This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3043444, IEEE Transactions on Circuits and Systems for Video Technology

Exploiting Global Camera Network Constraints for Unsupervised Video Person Re-identification

Xueping Wang, Rameswar Panda, Min Liu, Yaonan Wang and Amit K. Roy-Chowdhury, *Fellow, IEEE*

***Abstract*—Many unsupervised approaches have been proposed recently for the video-based re-identification problem since annotations of samples across cameras are time-consuming. However, higher-order relationships across the entire camera network are ignored by these methods, leading to contradictory outputs when matching results from different camera pairs are combined. In this paper, we address the problem of unsupervised video-based re-identification by proposing a consistent cross-view matching (CCM) framework, in which global camera network constraints are exploited to guarantee the matched pairs are with consistency. Specifically, we first propose to utilize the first neighbor of each sample to discover relations among samples and find the groups in each camera. Additionally, a cross-view matching strategy followed by global camera network constraints is proposed to explore the matching relationships across the entire camera network. Finally, we learn metric models for camera pairs progressively by alternatively mining consistent cross-view matching pairs and updating metric models using these obtained matches. Rigorous experiments on two widely-used benchmarks for video re-identification demonstrate the superiority of the proposed method over current state-of-the-art unsupervised methods; for example, on the MARS dataset, our method achieves an improvement of 4.2% over unsupervised methods, and even 2.5% over one-shot supervision-based methods for rank-1 accuracy.**

***Index Terms*—Video person re-identification, Consistent constraints, Cross-view label estimation**

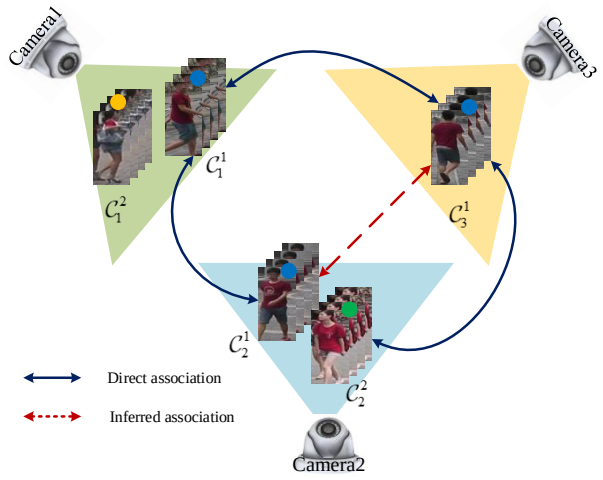

Fig. 1. An illustrative example of the contradictory matches in a camera network. Different dots indicate the identity associations. $\mathcal{C}_r^i$ denotes sample i captured in camera r. Assuming that the cross-camera positive matching associations $(\mathcal{C}_1^1, \mathcal{C}_2^1)$, $(\mathcal{C}_1^1, \mathcal{C}_3^1)$ and $(\mathcal{C}_2^2, \mathcal{C}_3^1)$ can be obtained independently by using some label estimation methods shown in blue lines. We can infer that $(\mathcal{C}_2^1, \mathcal{C}_3^1)$ is also a positive match because they are matched to the same sample $\mathcal{C}_1^1$ shown in red line. However when combining them together, there is an infeasible scenario that indicates that $\mathcal{C}_2^1$ and $\mathcal{C}_2^2$ are with the same label. Best viewed in color.

I. Introduction

PERSON re-identification (re-id) is a cross-camera instance retrieval problem which aims at searching persons across multiple cameras [1], [2]. In recent years, video-based person re-id has attracted increasing attention because video data provides richer information than images and it is easier to obtain than before. Some video-based person re-id methods have been proposed and achieved impressive results [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], however, their performance largely depends on huge amount of labeled data which are difficult to collect in real world applications. Consequently, in this work, we aim to develop a fully unsupervised solution for video person re-id that does not require any identity labels.

Recently, some unsupervised person re-id methods have been proposed that exploit unlabeled data progressively by assigning pseudo-labels and updating the re-id model in an alternative manner [18], [19], [20]. Despite promising results on common benchmarks, most of these methods are not fully unsupervised and still require some label information, such as source domain labeled data (domain adaption-based unsupervised method) [21], [22], to train a model, which limits the scalability of prior methods in practical applications. In recent years, some cross-camera matching methods have been proposed for person re-id or object tracking in a camera network and they achieved impressive performance [23], [24], [25]. However, most of these methods only consider the intra-camera and inter-camera matching correlations of samples independently [23], [24], [25], but ignore the higher-order relationships across the entire camera network. This may lead to contradictory outputs when matching results from different camera pairs are combined.

To illustrate this further, consider Figure 1 which shows a camera network containing 3 cameras and each of them capture 1-2 persons. Assume that the cross-camera positive matching associations between $(\mathcal{C}_1^1, \mathcal{C}_2^1)$, $(\mathcal{C}_2^2, \mathcal{C}_3^1)$ and $(\mathcal{C}_1^1, \mathcal{C}_3^1)$ can be obtained independently by using some label estimation methods ($\mathcal{C}_r^i$ denotes ith person captured in camera r). We can infer that $(\mathcal{C}_2^1, \mathcal{C}_3^1)$ is also a positive matching pair because

Xueping Wang, Min Liu and Yaonan Wang are with the College of Electrical and Information Engineering at Hunan University and National Engineering Laboratory for Robot Visual Perception and Control Technology, Changsha, Hunan, China. (*Corresponding author: Min Liu*)

Amit K. Roy-Chowdhury and Rameswar Panda are with the Department of Electrical and Computer Engineering at the University of California, Riverside.

E-mails: (wang_xueping@hnu.edu.cn, rpand002@ucr.edu, liu_min@hnu.edu.cn, yaonan@hnu.edu.cn, amitrc@ece.ucr.edu)

they are matched to the same person $\mathcal{C}_1^1$. However, when these matches from different camera pairs are combined, it leads to an infeasible scenario - $\mathcal{C}_2^1$ and $\mathcal{C}_2^2$ are with the same label. It is hard to distinguish which matches are reliable. Few recent methods [26], [27], [28] introduce global camera network constraints into person re-id task for reducing the unreliable matches by exploring high-order relationships in a camera network. However, they require a large number of labeled samples to train their models or the complex optimization method. Motivated by this, we ask an important question in this paper: *Can we develop a reliable cross-camera label estimation strategy, in which the matches are with a guarantee of consistency, for improving the performance of unsupervised re-id without requiring any labeled samples?* This is an especially important problem in the context of many person re-id systems involving large number of cameras.

To address such problems, in this paper, we propose a consistent cross-view matching framework by exploiting global camera network constraints for unsupervised video person re-id. First, the proposed method is fully unsupervised. We propose to use a first neighbor-based clustering strategy [29] to discover the intra-camera label relations and then cross-view matching to explore the inter-camera correlations without requiring any labeled samples for model learning. Second, our approach generates cross-view matches with a guarantee of consistency. Specifically, global camera network constraints are introduced into the cross-view matches to obtain the reliable matching pairs, including a definition for the reliability of matches to reduce the false positive ones. Finally, we learn metric models for camera pairs progressively by using an iterative updating framework which iterates between consistent cross-view matching and metric models learning.

To summarize, the contributions of our work are as follows.

- We propose a fully unsupervised, consistent cross-view matching framework, for video person re-id, in which the estimated cross-camera positive matching pairs follow the notion of consistency.
- We propose a definition for reliability of the cross-view matches via introducing global network constraints, which can reduce the incorrect matches significantly.
- Extensive experiments demonstrate that our approach outperforms the state-of-the-art unsupervised methods on MARS and DukeMTMC-VideoReID datasets and is very competitive while comparing with one-shot supervision-based methods.

