neural network. Unlike other methods [1, 17, 34, 37] aiming for local optimum, CADL aims for the global optimum and thus can improve the overall performance. Experimental results on three datasets including the Market-1501, WARD, and RAiD are presented to show the effectiveness of the proposed approach.

2. Related Work

Person Re-identification: Recently, numerous methods have been proposed for person re-identification from two aspects: image-based [19, 24, 36, 40, 41, 45–47] and video-based [13, 23]. Image-based methods focus on seeking discriminative feature descriptors and metrics for pedestrian matching. Representative features in person re-identification include color histograms [19,40,45,46], color names [41, 47], local binary patterns (LBP) [14, 40], scale invariant feature transform [22, 45, 46] and scale invariant local ternary patterns [19,21]. Typical metric learning methods for person re-identification are locally adaptive decision functions (LADF) [18], cross-view quadratic discriminant analysis (XQDA) [19], probabilistic relative distance

comparison (PRDC) [49], metric learning with accelerated proximal gradient (MLAPG) [20], local fisher discriminant analysis (LFDA) [26] and its kernel variant (k-LFDA) [39]. Video-based methods focus on effective video modeling and matching techniques to reduce the influences of occlusion and illumination changes. Representative methods in this category include conditional random field, space-time feature description [2], video ranking function [38], and top-push constrained matching [43].

Deep Learning: Deep learning has achieved great successes on various computer vision applications such as image classification [11, 15, 30, 32], object detection [7, 8, 10, 27, 28], face recognition [29, 31, 33], *etc*. There has also been growing number of methods which apply deep learning for person re-identification in recent years. For example, Yi *et al*. [42] proposed a siamese CNN (S-CNN) deep architecture for person re-identification, where three S-CNNs were employed for deep feature learning. Ahmed *et al*. proposed a cross-input neighborhood difference method [1] to extract the cross-view relationships of the features. Li *et al*. proposed a deep filter pairing neural network (FPNN) [17] to jointly handle misalignment, photometric and geometric transforms, occlusions and background clutter. Cheng *et al*. proposed a framework to deal with local body-parts based features and the global features [4]. More recently, Wang *et al*. [37] proposed a framework containing one shared sub-network together with two sub-networks that extract single image and cross image representations respectively. Xiao *et al*. proposed a domain guided drop out (DGD) method [39] to improve feature learning by selecting the neurons specific to certain domains. Varior *et al*. [35] proposed a long short term memory (LSTM) method to process image regions sequentially and enhance the discriminative capability of local feature representation by leveraging contextual information. Varior *et al*. [34] proposed a gated Siamese CNN architecture to selectively emphasize fine common local patterns by comparing the mid-level features across pairs of images.

3. Approach

3.1. Problem Formulation

Assume there are m cameras in the camera network. The number of possible camera pairs is $\binom{m}{2} = \frac{m(m-1)}{2}$. For simplicity, we first assume that the same n persons are present in each camera of the networks. We propose our framework by making all correct matches for all persons in the whole camera network.

Matrix Representation: Two types of matrices are used in our framework, where the superscript refers to camera ID, and the subscript refers to person ID.

Similarity score matrix: Let $\mathbf{C}^{a,b}$ denote the similarity score matrix between camera a and b. $\mathbf{C}$ is an n by n matrix, with row index representing the persons from camera a, and the column index representing the persons from camera b. Then the $(i,j)^{th}$ cell in $\mathbf{C}^{a,b}$ denotes the similarity score between person P_i^a and P_j^b.

Adjacency Matrix: Let $\mathbf{H}$ denote the adjacency matrix to represent the relationship between persons. $\mathbf{H}$ is an n by n matrix, with row index representing the persons from camera a, and the column index representing the persons from camera b. The elements $h_{i,j}^{a,b}$ of adjacency matrix $\mathbf{H}^{a,b}$ between camera a and b is either 0 or 1. If $h_{i,j}^{a,b} = 1$, we assume P_i^a and P_j^b are the same person. Otherwise, they represent different persons.

Globally Optimal Objective: As mentioned above, our framework aims to obtain the globally optimal match by maximizing the sum of similarities for all cameras. As two person images are considered as the same person only when the corresponding elements in $\mathbf{H}$ is 1. Therefore we formulate the following objective function:

$$\begin{aligned} \max_{\mathbf{H}} \mathbf{Sim} \quad &= \sum_{a,b=1,a<b}^{m} \mathbf{C}^{a,b} \cdot \mathbf{H}^{a,b} \\ \text{subject to:} \quad & \mathbf{H}_{i,j}^{a,b} \in \{0,1\}, \sum_{i=1}^{n} \mathbf{H}_{i,j}^{a,b} = 1, \\ & \sum_{j=1}^{n} \mathbf{H}_{i,j}^{a,b} = 1, \mathbf{H}_{i,k}^{a,c} \mathbf{H}_{k,j}^{c,b} \leq \mathbf{H}_{i,j}^{a,b}, \\ & \forall a,b,c = 1,2,...,m, a < b \end{aligned} \tag{1}$$

Notice that $a < b$ might be eliminated if we use probe and gallery setting, such as the Market-1501 dataset.

Constraints: We first consider the simplest situation where all persons appear in all cameras, so that the re-identification in the network can be modeled as a one-to-one match problem. Since each person can only be matched exactly to one person with no crossing, i.e. in the adjacency matrix, there should be exactly one 1 in each row and column. We present an example in Figure 1. When matching $P2$ in camera b to persons in camera c, we cannot match $P2$ to both $P3$ and $P5$ as $P3$ and $P5$ in the same camera are not the same person. We also need to keep the inter-camera consistency. As shown in Figure 1, if person $P1$ in camera a is considered to be the same person with both person $P2$ in camera b and person $P3$ in camera c, then person $P2$ and person $P3$ should be considered as the same person, resulting in the upper left green triangle. Otherwise, there will be a contradiction. For the case where there are more than 3 cameras, it can be easily demonstrated that we can guarantee the inter-camera consistency by checking all the triplet camera pairs, *e.g*. we keep consistency in a camera network which consists of camera $\{a,b,c,d\}$ by checking the consistency in $\{a,b,c\}, \{a,b,d\}, \{a,c,d\}$ and $\{b,c,d\}$. If $P1$ and $P2$, $P1$ and $P3$, $P2$ and $P4$ in Figure 1 are considered as the same person, we first re-identify $P2$ and $P3$ as the

same person, and then make sure that $P3$ and $P4$ are the same person.

To guarantee such consistency, we make certain constraints on the adjacency matrix $\mathbf{H}$. The constraints are described as follows:

1. *Binary Constraint*. While $\mathbf{H}$ should be theoretically binary, we first continualize $\mathbf{H}$ and iteratively adjust its value. We impose a binary constraint on $\mathbf{H}$, which is to help $\mathbf{H}$ approximate 0 or 1. It imposes larger penalty on elements far from 0 or 1, which is computed as:

$$\mathbf{J_B} = \|(\mathbf{H} - 0.5) \cdot (\mathbf{H} - 0.5) - 0.25\|_F^2 \tag{2}$$

2. *Row and Column Constraints*. As all n persons appear in all m cameras, the ground-truth $\mathbf{H}^*$ should have exactly one positive value every row and every column, due to one-to-one match. Take $\mathbf{e}$ as a n-length column vector full of ones, that is $\mathbf{e} = [1, 1, \cdots, 1]^\mathbf{T}$, we introduce row and column constraints as below:

$$\mathbf{J_R} = \|\mathbf{He} - \mathbf{e}\|_2^2, \mathbf{J_C} = \|\mathbf{e^T H} - \mathbf{e^T}\|_2^2 \tag{3}$$

Note that constraints 1 and 2 guarantee that each row and column of $\mathbf{H}$ has exactly one 1 element while other elements are 0.

3. *Triplet Constraints*. As there are more than 2 cameras in the camera network, there may exist inter-camera inconsistency in the network. As mentioned above, we only need to keep consistency in all camera triplets. We summarize this constraint in the matrix form. For a camera triplet, say a, b, c, we introduce the loop constraint as:

$$\mathbf{J_T}^{a,c,b} = \| \max\{0, \mathbf{H}^{a,c}\mathbf{H}^{c,b} - \mathbf{H}^{a,b}\}\|_F^2 \tag{4}$$

where $\forall 1 \leq a < b < c \leq m$.