Note that the cross-camera label estimation task usually suffers from large inter-camera variations due to different camera environments and self appearance changes of persons. Thus, we focus on unsupervised video re-id task as video tracklets contain much richer information than images, which helps to disambiguate difficult cases that arise when trying to recognise a person in a different camera. Different from other re-id methods, our method focuses on optimizing the cross-camera matching relations in a camera network with an unsupervised consistent cross-view matching framework. By introducing the global camera network constraints, the obtained matches will be more reliable than that considering inter-camera relations independently. This strategy can be used to other cross-view label estimation tasks, such as cross-camera person re-id and cross-view object tracking [23], [24], [25].

II. Related Work

A. Unsupervised Person Re-id

Unsupervised learning models have recently received much attention in person re-id task as they do not require manually labeled data. Most of the proposed unsupervised person re-id methods exploit unlabeled data progressively by assigning pseudo-labels and updating re-id models in an alternative manner. Fan et al. [19] proposed a *k*-means clustering-based method to select reliable images gradually and use them to fine tune a deep neural network to learn discriminative features for person re-id. Lin et al. [18], [30] proposed a hierarchical clustering-based feature embedding method by regarding sample labels as supervision signals to train a non-parametric convolutional neural network [31]. Liu et al. [32] presented a person re-id method which iterates between cross-camera tracklet association and feature learning. Li et al. [33] proposed a deep learning based tracklet association method by jointly learning per-camera tracklet association and cross-camera tracklet correlation to obtain the label information.

Some cross-camera matching methods have been proposed for person re-id and they obtained impressive performance. Lin et al. [25] proposed a cross-camera encouragement (CCE) term to assign different distances to image pairs from different cameras, which explores the cross-camera relations and overcomes the negative effect caused by the wrong clustering results. In [30], images from different camera styles are generated for data augmentation, so that the relations between cameras are inferred during unsupervised training.

Generative adversarial networks have also been adopted to train a camera style transfer model to bridge the gap between the labeled source domain and unlabeled target domain. Zhong et al. [34] introduced camera style adaptation as a data augmentation approach that smooths the camera style disparities. Deng et al. [35] translated the labeled images from source to target domain in an unsupervised manner and then trained re-ID models with the translated images by supervised methods.

Despite promising results on common benchmarks, most of these methods ignore the high-order relationships in a camera network and still require some person identity information for their models training. On the contrary, our approach introduces the global camera network constraints into person re-id models to explore more reliable cross-camera sample pairs and it is a fully unsupervised method that does not require any identity labeled data.

B. Graph Matching for Person Re-id

Graph matching has been widely used in computer vision and machine learning domains, such as shape matching and object recognition [36], [37], [38]. Recently, several works also introduce it into the person re-id task. Wu et al. proposed an unsupervised graph association method [39] to mine the cross-view relationships and reduce the damage of noisy

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3043444, IEEE Transactions on Circuits and Systems for Video Technology

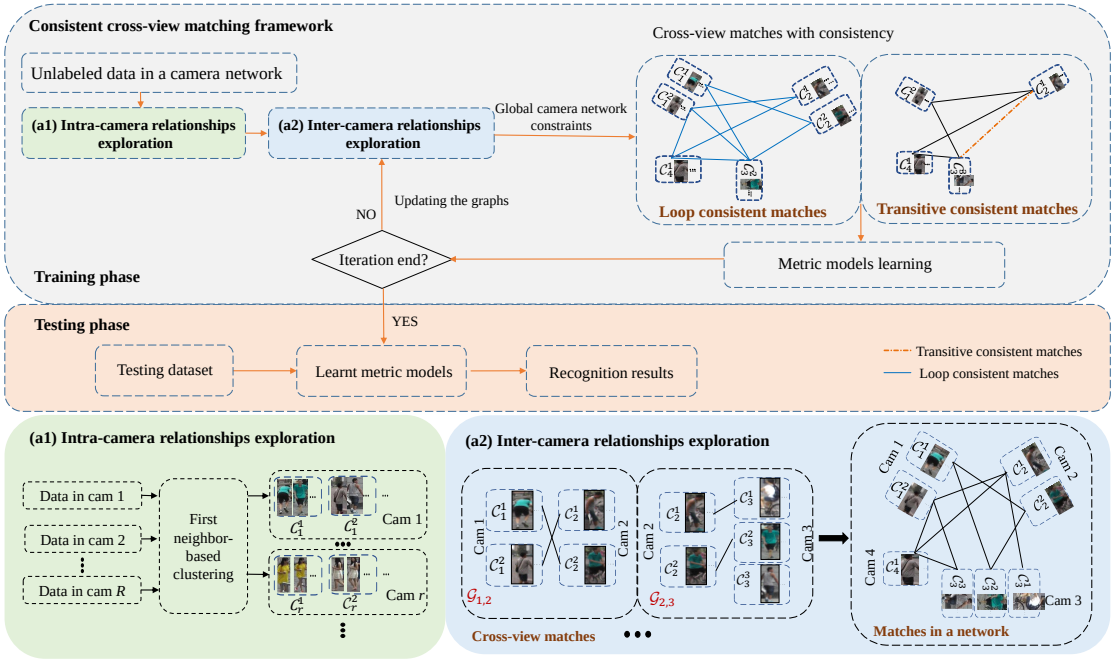

Fig. 2. Overview of our proposed method. This figure demonstrates the overall framework of the proposed approach. By introducing global camera network constraints into the matches in a camera network, we can select some reliable pairs with a guarantee of consistency. Thereafter, we learn metric models for camera pairs progressively by alternatively mining consistent cross-view matches and updating metric models. (a1) shows the intra-camera clustering for each camera. By using the first neighbor-based clustering algorithm, first neighbor relations can be obtained in each camera. According to Equation 1, the adjacency matrix can be obtained. Thereafter, the connected samples can be clustered together. $\mathcal{C}_r^i$ denotes ith cluster in camera r. (a2) illustrates the inter-camera relationships exploration across a camera network. There may be contradictory matches when combining all cross-view matches together, so we introduce global camera network constraints into these matches to obtain reliable pairs. Note that each image in this figure denotes one person tracklet.

associations. Ye et al. [6] presented a dynamic graph co-matching method to obtain the corresponding image pairs across cameras. Das et al. [27], [28] proposed a consistent re-id method in a camera network by considering the matching consistency to improve camera pairwise re-id performance. Following [27], Lin et al. [26] proposed a consistent-aware deep learning method by incorporating consistency constraints into deep learning framework for person re-id. Roy et al. [40] constructed a k-partite graph for the camera network and then exploited transitive information across the graph to select an optimal subset of image pairs for manual labeling. However, most of these methods often ignored the higher-order relationships in a camera network. The proposed approach introduced the global network constraints into cross-view matches, which can reduce the incorrect matches significantly and learn robust metric models for camera pairs progressively.

III. Consistent Cross-View Matching

In this section, we first explore the intra-camera label relationships by using a first neighbor-based clustering strategy. On top of that, we construct a graph for each pair of cameras and search for reliable cross-view matching pairs with consistency by introducing global camera network constraints. Finally, we learn the distance metric models for camera pairs progressively by alternatively mining consistent matches and updating the learned metric models. The overall framework of our proposed method is shown in Figure 2.

In camera r, we assume that there are N_r samples and denote it as $\mathcal{I}_r = \{I_r^1, I_r^2, ..., I_r^{N_r}\}$. A pre-trained feature embedding model $f(\cdot)$ is employed to extract features for the training samples $\mathcal{T}_r = \{T_r^1, T_r^2, ..., T_r^{N_r}\}$ and the extracted features are used as the inputs of our approach.

A. Intra-camera Relationships Exploration

In each camera, there is not much appearance variation between the samples with the same identity. So, we propose to utilize the first neighbor of each sample which can be obtained via fast approximate nearest neighbor methods (such as k-d tree) to explore the label relationships among samples and find the groups in each camera [29]. Specifically, given the indexes of the first neighbor of each sample in one camera, we define an adjacency matrix:

$$A(i,j) = \begin{cases} 1, & \text{if } i = k_j^1 \text{ or } j = k_i^1 \text{ or } k_i^1 = k_j^1; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where k_j^1 denotes that the first neighbor of sample j. The adjacency matrix links each sample i to its first neighbor via $j = k_i^1$, enforces symmetry via $k_j^1 = i$ and links samples $(i;j)$ that have the same neighbor with $k_i^1 = k_j^1$. Equation 1 for each camera returns a symmetric sparse matrix directly specifying a graph with connected components as the clusters (shown in Figure 2 (a1)). It is reasonable to regard each cluster as one person. So, one camera, e.g. camera r, can be denoted as $\mathcal{C}_r = \{\mathcal{C}_r^1, \mathcal{C}_r^2, ..., \mathcal{C}_r^{n_r}\}$ with n_r clusters/persons, where $\mathcal{C}_r^i$ is the ith cluster/person in camera r.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3043444, IEEE Transactions on Circuits and Systems for Video Technology

B. Inter-camera Relationships Exploration

1) Graph Construction: We construct a bipartite graph $\mathcal{G} = (U, V, E)$ for each pair of cameras where each part of the graph denotes one camera and the vertices are the obtained clusters/persons (in section A). For example, we could convert camera pair (p, q) into a graph $\mathcal{G}_{p,q} = (\mathcal{C}_p, \mathcal{C}_q, E_{M_{p,q}})$, where $\mathcal{C}_p = \{\mathcal{C}_p^1, \mathcal{C}_p^2, ..., \mathcal{C}_p^{n_p}\}$ and $\mathcal{C}_q = \{\mathcal{C}_q^1, \mathcal{C}_q^2, ..., \mathcal{C}_q^{n_q}\}$ denote camera p and q, respectively. Note that we will use the terms 'cluster', 'person' and 'vertex' interchangeably throughout our work. The edge $E_{M_{p,q}}$ is a matching cost matrix of camera pair (p, q) and each element $e_{M_{p,q}}^{i,j}$ describes the similarity of vertex pair $(\mathcal{C}_p^i, \mathcal{C}_q^j)$ which is computed through a minimum distance criterion that takes the shortest distance between samples in two clusters, as follows:

$$e_{M_{p,q}}^{i,j} = \min_{T_p^a \in \mathcal{C}_p^i, T_q^b \in \mathcal{C}_q^j} d_{M_{p,q}}(T_p^a, T_q^b) \tag{2}$$

where $M_{p,q}$ denotes a distance metric model learned using the estimated pairs with consistency from camera p and q, which is initialized with identity matrix, and $d_{M_{p,q}}(T_p^a, T_q^b) = (T_p^a - T_q^b)^T M_{p,q}(T_p^a - T_q^b)$.

2) Graph matching: We use the assignment matrix $X_{p,q}$ to represent the matching associations between the vertices across camera pair (p, q). Element $x_{p,q}^{i,j}$ in $X_{p,q}$ represents the matching association of the vertex $\mathcal{C}_p^i$ and $\mathcal{C}_q^j$, which is defined as follows:

$$x_{p,q}^{i,j} = \begin{cases} 1, & \text{if } \mathcal{C}_p^i \text{ and } \mathcal{C}_q^j \text{ are a matched pair;} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In a large camera network, it is common that one camera may not capture every person. In this situation, a person from one camera *p* can have at most one match from another camera *q*. In other words, the matching association values in every row or column of the assignment matrix $X_{p,q}$ can all be 0. As a result, the matching association constraints are as follows:

$$\sum_{j=1}^{n_q} x_{p,q}^{i,j} \leq 1, i = 1, 2, ..., n_p \text{ and } \sum_{i=1}^{n_p} x_{p,q}^{i,j} \leq 1, j = 1, 2, ..., n_q \tag{4}$$

where n_p and n_q are the number of persons/clusters in camera p and camera q, respectively.

To compute the assignment matrix across camera pairs, we formulate it as a binary linear programming with constraints as follows:

$$\begin{aligned} X_{p,q} = \arg\min_{x_{p,q}^{i,j}} & \sum_{i,j=1}^{n_p,n_q} e_{M_{p,q}}^{i,j} x_{p,q}^{i,j} \\ \text{subject to: } & x_{p,q}^{i,j} \in \{0, 1\}, \forall i = 1, ..., n_p, j = 1, ..., n_q \\ & \sum_{i=1}^{n_p} x_{p,q}^{i,j} \leq 1, \forall j = 1, ..., n_q \\ & \sum_{j=1}^{n_q} x_{p,q}^{i,j} \leq 1, \forall i = 1, ..., n_p \end{aligned} \tag{5}$$

The assignment matrix set $\mathbf{X} = \{X_{p,q} | p < q\}$ across the pair of cameras in a network can be obtained, where $X_{p,q} = \{x_{p,q}^{i,j} | i = 1, ..., n_p, j = 1, ..., n_q\}$.

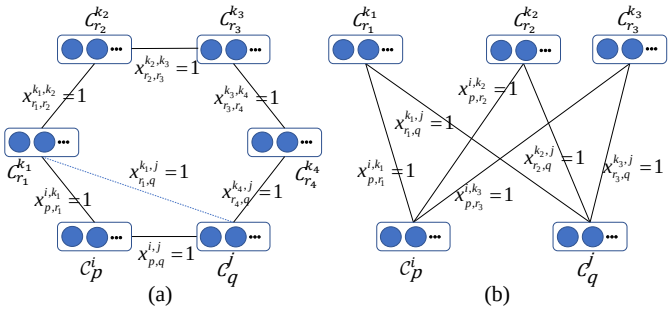

Fig. 3. An example of consistent cross-view matches. (a) demonstrates loop consistent constraint. If $x_{p,q}^{i,j} = 1$ and existing a person k_1 in camera r_1 satisfies $x_{p,r_1}^{i,k_1} x_{r_1,q}^{k_1,j} = 1$, the match $(\mathcal{C}_p^i, \mathcal{C}_q^j)$ is with consistency. (b) shows a transitive inference consistency pair $(\mathcal{C}_p^i, \mathcal{C}_q^j)$ and $\text{RT}_{p,q}^{i,j} = 3$.

C. Global Camera Network Constraints

Existing methods, like Hungarian algorithm [41] can be directly used to solve the above binary linear programming problem. However, it is hard to ensure that the obtained matching associations are reliable because Hungarian algorithm will try to get as many matching associations as possible. Thus, the assignment matrix may contain a lot of false positive matches. In addition, these cross-view matched pairs also do not consider matching consistency in a network of camera. It may lead to contradictory outputs when matching associations from different camera pairs are combined as shown in Figure 1. To address this problem, we introduce global camera network constraints including loop consistency constraints and transitive inference consistency constraints into these cross-view matches, which will guarantee the obtained cross-view matching pairs are with consistency.