We see that if and only if the $(i,k)^{th}$ cell of $\mathbf{H}^{(a,c)}$ and the $(k,j)^{th}$ cell of $\mathbf{H}^{(c,b)}$ are 1, will the $(i,j)^{th}$ cell of $\mathbf{H}^{(a,c)}\mathbf{H}^{(c,b)}$ be 1. In this case, the $(i,j)^{th}$ cell of $\mathbf{H}^{(a,b)}$ should be 1 to guarantee the consistency. There are totally m-2 terms for every a, b, and we take the average of these terms.

To sum up, we formulate the following optimization objective function to achieve the above goal:

$$\begin{aligned}\min_{\mathbf{H}} \mathbf{J_1} \quad &= \quad \sum_{a,b=1}^{m} (-\|\mathbf{H}^{a,b} \cdot \mathbf{C}^{a,b}\|_F^2) \\ &+ \quad \sum_{a,b=1}^{m} (\alpha \mathbf{J_B}^{a,b} + \beta(\mathbf{J_R}^{a,b} + \mathbf{J_C}^{a,b})) \\ &+ \quad \mu \frac{1}{m-2} \sum_{a,c,b}^{m} \mathbf{J_T}^{a,c,b}\end{aligned} \tag{5}$$

Having obtained the globally optimal match, we expect the result exactly the same as the ground-truth. Specifically, we compute the loss between the obtained result and the ground-truth as follows:

$$\arg\min_{f} \mathbf{J_2} = (\|\mathbf{H} - \mathbf{H}^*\|_F^2) \tag{6}$$

3.2. The Overall Framework

We present the whole framework in a recurrent manner, as shown in Figure 2.

We first take a subset of persons from different cameras as training samples and feed them into f, which is a neural network in this work. We take the output of f as features and obtain a similarity matrix $\mathbf{C}^{a,b}$ for each camera pair $a, b, \forall a, b = 1, ..., m$. More specifically, we use the cosine metric to calculate the similarity between the feature $\mathbf{x}_i$ and $\mathbf{x}_j$, :

$$cos = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}, sim = \frac{cos + 1}{2} \tag{7}$$

Then we obtain the best $\mathbf{H}$ that maximizes the global similarity and minimizes the constraint items. The algorithm to obtain $\mathbf{H}$ will be described later. According to the obtained $\mathbf{H}$, we calculate the difference using 6 and propagate the difference between the obtained $\mathbf{H}$ and the ground-truth adjacency matrix $\mathbf{H}^*$ to train the neural network. More specifically, we intuitively consider $\mathbf{dC} = \mathbf{dH}$, so the $(i,j)^{th}$ entry of $\mathbf{H}$ and features $\mathbf{x}_i$, $\mathbf{x}_j$ can be computed as:

$$\frac{\partial \mathbf{J_2}}{\partial \mathbf{x}_i} = (\mathbf{H}_{i,j} - \mathbf{H}^*_{i,j}) \cdot \frac{1}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|} \cdot (\mathbf{x}_j - \frac{\mathbf{x}_i^T \mathbf{x}_j \mathbf{x}_i}{\mathbf{x}_i^T \mathbf{x}_i}) \tag{8}$$

$$\frac{\partial \mathbf{J_2}}{\partial \mathbf{x}_j} = (\mathbf{H}_{i,j} - \mathbf{H}^*_{i,j}) \cdot \frac{1}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|} \cdot (\mathbf{x}_i - \frac{\mathbf{x}_j^T \mathbf{x}_i \mathbf{x}_j}{\mathbf{x}_j^T \mathbf{x}_j}) \tag{9}$$

We fix the training person set and use the adjusted neural network to repeat the above process. In this manner, we train our neural network by iteratively optimize $\mathbf{H}$ and $\mathbf{C}$ to meet the globally optimal result for this batch of persons.

The way to obtain the globally optimal association matrix from $\mathbf{C}$ in different camera pairs is not trivial. We denote the relationship between $\mathbf{H}$ and $\mathbf{C}$ as: $\mathbf{H} = \Phi(\mathbf{C}, \alpha, \beta, \mu)$. It's hard to obtain Φ explicitly. As an alternative, we proposed a gradient descent method to gradually approximate the optimal $\mathbf{H}$ that maximizes the global similarity and minimizes the constraint items at the same time.

We compute the following derivatives:

$$\begin{aligned}\frac{\partial \mathbf{J_1}}{\partial \mathbf{H}^{a,b}} = &- \mathbf{H}^{a,b} \cdot \mathbf{C}^{(a,b)2} \\ &+\alpha((\mathbf{H}^{a,b} - 0.5)^{\cdot 2} - 0.25) \cdot (\mathbf{H}^{a,b} - 0.5) \\ &+\beta((\mathbf{H}^{a,b}\mathbf{e} - \mathbf{e})\mathbf{e^T} + \mathbf{e}(\mathbf{e^T}\mathbf{H}^{a,b} - \mathbf{e^T})) \\ &+\mu \sum_{c}^{m} -(max\{0, \mathbf{H}^{a,c}\mathbf{H}^{c,b} - \mathbf{H}^{a,b}\})\end{aligned} \tag{10}$$

Algorithm 1 details the proposed method.

First, we initialize all $\mathbf{H}$ matrices with every entry equals to $\frac{1}{n}$, so that the sum of each row and column is 1, which meets row and column constraints. Then, we get the gradient $\mathbf{dH}$ so that we perform gradient descent over $\mathbf{H}$. After that we impose double ReLU to $\mathbf{H}$ to help the matrix converge faster.

3.3. Extension to General Cases

In real-world scenarios, we do not always ensure that every person appear in every camera, so the number of persons in different cameras is different, like the Market-1501 dataset. Here we generalize our algorithm to the situation where only part of people appear in a certain camera. This situation can be easily obtained by adapting the row and column constraints in the objective function and the corresponding gradient. In this case, $\mathbf{C}$ and $\mathbf{H}$ matrices are not always square, and some rows or columns of $\mathbf{H}$ may be all zeros. For example, person P_1 appears in camera a but does not appear in camera b, the corresponding row and column will be all zeros as no match is found. Instead of keeping the sum of every row and column to 1, we only need to keep the sum 0 or 1. The modified row and column constraints are described as follows:

$$\mathbf{J_R}' = (\|(\mathbf{H}e - 0.5e) \cdot (\mathbf{H}e - 0.5e) - 0.25e\|_2^2 \quad (11)$$

$$\mathbf{J_C}' = \|(e^T\mathbf{H} - 0.5e^T) \cdot (e^T\mathbf{H} - 0.5e^T) - 0.25e^T\|_2^2 \quad (12)$$

We see that if the sum of each row and column is 0 or 1, no penalty will be exerted. We derive the gradient by using the same rule. Due to the length limitation, we put the complete objective and the modified gradient in the supplementary.

Algorithm 1 CADL

Input: Training images set, label index, learning rate lr, iterative number I_t, iterative number I_H and parameter α, β, μ.

Output: neural network: f

Initialize f
for $iter < I_t$ **do**
  Select a batch of persons $\mathbf{P} = \{P_i\}$
  Extract features $\mathbf{X} = \{\mathbf{x}_i\}$ for $\mathbf{P}$
  Calculate $\mathbf{C}^{a,b}$ using (7)
  Initialize $\mathbf{H}^{a,b}$
  for $i < I_H$ **do**
    for $\forall a, b = 1, ..., m, a < b$ **do**
      Generate $\mathbf{dH}^{a,b}$ using (10)
    end for
    for $\forall a, b = 1, ..., m, a < b$ **do**
      $\mathbf{H}^{a,b} \leftarrow \mathbf{H}^{a,b} - lr * \mathbf{dH}^{a,b}$
      $\mathbf{H}^{a,b} \leftarrow max(\mathbf{H}^{a,b}, 0)$
      $\mathbf{H}^{a,b} \leftarrow min(\mathbf{H}^{a,b}, 1)$
    end for
  end for
  $\mathbf{dH}^{a,b} \leftarrow \mathbf{H}^{a,b} - \mathbf{H}^{a,b*}, \forall a, b = 1, ..., m$
  Back propagation $\mathbf{dH}$ to adjust f using (8-9)
  Repeat 4 - 18 until reaching global optimum
end for
return neural network: f

4. Experiments

4.1. Datasets and Protocols

We conducted experiments on three popular person re-identification datasets that have three or more cameras to form a camera network, Market-1501 [47], RAiD [3, 5] and WARD [25], having 6, 4, and 3 cameras respectively. While several state-of-the-art methods for person re-identification *e.g.*, [4, 19, 35, 37] evaluated their performance on other popular datasets (*e.g.* i-LIDS [48], VIPeR [9], CUHK01 [16], CUHK03 [17]), these datasets do not fit our purpose since they only provide images from two different cameras for a certain sequence. Here we give a brief description of these three datasets.