1) Loop consistent matches: Given two vertices $\mathcal{C}_p^i$ and $\mathcal{C}_q^j$ from a camera pair (p, q) in a camera network, it can be noted that for consistency, logical 'AND' relationship between the association value $x_{p,q}^{i,j}$ and the set of association values $\{x_{p,r_1}^{i,k_1}, x_{r_1,r_2}^{k_1,k_2}, ..., x_{r_n,q}^{k_n,j}\}$ across possible vertices in different cameras has to be maintained, where $r_1, ..., r_n, p, q$ denote cameras in a network and $k_1, ..., k_n, i, j$ represent the persons captured by corresponding cameras. In other words, the association value $x_{p,q}^{i,j}$ between the two vertices $\mathcal{C}_p^i$ and $\mathcal{C}_q^j$ has to be 1, and it has to satisfy the indirect matching association $x_{p,r_1}^{i,k_1} x_{r_1,r_2}^{k_1,k_2} ... x_{r_4,q}^{k_4,j} = 1$ as shown in Figure 3 (a). In [27], [28], it has been proven that if the loop consistency constraint is satisfied for every triplet of cameras, it automatically ensures consistency for every possible combination of cameras taking 3 or more of them. Thus, the consistent matching pair $\mathcal{C}_p^i$ and $\mathcal{C}_q^j$ in the network of cameras has to satisfy the direct cross-view matching association $x_{p,q}^{i,j} = 1$ and a person k in camera r should satisfy: $x_{p,r}^{i,k} x_{r,q}^{k,j} = 1$ as shown in Figure 3(a).

$$x_{p,q}^{i,j} = 1 \text{ and } \exists \mathcal{C}_r^k, x_{p,r}^{i,k} x_{r,q}^{k,j} = 1, r \neq p, q \tag{6}$$

2) Transitive inference consistent matching: Transitive inference among person identities across multiple cameras and their logical consequences are strongly informative properties [40]. We exploit the transitive relations for enhancing the

performance of our cross-view matches. To illustrate the idea, let us consider a plausible scenario as shown in Figure 3(b). Assuming we have positive cross-view matches $(\mathcal{C}_p^i, \mathcal{C}_{r_1}^{k_1})$ and $(\mathcal{C}_{r_1}^{k_1}, \mathcal{C}_q^j)$, then according to the transitive inference we can directly infer that $\mathcal{C}_p^i$ and $\mathcal{C}_q^j$ also have the same label, i.e., $x_{p,r_1}^{i,k_1} x_{r_1,q}^{k_1,j} = 1 \Rightarrow x_{p,q}^{i,j} = 1$. Obviously, by introducing transitive inference, we can increase the number of cross-view matching pairs in a camera network. Usually, in a camera network, with more than two cameras, we define the reliability of the transitive inference-based cross-view matches

$$\mathrm{RT}_{p,q}^{i,j} = \sum_r \sum_{k=1}^{n_r} x_{p,r}^{i,k} x_{r,q}^{k,j}, p \neq q \text{ and } r \neq p, q \tag{7}$$

where *p, q* and *r* are cameras in a network. n_r is the number of persons/clusters in the camera *r*. $\mathrm{RT}_{p,q}^{i,j}$ denotes the reliability of the pair $(\mathcal{C}_p^i, \mathcal{C}_q^j)$. The larger the value $\mathrm{RT}_{p,q}^{i,j}$ is, the more reliable the transitive inference-based match is, as shown in Figure 3(b), i.e. $\mathrm{RT}_{p,q}^{i,j} = 3$. When the reliability value $\mathrm{RT}_{p,q}^{i,j}$ satisfies $\mathrm{RT}_{p,q}^{i,j} > 1$, we regard the matching pair $(\mathcal{C}_p^i, \mathcal{C}_q^j)$ as a transitive inference consistent match.

Note that loop consistency can be regarded as a specific form of the transitive inference consistency constraints. Therefore, combining them together, we define a metric for measuring the reliability of cross-camera matched pairs as follows:

$$\mathrm{RLT}_{p,q}^{i,j} = x_{p,q}^{i,j} + \sum_r \sum_{k=1}^{n_r} x_{p,r}^{i,k} x_{r,q}^{k,j}, r \neq p, q, \tag{8}$$

where $\mathrm{RLT}_{p,q}^{i,j}$ represents the reliability of the cross-view matched pair $(\mathcal{C}_p^i, \mathcal{C}_q^j)$. The larger the value is, the more reliable the match is. With this reliability score, we obtain the consistent assignment matrices $\hat{X}_{p,q} = \{\hat{x}_{p,q}^{i,j} | i = 1, ..., n_p, j = 1, ..., n_q\}$ to learn metric models for camera pairs as,

$$\hat{x}_{p,q}^{i,j} = \begin{cases} 1, & \text{if } \mathrm{RLT}_{p,q}^{i,j} > \theta; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where θ is a threshold that is used to balance the quality and quantity of the selected matches. Obviously, with the increase in θ value, the selected pairs will be more reliable, however, the number of the matches will be less for training. Thus, we can obtain sufficient and reliable cross-view matches in a camera network by selecting a suitable θ value for the unsupervised video person re-id task.

D. Metric Learning with Consistent Matches

Given a consistent assignment matrix $\hat{X}_{p,q}$ for camera pair (p, q), the corresponding metric model $M_{p,q}$ could be learned to update its matching cost matrix $E_{M_{p,q}}$. In this paper, we use the log-logistic metric learning as the loss function [42],

$$f_{M_{p,q}}(\mathcal{C}_p^i, \mathcal{C}_q^j) = \log(1 + e^{\hat{x}_{p,q}^{i,j}(e_{M_{p,q}}^{i,j} - \mu_{p,q})}) \tag{10}$$

where $e_{M_{p,q}}^{i,j}$ is the minimum distance between clusters $\mathcal{C}_p^i$ and $\mathcal{C}_q^j$ as calculated in Equation 2. $\mu_{p,q}$ is the average distance of all consistent matches from camera pair (p, q). Accordingly, for the camera pair (p, q), the overall cost function is

$$F(M_{p,q}; \hat{X}_{p,q}) = \sum_{i=1}^{n_p} \sum_{j=1}^{n_q} w_{i,j} f_{M_{p,q}}(\mathcal{C}_p^i, \mathcal{C}_q^j), M_{p,q} \succeq 0 \tag{11}$$

where $w_{i,j}$ is utilized to handle the imbalanced positive and negative pairs, i.e. $w_{i,j} = \frac{1}{N_{pos}}$ if $\hat{x}_{p,q}^{i,j} = 1$, and $\frac{1}{N_{neg}}$ otherwise, and N_{pos} and N_{neg} are the number of consistent matches and negative pairs.

During testing, we compute the distance of each query-gallery pair (T_{qu}, T_{ga}) by taking the minimum value under different pair-wise distance metric models as follows:

$$D(T_{qu}, T_{ga}) = \min_{\substack{p,q=1,...,R \\ p<q}} \{d_{M_{p,q}}(T_{qu}, T_{ga})\} \tag{12}$$

E. Iterative Updating

In this work, we learn metric models for camera pairs progressively by alternatively mining consistent cross-view matches and updating metric models. In each iteration, the learned metric models are used to update the corresponding matching cost matrix in Equation 5 for better exploring inter-camera relationships in a new iteration. Thereafter, the updated consistent cross-view matching correlations could be used to update the previous metric models. Finally, the reliable cross-view matches with consistency and distance metric models can be obtained.

1) Convergence Analysis: Note that we have two objective functions F and G for optimizing M and X in each iteration. To ensure the overall convergence of the proposed method, we adopt the same strategy as discussed in [43]. M can be optimized by choosing a suitable working step size $\eta \leq L$, where L is the Lipschitz constant of the gradient function $\bigtriangledown F(M; \hat{X})$. Thus, it could ensure $F(M^t; \hat{X}^{t-1}) \leq F(M^{t-1}; \hat{X}^{t-1})$, a detailed proof is shown in [44]. For $X_{p,q}^t$, the updating procedure at iteration *t* is constrained by keeping updating $E_{M_{p,q}}^t$ until a better $X_{p,q}$ is obtained, which satisfies $G(X_{p,q}^t; M_{p,q}^t) \leq G(X_{p,q}^{t-1}; M_{p,q}^t)$ where $G(X_{p,q}; M_{p,q}) = E_{M_{p,q}} X_{p,q}$, and the convergence analysis has been verified from [43].

2) Complexity Analysis: The major computational cost of our CCM comes from the intra-camera and cross-view label estimation. We assume that the number of samples in each camera is *n*. The intra-camera label estimation complexity is $\mathcal{O}(n\log n)$ [29]. After intra-camera clustering, we assume that each camera has *m* clusters, then the cross-view matching time complexity is $\mathcal{O}(m^3)$ [41]. Updating metric model M with accelerated proximal gradient is extremely fast as illustrated in [44]. So, the total complexity of our framework for each camera pair is $\mathcal{O}(n\log n) + \mathcal{O}(m^3)$. It may be noted that $m << n$ in a practical scenario. In Section IV-B, we analyze the real time cost of the proposed method on intra- and inter-camera label estimation.

IV. Experimental Results

A. Experimental Settings

1) Datasets: We use two publicly available video re-id datasets for experiments such as MARS dataset and DukeMTMC-VideoReID dataset. **MARS** [45] is captured by 6 cameras and contains 20,715 video tracklets of 1,261 identities. **DukeMTMC-VideoReID** (Duke-VideoReID) [46] is from the DukeMTMC dataset [47] which is captured by 8

Algorithm 1 Consistent Cross-view Matching

Input: Samples in each camera: $\mathcal{I}_r = \{I_r^1, I_r^2, ..., I_r^{N_r}\}$; pre-trained feature extractor: $f(\cdot)$; initialized distance metric model $M_{p,q}^0 = I$ for camera pair (p, q); the number of iterations: *maxIter*.

Output: Distance metric models $M_{p,q}$ for camera pairs.

$f(\cdot) : \mathcal{I}_r \rightarrow \mathcal{T}_r = \{T_r^1, T_r^2, ..., T_r^{N_r}\}$; //Feature extraction;
Intra-camera Relationships Explorations:
$\mathcal{T}_r \rightarrow \mathcal{C}_r = \{\mathcal{C}_r^1, \mathcal{C}_r^2, ..., \mathcal{C}_r^{n_r}\}$; // intra-camera clustering with Eq. (1);
Inter-camera Relationships Explorations:
construct graphs $\mathcal{G}_{p,q} = (\mathcal{C}_p, \mathcal{C}_q, E_{M_{p,q}^0})$ for camera pairs;
compute the assignment matrix $X_{p,q}^0$ with Eq. (5);
compute the consistent matches $\hat{X}_{p,q}^0$ with Eq. (8-9); // Global camera network constraints
for $t = 1$ to $maxIter$ **do**
    update the distance metric model $M_{p,q}^t$ with Eq. (11);
    update assignment matrix $X_{p,q}^t$ with Eq. (5);
    if $G(X_{p,q}^t; M_{p,q}^t) \leq G(X_{p,q}^{t-1}; M_{p,q}^t)$ **then**
        break;
    end if
    update the consistent matches $\hat{X}_{p,q}^t$ with Eq. (8-9);
end for

cameras and contains 4,832 tracklets of 1,404 identities. It may be noted that the proposed global camera network constraints are based on the triangle relationships in a camera network. That means we need the matching associations between every two of the three cameras at each time. Therefore, the proposed method can be used to the scenarios where there are more than two cameras in a camera network.

2) Feature extraction: In the training stage, we first employ a pre-trained feature embedding model to extract features for the training samples and use them as the inputs of our approach. In this paper, both hand-crafted features (LOMO) [48] and deep convolutional neural network (CNN) features are considered for evaluating the performance of our proposed method. The LOMO feature descriptor is of 26,960 dimensions and we use the principal component analysis (PCA) method [49] to reduce the dimension to 600. We adopt the pre-trained unsupervised feature embedding model in [18] which is designed for unsupervised person re-id to extract the deep CNN features and then ℓ_2 normalize it for all experiments. For the video-based datasets, we conduct mean-pooling for each tracklet to get more robust video feature representations.

3) Evaluation metrics: We follow the standard training/testing split [45], [46] of the two datasets to train and test the proposed model. The Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) are utilized to evaluate the performance of each method.

4) Implementation details: In this paper, we employ Hungarian algorithm [41] to solve the binary linear programming problem and the log-logistic metric learning (MLAPG) [42] for matching persons in re-id. Note that our approach is not specific to any type of matching and metric learning algorithms used for person re-id. During training, we learn a metric model for each pair of cameras using their corresponding matched pairs, and in testing, we match each query-gallery pair using the learned metric models and take the minimum value for each pair. All the reported results are based on $\theta = 1$ and deep CNN features. Moreover, on the MARS dataset, we evaluate the performance of the proposed method using both LOMO and CNN features. Note that our method does not require any labeled samples during model training and on the DukeMTMC-VideoReID dataset, we conduct cross-view matching directly without intra-camera clustering as the number of samples captured by each camera is small. The specific training process of the proposed consistent cross-view matching method can be found in Algorithm 1.

TABLE I
RANK-1, -5, -10 ACCURACY (%) AND MAP (%) PERFORMANCE USING DIFFERENT INTRA-CAMERA CLUSTERING METHODS ON THE MARS DATASET.

MARS	R1	R5	R10	mAP
DBSCAN	60.9	72.6	77.2	37.4
HDBSCAN	62.1	74.2	78.5	38.3
OURS + FN-C	65.3	77.8	81.3	41.2
OURS + Per-C	66.0	77.8	81.9	42.3

OURS+ FN-C represents that we use the first neighbor-based strategy for intra-camera samples clustering in our framework; OURS + Per-C denotes that we employ the perfect intra-camera associations in our framework.

B. Evaluation of Label Estimation

In this section, we evaluate the intra- and cross-camera label estimation performance of our proposed method. Specifically, on the MARS dataset, we first measure the intra-camera label estimation performance and observe that 70.9% samples of the same identities are clustered correctly indicating that our first neighbor-based clustering method is efficient for intra-camera label estimation. In addition, to better evaluate the advantages of the first neighbor-based intra-camera clustering strategy, we compare the recognition performance with other clustering methods, such as DBSCAN [50] and HDBSCAN [51], [52] on the MARS dataset. From Table I, it can be seen that our method outperforms the other clustering methods consistently. Comparing to HDBSCAN, the recognition performance is improved by 3.2% and 2.9% for rank-1 accuracy and mAP score, respectively. We further evaluate our method with the perfect intra-camera clustering performance (Per-C) which means that samples with the same identity are grouped together manually in each camera. It can be seen from Table I that by using the perfect intra-camera associations, the rank-1 accuracy and mAP score are just increased by 0.7% and 1.1% compared to the first neighbor based clustering strategy (FN-C). We may conclude that the first neighbor based clustering method is effective for intra-camera associations exploration.

We next validate the advantages of the proposed consistent cross-view matching approach for cross-camera label estimation. Specifically, we assume that in each camera, we can group all samples with the same identity together. It may be noted that this assumption just works in this subsection.

On top of the perfect intra-camera clustering results, Table II reports the performance of cross-camera label estimation with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3043444, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II
PERFORMANCE OF CROSS-VIEW MATCHING WITH/WITHOUT GLOBAL CAMERA NETWORK CONSTRAINTS ON TWO DATASETS.

Dataset	Setting	Pr (%)	Re (%)	F1 (%)
MARS	w/o GNC	59.5	82.7	69.2
	w/ GNC	72.3	76.8	74.5
Duke-VideoReID	w/o GNC	21.7	81.5	34.3
	w/ GNC	50.9	55.0	52.9

w/o GNC: cross-view matching without global network constraints; w/ GNC: cross-view matching with global network constraints.