Market-1501 [47]: The Market-1501 dataset has 32,668 bounding boxes of 1501 persons captured from 6 cameras in front of a supermarket in Tsinghua University. The dataset employs Deformable Part Model (DPM) as pedestrian detector to obtain the bounding boxes, which is not as ideal as human annotated ones but is more close to the real-world scenario. Besides the true positive bounding boxes, the dataset also provides false alarm detection results, which makes the dataset more challenging. The standard evaluation protocol in [47] treats the re-identification problem as an image search problem, which is not very suitable to evaluate our framework. Hence we proposed a special protocol by evaluating the matching accuracy for all 30 camera pairs between the test and probe sets. Since there are so many results, it is not possible to show all of them, we calculated weighted average and variance to evaluate the performance (the full results are given in the supplementary). In some scenarios, to evaluate the effectiveness of our method on pairwise re-identification, we also conducted experiments on the standard protocol where rank-1 accuracy and mean average precision (mAP) were measured to show that even only training with our method in such cases.

RAiD [3, 5]: Re-identification Across indoor-outdoor Dataset (RAiD) has 6920 bounding boxes of 43 identities captured by 4 cameras. The cameras are numbered as 1,2,3 and 4 where camera 1 and 2 are indoor while camera 3 and 4 are outdoor. The images are affected by very large illumination variations between indoor and outdoor situations. Following the same protocol in [3, 5], 21 persons were used for training and 20 were used in testing. This dataset is

Table 1. Performance comparison of state-of-the art algorithms on the Market-1501 dataset using our own protocol. We measure the matching accuracy for all 30 camera pairs and calculate the weighted average and variance. Our method is more accurate and balanced between different camera pairs.

Method	Weight Acc.	Var.
BoW [47] - (SQ)	21.14	0.0321
Ours - Pretrained - (SQ)	27.29	0.0221
Ours - Contrastive - (SQ)	46.23	0.0084
Ours - Cosine - (SQ)	52.88	0.0049
Ours - CADL - (SQ)	**60.09**	**0.0111**
BoW [47] - (MQ)	27.76	0.0258
Ours - Pretrained - (MQ)	40.51	0.0189
Ours - Contrastive - (MQ)	59.16	0.0168
Ours - Cosine - (MQ)	72.02	0.0043
Ours - CADL - (MQ)	**81.15**	**0.0039**

Table 2. Performance comparison of state-of-the art algorithms for the Market-1501 dataset using standard protocol. The proposed CADL framework not only outperforms the pretrained model and contrastive loss trained model, but also outperforms all the state-of-the-arts by large margins.

Method	Rank 1	mAP
SDALF [6]	20.53	8.20
eSDC [46]	33.54	13.54
BoW [47] - (SQ)	34.40	14.09
DNS [44] - (SQ)	61.02	35.68
Gated Siamese [34]	65.88	39.55
Ours - Pretrained - (SQ)	35.65	12.82
Ours - Contrastive - (SQ)	58.28	41.28
Ours - Cosine - (SQ)	**73.84**	**47.11**
BoW [47] - (MQ)	42.14	19.20
BoW + HS [47] - (MQ)	47.25	21.88
S-LSTM [35] -(MQ)	61.60	35.31
DNS [44] - (MQ)	71.56	46.03
Gated Siamese [34] - (MQ)	72.92	45.39
Ours - Pretrained - (MQ)	41.57	15.97
Ours - Contrastive - (MQ)	69.09	49.20
Ours - Cosine - (MQ)	**80.85**	**55.58**

challenging for deep learning methods due to the small data size. It is impossible to train a deep neural network on such a small set of data, so we took the pretrained model provided in [39] and fine-tuned it on RAiD. We conducted 20 experiments without applying the provided masks and took average as our result.

WARD [25]: The WARD dataset has 4786 images of 70 persons acquired in a real surveillance scenario in three non-overlapping cameras. The dataset also has a huge illumination variation apart from resolution and pose changes. Following the protocol proposed in [3,5], we conducted our experiment in the same way as RAiD for 20 times. Again, we did not apply the provided masks. And our method achieved almost 100% accuracy on all 3 camera pairs.

4.2. Settings

Due to the limitation of GPU memory, it is impossible to feed all person images for all cameras into CNN at once, so in our experiments, we considered 3 cameras each time and took 30 persons per camera to feed into our CNN for WARD and Market-1501 datasets. We considered all 4 cameras and took 21 persons per camera to feed into CNN for RAiD dataset. We set the initial learning rate as 0.01 and reduced by a factor of 0.1 every 10,000 batches. After careful adjustment, we set the hyper-parameter as $\alpha = 0.01, \beta = 1, \mu = 1$. We set momentum = 0.9 and weight decay = 0.005. All the experiments were conducted using Caffe [12] without data augmentation.

Due to the small size of WARD and RAiD dataset, it is infeasible to train a neural network from the scratch, so we performed our method by fine-tuning pretrained model using CADL. More specifically, we chose the pretrained model provided in [39]. The CNN starts with 4 concatenated convolution layers followed by a pooling layer. Next is a series of 6 inception units. After the final fully connected layer, the CNN produces features of length 256. The detailed structure of the pretrained CNN is described in our supplementary materials. Unlike other pretrained model trained on ImageNet, it takes input of size 144×56 which is much more suitable to due with person images. We give a short analysis proving its advantage over other pretrained model in the supplementary.

For our framework, we evaluated 4 results on datasets:

1. **Ours - Pretrained**. This is the result obtained using the original pretrained model from [39]. We conducted this experiment to show the effectiveness of the pretrained model on three datasets.

2. **Ours - Contrastive**. We fine-tuned the pretrained model using contrastive loss to provide a baseline for evaluating CADL. Contrastive loss is widely used in siamese neural network structure, *e.g.* [35, 42], and can well represent the results for regular training approach. We randomly selected the camera and person ID, and the margin was kept as 1.0, the ratio between negative sample pairs and positive sample pairs are set as 2.

3. **Ours - Cosine**. In some scenarios, we need to perform pairwise re-identification, *e.g.* the standard protocol for Market-1501. So we trained our model using CADL and conducted pairwise re-identification using cosine metric. We took the rank-1 result to evaluate the performance. In this experiment, we only trained our model using CADL and experimental results show that CADL can well boost

Table 3. Performance comparison of state-of-the art algorithms for the RAiD dataset, we report the rank-1 results for all camera pairs.

Method	Cam12	Cam13	Cam14	Cam23	Cam24	Cam34	Avr. Acc.	Var
FT [3, 5]	74.00	26.00	41.00	41.00	52.00	61.00	49.17	0.0287
ICT + NCR [3, 5]	89.00	60.00	66.00	60.00	71.00	68.00	69.00	0.0115
FT + NCR [3, 5]	86.00	67.00	68.00	75.00	74.00	79.00	78.83	0.0050
Ours - Pretrained	47.14	60.71	31.19	51.90	25.24	73.33	48.25	0.0324
Ours - Contrastive	48.33	60.24	50.92	67.14	55.00	67.62	58.21	0.0067
Ours - Cosine	67.14	88.81	67.62	88.57	58.57	93.57	77.38	0.0214
Ours - CADL	**97.50**	**98.25**	**100.00**	**96.00**	**95.50**	**96.00**	**97.21**	**0.0003**

Table 4. Performance comparison of state-of-the art algorithms for the WARD dataset, we report the rank-1 results for all camera pairs. The results are averaged from 20 experiments and our method reach almost 100% performance.

Method	Cam12	Cam13	Cam23
ICT + NCR [3, 5]	40.00	26.85	36.57
FT + NCR [3, 5]	57.14	45.14	61.71
Ours - Pretrained - (SQ)	51.29	77.29	65.43
Ours - Contrastive - (SQ)	63.57	70.71	78.14
Ours - Cosine - (SQ)	93.57	90.14	94.57
Ours - CADL - (SQ)	**99.57**	**99.29**	**99.71**

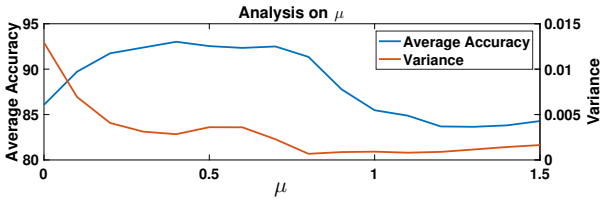

Figure 3. An analysis on μ. We conducted an experiment on RAiD dataset by changing μ from 0 to 1.5, and plot the average accuracy and variance. From the curve we can see that at about $\mu = 0.8$ we have the highest accuracy and the smallest variance.

the performance in the training stage.