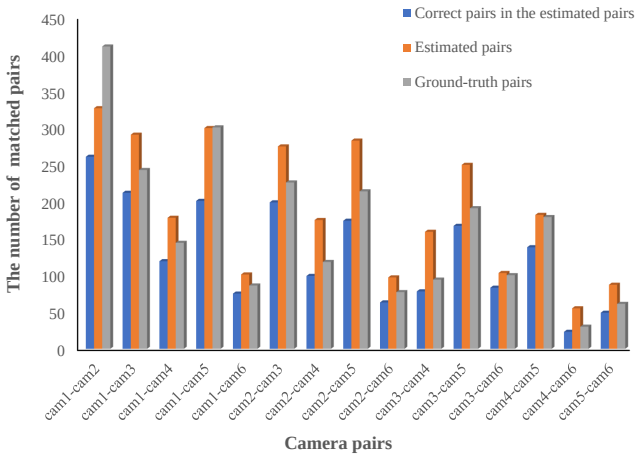

Fig. 4. The number of cross-camera matches on the MARS dataset.

or without global network constraints (GNC). The standard precision (Pr), recall (Re) and F1-score (F1) are utilized to illustrate the performance of the proposed consistent cross-view matching approach across a camera network. We can see that by introducing global network constraints the precision score is improved by a large margin, especially, the improvement of 12.8% and 29.2% can be obtained on MARS and DukeMTMC-VideoReID datasets, respectively and on the MARS dataset, 72.3% matched pairs are the correct matches. Moreover, it can be observed that by introducing the global network constraints into cross-view matches, the recall value drops a lot, but the F1-score is significantly improved. We believe this is due to Hungarian algorithm tries to obtain as many matching associations as possible across camera pairs and hence produces many false positive pairs.

We further demonstrate some specific cross-camera label estimation results on the MARS dataset as shown in Figure 4. In the figure, it can be seen that the proposed method obtains most of the correct pairs and reject the false positive matches across all the pairs of cameras. For example, camera #1 captures 520 different persons, camera #6 captures 104 persons on the MARS dataset, and there are 86 persons in common. That means 452 outliers will affect their matching performance. However, Figure 4 shows us that 101 matches can be obtained by our method, including 75 correct pairs among all 86 ground-truth matches. This again demonstrates that our method is very effective for cross-camera label estimation in a network of cameras.

TABLE III
RANK-1 ACCURACY (%) AND MAP SCORE (%) ON DIFFERENT θ VALUES.

Dataset	$\theta = 0$		$\theta = 1$		$\theta = 2$	
	R1	mAP	R1	mAP	R1	mAP
MARS	63.7	35.7	65.3	41.2	61.1	38.5
Duke-VideoReID	73.1	63.9	76.5	68.7	75.4	67.0

TABLE IV
RANK-1, -5, -10 ACCURACY (%) AND MAP (%) PERFORMANCE USING CROSS-VIEW DISTANCE METRIC MODELS.

MARS	R1	R5	R10	mAP
OURS w/o cross-view	65.2	77.1	81.1	40.4
OURS w cross-view	65.3	77.8	81.3	41.2
DukeMTMC-VideoReID	R1	R5	R10	mAP
OURS w/o cross-view	73.6	87.1	90.2	65.5
OURS w cross-view	76.5	89.6	91.9	68.7

w cross-view: training a metric model for each camera pair. w/o cross-view: training a global metric model for the entire dataset.

We also measure the real time cost of our method on label estimation. Specifically, on the MARS dataset, for each camera, it takes 0.27 seconds for intra-camera label estimation on average and it takes an average of 45.12 seconds for cross-camera label estimation. It may be noted that all experiments are conducted on an i5-7200U CPU.

C. Evaluation of Different Reliability Values θ

The quality and quantity of the estimated matching pairs are very important for learning an efficient pair-wise metric models in unsupervised person re-id. Both of them are related to the reliability value (θ) of the matches. Thus, we compare the recognition performance under different θ values to select the optimal one. Table III shows the Rank-1 accuracy and mAP score under 3 different reliability values. We can see that the recognition performance fluctuates a little with the increase in θ values because of the trade-off between the quantity and reliability of the matched pairs. A small reliability value means that we can collect most of cross-view matches, but the learned metric models using these pairs may not perform well because it also introduces massive false positive pairs. As the θ value increases, the number of matched pairs will decline, but the reliability of the matches will increase. We observe that when $\theta = 1$, we obtain the best recognition performance with 65.3% and 41.2% for rank-1 accuracy and mAP score respectively on the MARS dataset, similarly 76.5% and 68.7% on the DukeMTMC-VideoReID dataset.

D. Evaluation of Cross-view Metric Model

In the proposed consistent cross-view matching framework, we train a separate metric model for each pair of cameras. In this section, we show the effectiveness of the cross-view metric models over a global metric model for the entire dataset.

From Table IV, it can be seen that the cross-view metric models perform better than that training a global metric model for the entire dataset. Specifically, on the MARS dataset, the rank-1 accuracy and mAP score are improved by 0.1% and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3043444, IEEE Transactions on Circuits and Systems for Video Technology

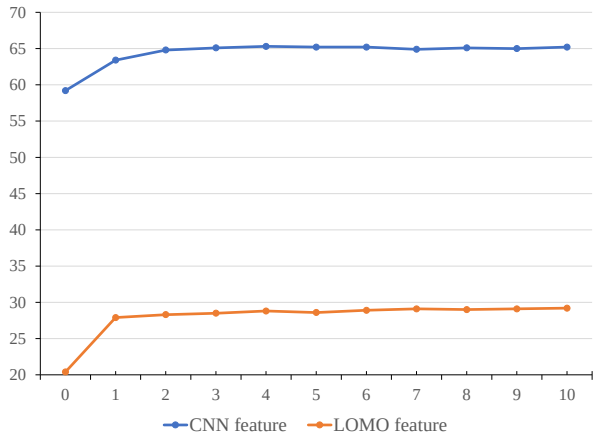

Fig. 5. Rank-1 accuracy of the proposed method on the MARS dataset with the LOMO feature and CNN feature at each iteration.

0.8%, respectively. On the DukeMTMC-VideoReID dataset, the rank-1 accuracy is increased from 73.6% to 76.5% (1.8% difference), and from 65.5% to 68.7% (1.5% difference) for mAP score.

E. Evaluation of Iterative Updating

Figure 5 shows the rank-1 accuracy of the proposed method using LOMO feature and CNN feature on the MARS dataset at each iteration. We iterate each experiment 10 times. As may be observed from the plot, with the increase in the number of iterations, the rank-1 accuracy is improved from 59.2% to 65.3% for CNN feature and 20.4% to 29.2% for LOMO feature, and after that, the plots are almost stable. By introducing the conditions of iteration end, our experiments will end at the 4th iteration and the 7th iteration for CNN feature and LOMO feature, respectively.

F. Comparison to the SOTA Methods

MARS Dataset. We compare our approach with several state-of-the-art person re-id methods that fall into two main categories: *unsupervised methods* such as GRDL [53], UnKISS [54], DGM+ [6] using LOMO feature, DGM+ [6] using deep IDE features, OIM [31], DAL [10], BUC [18], UTM [13], TAULD [33], UTAL [55], TSSL [14], CCE [25] and *semi-supervised methods* such as UGA (intra-camera supervision) [39], Progressive Learning (one-shot setting) [56], Stepwise (one-shot setting) [32], RACE (one-shot setting) [5] and EUG (one-shot setting) [46]. As seen from Table V, while comparing with unsupervised alternatives, we evaluate our method in two different settings: (1) methods based on hand-crafted features: the proposed method significantly outperforms all the compared methods; comparing to DGM+, we achieve 4.5% and 0.5% improvement using the same LOMO feature in rank-1 accuracy and mAP score, respectively; (2) methods based on deep learning: our method also obtains the best recognition performance 65.3% for rank-1 and 41.2% for mAP while comparing to fully unsupervised deep learning based alternatives, especially, comparing to BUC, the rank-1 accuracy and mAP score are improved by 4.2% and 3.2%,

TABLE V
RANK-1, -5, -10 ACCURACY (%) AND MAP (%) WITH SOME UNSUPERVISED AND SEMI-SUPERVISED APPROACHES ON THE MARS DATASET.