4. **Ours - CADL**. We trained and tested with CADL framework. After training and calculating the similarities, we applied the gradient descent algorithm to approximate the globally optimal match. By this experiment we obtained the best result where average accuracy is highest with smallest variance. Notice that this result is not available for the standard protocol of Market-1501, which conducts pairwise re-identification.

4.3. Results and Analysis

Results: The results of Market-1501, RAiD and WARD datasets are given in table 1 & 2, table 3, and table 4 respectively.

The results show that CADL outperforms all the traditional or deep methods on the 3 datasets. For Market-1501, which is very challenging dataset due to the large person numbers and distractors, our method reaches state-of-the-art performance both in the average accuracy and the variance. Under our protocol, we obtained a more accurate and balanced CNN for different camera pairs using CADL. Under the standard protocol where many state-of-the-art methods are evaluated, we outperforms all the state-of-the-arts by 8%. RAiD dataset has 4 cameras and only 22 persons for training, which is difficult for deep learning based methods. Nevertheless, our framework can address the problem of overfitting effectively, and achieved remarkable result. For the WARD dataset, CADL reached over 99% accuracy for all the camera pairs and outperformed the state-of-the-art by large margins. As shown in Table 4, our method has significant performance gains compared to the pretrained model and contrastive loss based method.

Even without obtaining globally optimal match in the testing stage, *i.e*. only training with CADL and performing pairwise re-identification, as reported in **Ours - Cosine**, large improvements were gained compared to contrastive loss and pretrained models. Furthermore, our framework outperforms all the state-of-the-art methods. This shows that our method is very effective for re-identification in a camera network. Training with CADL can significantly boost the performance of both pairwise re-identification or camera network based re-identification. By implementing our algorithm after similarity is generated, we can further improve the result and bring down the variance (see **Ours - CADL**).

Hyper-Parameters: We analyzed the effect of parameter setting. From previous experiments we found that the values of α and β cannot vary too much, otherwise the algorithm will fail. So we conducted an experiment on μ to show the effect of inter-camera information. By introducing μ we can bring down the inconsistency between different camera pairs and balance the performance. So a proper μ can help to increase the average accuracy and bring down the variance. We conducted the experiment on the RAiD dataset,

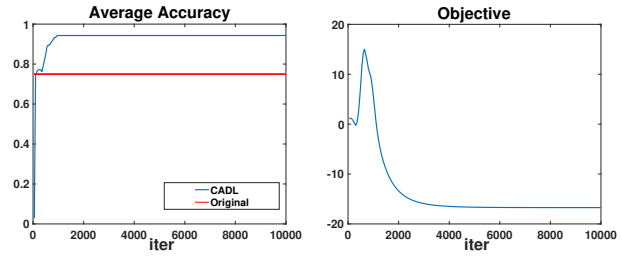

Figure 4. An example to show the convergency and effectiveness of our algorithm. The blue curve is the average accuracy of obtained $\mathbf{H}$, and the red curve is the original accuracy of $\mathbf{C}$. We can see that accuracy improvement is gained and the objective J goes down smoothly.

and plotted the average accuracy for $\binom{4}{2} = 6$ camera pairs and the variance along μ in Figure 3. The figure shows that as μ increases, the average accuracy first increases to the peak due to the obtained consistency, and then decreases for too large μ. As for the variance, it decreases as μ increase from 0 to 1. Taking accuracy and variance into account, we can reach an optimal result by setting $\mu = 0.8$ approximately.

Convergency and Effectiveness of H Approximation: To demonstrate the convergency and effectiveness of our gradient descent algorithm, we first conduct a small experiment on dataset WARD. After careful adjustment, we set the hyper-parameter as $\alpha = 0.01, \beta = 1, \mu = 1$. We initialized all $\mathbf{H}$ matrices with every entry equals to $\frac{1}{n}$. We plotted the accuracy of 3 camera pairs and the overall loss in Figure 4. We can see that as iterations goes on, the loss drops smoothly, and the average accuracy climbed up above the original accuracy (obtained by matching the sample with smallest distance).

As we initialized the adjacency matrix to meet the row and column constraints, the initial objective is very small. At the beginning of the optimization, the objective increases due to the increase of row and column loss. However, the objective decreases significantly later and converges to a good level for 3,000 iterations.

Further experiments show that our algorithm is robust and insensitive to scale.