Methods	Labels	R1	R5	R10	mAP
GRDL [53]	None	19.3	33.2	41.6	9.6
UnKISS [54]	None	22.3	37.4	47.2	10.6
DGM+LOMO [6]	None	24.7	39.4	47.0	11.7
OURS+LOMO	None	**29.2**	**44.3**	**50.5**	**12.2**
OIM [31]	None	33.7	48.1	54.8	13.5
UTM [13]	None	39.7	53.2	-	20.1
DGM+IDE [6]	None	48.1	64.7	71.1	29.2
DAL [10]	None	49.3	65.9	72.2	23.0
BUC [18]	None	61.1	75.1	80.0	38.0
TAULD [33]	None	43.8	59.9	72.8	29.1
UTAL [55]	None	49.9	66.4	77.8	35.2
TSSL [14]	None	56.3	-	-	30.5
CCE [25]	None	62.8	77.2	80.1	**43.6**
OURS	None	**65.3**	**77.8**	**81.3**	41.2
UGA [39]	Intra-camera	59.9	-	-	40.5
Prog. Learning [56]	One-shot	62.8	75.2	80.4	42.6
Stepwise [32]	One-shot	41.2	55.5	-	19.6
RACE [5]	One-shot	43.2	57.1	62.1	24.5
EUG [46]	One-shot	62.6	74.9	**82.5**	42.4
TCPL [57]	One-shot	65.2	77.5	-	**43.6**
OURS	None	**65.3**	**77.8**	81.3	41.2

None denotes fully unsupervised methods; One-shot assumes a singular labeled tracklet for each identity along with a large pool of unlabeled samples; Intra-camera setting works with labels which are provided only for samples within an individual camera view.

TABLE VI
RANK-1, -5, -10 ACCURACY (%) AND MAP (%) WITH SOME UNSUPERVISED AND ONE-SHOT SUPERVISED APPROACHES ON THE DUKEMTMC-VIDEOREID DATASET.

Methods	Labels	R1	R5	R10	mAP
OIM [31]	None	51.1	70.5	76.2	43.8
DGM+IDE [6]	None	42.3	57.9	69.3	33.6
TAUDL [33]	None	26.1	42.0	57.2	20.8
UTAL [55]	None	48.3	62.8	76.5	36.6
TSSL[14]	None	73.9	-	-	64.6
BUC [18]	None	69.2	81.1	85.8	61.9
CCE [25]	None	76.4	88.7	91.0	**69.3**
OURS	None	**76.5**	**89.6**	**91.9**	68.7
Stepwise [32]	One-shot	56.2	70.3	79.2	46.7
EUG [46]	One-shot	72.7	84.1	-	63.2
Prog. Learning [56]	One-shot	72.9	84.3	88.3	63.3
OURS	None	**76.5**	**89.6**	**91.9**	**68.7**

respectively.Compared with the recent CCE [25], our method improves the rank-1 accuracy from 62.8% to 65.3%. As expected, the proposed method performs better while using the deep CNN features compared to the handcrafted LOMO features. Moreover, from Table V, we observe that the proposed method is also very competitive while comparing with the semi-supervised methods without requiring any person identity information. Comparing to EUG (one-shot setting), our method achieves 2.7% improvement in rank-1 accuracy. It may be noted that any unsupervised feature embedding learning-based person re-id models [18], [25] can be used as our feature extractors, and the better the feature extractors are, the better our method performs.

TABLE VII
ABLATION STUDIES OF THE PROPOSED METHOD ON THE MARS AND DUKEMTMC-VIDEOREID DATASETS. RANK-1,-5,-10 ACCURACIES(%) AND MAP (%) ARE REPORTED

MARS	R1	R5	R10	mAP
Baseline	61.1	75.1	80.0	38.0
OURS w/ CM	64.6	76.7	80.7	39.8
OURS w/ CM+GNC	65.3	77.8	81.3	41.2
DukeMTMC-VideoReID	R1	R5	R10	mAP
Baseline	69.2	81.1	85.8	61.9
OURS w/ CM	72.2	86.2	89.3	64.3
OURS w/ CM+GNC	76.5	89.6	91.9	68.7

Baseline: Recognition performance is measured by directly using the Euclidean distance (ℓ_2 distance). w/ CM: Cross-view matching by introducing the graph matching into baseline. GNC: global network constraints.

DukeMTMC-VideoReID Dataset. We also evaluate our method on a larger video person re-id dataset - DukeMTMC-VideoReID dataset which is captured with 8 different cameras by comparing with several state-of-the-art methods such as OIM [31], DGM+ [6], Stepwise [32], EUG [46], Progressive Learning [56], BUC [18], TAULD [33], UTAL [55], TSSL [14] and CCE [25]. Results in Table VI shows the superiority of the proposed unsupervised framework over all the compared methods (unsupervised or one-shot supervised methods). We achieve the best recognition performance with rank-1 accuracy of 76.5% and mAP score of 68.7%, respectively. Comparing to BUC (unsupervised), the proposed method achieves 7.2% and 6.8% improvement for rank-1 accuracy and mAP score, respectively. Comparing EUG (one-shot setting), the recognition performance is improved from 72.7% to 76.5% for rank-1 accuracy and 63.2% to 68.7% for mAP score, respectively.

G. Ablation Studies

To better evaluate the effectiveness of our proposed method, we conduct ablation studies on the DukeMTMC-VideoReID dataset and MARS dataset. As shown in Table VII, Baseline denotes that we measure the recognition performance using the Euclidean distance (ℓ_2 distance) directly on the extracted features. We first show the effect caused by the cross-view matching (CM) via introducing the graph matching into the Baseline. On the DukeMTMC-VideoReID dataset, it can be seen that "Ours w/ CM" improves the recognition performance from 69.2% to 72.2% for rank-1 accuracy and 61.9% to 64.3% for mAP score, similarity, 3.5% and 1.8% improvement on the MARS dataset. This demonstrates that the cross-view matching is an effective way to improve the performance by exploiting the similarity relationships across camera pairs. Furthermore, we validate the effect caused by the global network constraints (GNC). As shown in Table VII, by introducing the GNC, the recognition performance is improved consistently. "Ours w/ CM+GNC" achieves the best recognition performance with rank-1 accuracy of 76.5% and mAP score of 68.7% on the DukeMTMC-VideoReID dataset, and 65.3% and 41.2% on the MARS dataset.

The visualization of intra-camera and inter-camera label estimation results with/without global network constraints on MARS is shown in Figure 6. It can be seen that the first neighbor-based clustering strategy is effective to group the similar samples in each camera. When introducing global network constraints, the outliers can be removed significantly. Note that we set $\theta = 1$ and iteration $= 0$ to show the results of cross-camera label estimation.

Fig. 6. Visualization of intra-camera and inter-camera label estimation with/without global network constraints on the MARS dataset. Samples in each box denote the intra-camera clustering results and $\mathcal{C}_r^i$ represents the ith cluster in camera r. The first and second rows show the cross-camera matching performance without and with global network constraints, respectively. Note that each image in this figure denotes one person tracklet and we illustrate the matching performance with one person/cluster.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a fully unsupervised consistent cross-view matching framework for video-based person re-identification. We first propose to use first neighbor of each sample to explore the correlations in each camera and then global camera network constraints are introduced into cross-view matches to reduce the wrong matches significantly. We then present a definition of cross-view matching reliability, which can be used to balance the performance between the quality and quantity of the estimated pairs. In addition, our consistent cross-view matching method is embedded into an iterative framework which iterates between the cross-camera label estimation and metric models learning. Rigorous experiments on two standard video person re-id datasets show the advantages of our approach over the state-of-the-art methods.