Comparison with NCR: Previous work [3, 5] on camera network also presents a method to obtain $\mathbf{H}$, named NCR, by solving a binary integer programming problem. It gains performance improvement. As mentioned above our framework CADL exploited such consistent-ware information throughout the training and testing process, but here, we would like to give a comparison by only generating globally optimal $\mathbf{H}$ in the testing stage to compare the performance of two methods. Using the standard protocol provided in [3, 5], we test the performance and speed using the similarity matrix obtained from contrastive loss on WARD. The experiment was conducted 20 times using MATLAB on i7-4750HQ with 8GB RAM. The results are shown in Table 5. We can see that our method CADL outperforms NCR in both accuracy and speed. Our algorithm is much faster and stable than NCR, and obtain little accuracy improvement.

Table 5. Comparison between NCR and our CADL.

	Cam12	**Cam13**	**Cam23**	**Time**
Original	63.57	70.71	78.14	-
NCR	79.43	85.14	85.14	5.2558
Ours	**82.29**	**86.86**	**88.57**	**0.3768**

Furthermore, as the number of persons increases, the number of constraints for BIP will increase significantly. Suppose there are m cameras and n persons in the camera network, the number of constraints is approximately $n^5 * m^4$ (as proposed in the supplementary of [3,5]), which makes it impossible to solve for scenario where there are large number of people, *e.g.* in Market-1501, there are 6 cameras and 1501 persons, so the total number of constraints will reach 10^{16}. On the contrast, our gradient descent based algorithm can work robustly on different scales.

5. Conclusion

In this paper, we proposed the first end-to-end consistent-aware deep learning (CADL) method for person re-identification in a camera network. We solved the problem of person re-identification in a camera network by exploiting intra-camera and inter-camera consistent-aware information both in the training and testing stages. We also presented a gradient descent based algorithm to obtain globally optimal matching that maximizes the global similarity satisfying the consistency constraints. Experimental results have been proposed to validate the effectiveness of our method.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001004, the National Natural Science Foundation of China under Grants 61672306, 61572271, 61527808, 61373074 and 61373090, the National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015. 2, 3

[2] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*, pages 179–184, 2011. 3

[3] A. Chakraborty, A. Das, and A. K. Roy-Chowdhury. Network consistent data association. *TPAMI*, 38(9):1859–1871, 2016. 2, 5, 6, 7, 8

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016. 3, 5

[5] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*, pages 330–345, 2014. 2, 5, 6, 7, 8

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010. 6

[7] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 3

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 3

[9] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, volume 3, 2007. 5

[10] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, pages 346–361, 2014. 3

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 6

[13] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, pages 4516–4524, 2015. 1, 2

[14] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012. 2

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 3

[16] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, pages 31–44, 2012. 5

[17] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2, 3, 5

[18] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, pages 3610–3617, 2013. 2

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 1, 2, 5

[20] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015. 3

[21] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *CVPR*, pages 1301–1306, 2010. 2

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[23] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, pages 1137–1145, 2015. 1, 2

[24] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, pages 11–pages, 2012. 1, 2

[25] N. Martinel and C. Micheloni. Re-identify people in wide area camera network. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 31–36, 2012. 5, 6

[26] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, pages 3318–3325, 2013. 3

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 3

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 3

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 3

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[31] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 3

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 3

[33] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 3

[34] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016. 2, 3, 6

[35] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, pages 135–153, 2016. 3, 5, 6

[36] R. R. Varior, G. Wang, J. Lu, and T. Liu. Learning invariant color features for person reidentification. *IEEE Transactions on Image Processing*, 25(7):3395–3410, 2016. 1, 2

[37] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. 2, 3, 5

[38] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014. 3

[39] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016. 3, 6

[40] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16, 2014. 1, 2

[41] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*, pages 536–551, 2014. 1, 2

[42] D. Yi, Z. Lei, S. Liao, S. Z. Li, et al. Deep metric learning for person re-identification. In *ICPR*, volume 2014, pages 34–39, 2014. 3, 6

[43] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2016. 3

[44] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016. 6

[45] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, pages 2528–2535, 2013. 1, 2

[46] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013. 1, 2, 6

[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1, 2, 5, 6

[48] W. S. Zheng, S. Gong, and T. Xiang. Associating groups of people. *BMVA*, pages 23–1, 2009. 5

[49] W. S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, pages 649–656, 2011. 3