The proposed method learns metric models for camera pairs progressively, which relies on the pre-trained feature embedding models. In the future, we will try to design an end-to-end model that jointly optimizes the feature extractor and distance metric models by mining the consistent cross-camera pairs for person re-id task.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61771189 and the Hunan Provincial Natural Science Foundation of China under Grant No.2018JJ2060, in part by ONR grant N00014-19-1-2264 and NSF grant 1544969, and in part by China Scholarship Council.

REFERENCES

[1] A. K. Roy-Chowdhury and B. Song, "Camera networks: The acquisition and analysis of videos over wide areas," *Synthesis Lectures on Computer Vision*, vol. 3, no. 1, pp. 1–133, 2012.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[3] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4192–4205, 2019.

[4] Y. Rao, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition and person re-identification," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 701–718, 2019.

[5] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proceedings of European Conference on Computer Vision*, 2018, pp. 170–186.

[6] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2976–2990, 2019.

[7] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3958–3967.

[8] D. Ouyang, J. Shao, Y. Zhang, Y. Yang, and H. T. Shen, "Video-based person re-identification via self-paced learning and deep reinforcement learning framework," in *Proceedings of ACM International Conference on Multimedia*, 2018, pp. 1562–1570.

[9] M. Ye and P. C. Yuen, "Purifynet: A robust person re-identification model with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2655–2666, 2020.

[10] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification," in *Proceedings of the British Machine Vision Conference*, 2018.

[11] P. Li, P. Panb, P. Liuc, M. Xu, and Y. Yang, "Hierarchical temporal modeling with mutual distance matching for video based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[12] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5142–5150.

[13] C. Riachy, F. Khelifi, and A. Bouridane, "Video-based person re-identification using unsupervised tracklet matching," *IEEE Access*, vol. 7, pp. 20 596–20 606, 2019.

[14] G. Wu, X. Zhu, and S. Gong, "Tracklet self-supervised learning for unsupervised person re-identification." in *Proceedings of AAAI Conference on Artificial Intelligence*, 2020, pp. 12 362–12 369.

[15] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Transactions on Image Processing (TIP)*, 2020.

[16] X. Wang, S. Paul, D. S. Raychaudhuri, M. Liu, Y. Wang, A. K. Roy-Chowdhury *et al.*, "Learning person re-identification models from videos with weak supervision," *arXiv preprint arXiv:2007.10631*, 2020.

[17] M. Ye, J. Shen, X. Zhang, P. C. Yuen, and S.-F. Chang, "Augmentation invariant and instance spreading feature for softmax embedding," *IEEE TPAMI*, 2020.

[18] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 8738–8745.

[19] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 4, pp. 83:1–83:18, 2018.

[20] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2148–2157.

[21] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6112–6121.

[22] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8222–8231.

[23] Z. Zhang and V. Saligrama, "Prism: Person reidentification via structured matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 499–512, 2016.

[24] C.-T. Chu and J.-N. Hwang, "Fully unsupervised learning of camera link models for tracking humans across nonoverlapping cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 979–994, 2014.

[25] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3390–3399.

[26] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5771–5780.

[27] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *Proceedings of European Conference on Computer Vision*, 2014, pp. 330–345.

[28] A. Chakraborty, A. Das, and A. K. Roy-Chowdhury, "Network consistent data association," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1859–1871, 2015.

[29] S. Sarfraz, V. Sharma, and R. Stiefelhagen, "Efficient parameter-free clustering using first neighbor relations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8934–8943.

[30] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Transactions on Image Processing*, vol. 29, pp. 5481–5490, 2020.

[31] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3415–3424.

[32] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2429–2438.

[33] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 737–753.

[34] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157–5166.

[35] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 994–1003.

[36] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2005, pp. 26–33.

[37] J. Yan, M. Cho, H. Zha, X. Yang, and S. M. Chu, "Multi-graph matching via affinity optimization with graduated consistency regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 6, pp. 1228–1242, 2015.

[38] Z. Zhang, Q. Shi, J. McAuley, W. Wei, Y. Zhang, and A. Van Den Hengel, "Pairwise matching through max-weight bipartite belief propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1202–1210.

[39] J. Wu, Y. Yang, H. Liu, S. Liao, Z. Lei, and S. Z. Li, "Unsupervised graph association for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8321–8330.

[40] S. Roy, S. Paul, N. E. Young, and A. K. Roy-Chowdhury, "Exploiting transitivity for learning person re-identification models on a budget," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7064–7072.

[41] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[42] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3685–3693.

[43] Y. Tian, J. Yan, H. Zhang, Y. Zhang, X. Yang, and H. Zha, "On the convergence of graph matching: Graduated assignment revisited," in *Proceedings of European Conference on Computer Vision*, 2012, pp. 821–835.

[44] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[45] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 868–884.

[46] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5177–5186.

[47] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 17–35.

[48] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[49] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[50] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proceedings of International Conference on Knowledge Discovery and Data Mining*, vol. 96, no. 34, 1996, pp. 226–231.

[51] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2013, pp. 160–172.

[52] R. J. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 1, pp. 1–51, 2015.

[53] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised ℓ_1 graph learning," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 178–195.

[54] F. M. Khan and F. Bremond, "Unsupervised data association for metric learning in the context of multi-shot person re-identification," in *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2016, pp. 256–262.

[55] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1770–1782, 2019.

[56] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872–2881, 2019.

[57] D. S. Raychaudhuri and A. K. Roy-Chowdhury, "Exploiting temporal coherence for self-supervised one-shot video re-identification," *arXiv preprint arXiv:2007.11064*, 2020.

Xueping Wang is currently pursuing the Ph.D. degree with the College of Electrical and Information Engineering, Hunan University, China. His research interests include computer vision, person re-identification, and adversarial attack and defense methods.

Rameswar Panda graduated from University of California, Riverside with a Ph.D. in Electrical and Computer Engineering in 2018. Previously, he received his Bachelors and Masters degree from Biju Patanaik University of Technology, India and Jadavpur University, India. He is currently a researcher at IBM Research AI (MIT-IBM Watson AI Lab). His main research interests include computer vision, machine learning, video summarization, person re-identification and multimedia.

Min Liu is a professor at Hunan University. He received his bachelor degree from Peking University and Ph.D. degree in Electrical Engineering from the University of California, Riverside in 2012. He was a research intern in HHMI Janelia Farm Research Campus and a research scientist at the University of California, Santa Barbara. His research interests include computer vision and biomedical image analysis. Dr. Liu is an Associate Editor of BMC Bioinformatics.

Yaonan Wang received the Ph.D. degree in electrical engineering from Hunan University, Changsha, China, in 1994. Since 1995, he has been a Professor with the College of Electrical and Information Engineering, Hunan University. From 1994 to 1995, he was a Post-Doctoral Research Fellow with the Normal University of Defense Technology, Changsha. From 1998 to 2000, he was supported as a Senior Humboldt Fellow by the Federal Republic of Germany at the University of Bremen, Bremen, Germany. From 2001 to 2004, he was a Visiting Professor at the University of Bremen. He is a member of the Chinese Academy of Engineering. His research interests include robotics and image processing.

Amit K. Roy-Chowdhury received the Bachelors degree in Electrical Engineering from Jadavpur University, Calcutta, India, the Masters degree in Systems Science and Automation from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park. He is a Professor of Electrical and Computer Engineering and a Cooperating Faculty in the Department of Computer Science and Engineering, University of California, Riverside. His broad research interests include computer vision, image processing, and vision-based statistical learning, with applications in cyber-physical, autonomous and intelligent systems. He is a coauthor of two books: Camera Networks: The Acquisition and Analysis of Videos over Wide Areas, and Recognition of Humans and Their Activities Using Video. He is the editor of the book Distributed Video Sensor Networks. He has been on the organizing and program committees of multiple computer vision and image processing conferences and is serving on the editorial boards of multiple journals. He is a Fellow of the IEEE and IAPR.