

# Joint Person Re-identification and Camera Network Topology Inference in Multiple Cameras

Yeong-Jun Cho, Su-A Kim\*, Jae-Han Park\*, Kyuewang Lee, *Student Member, IEEE*  
and Kuk-Jin Yoon, *Member, IEEE*

**Abstract**—Person re-identification is the task of recognizing or identifying a person across multiple views in multi-camera networks. Although there has been much progress in person re-identification, person re-identification in large-scale multi-camera networks still remains a challenging task because of the large spatio-temporal uncertainty and high complexity due to a large number of cameras and people. To handle these difficulties, additional information such as camera network topology should be provided, which is also difficult to automatically estimate, unfortunately. In this study, we propose a unified framework which jointly solves both person re-identification and camera network topology inference problems with minimal prior knowledge about the environments. The proposed framework takes general multi-camera network environments into account and can be applied to online person re-identification in large-scale multi-camera networks. In addition, to effectively show the superiority of the proposed framework, we provide a new person re-identification dataset with full annotations, named **SLP**, captured in the multi-camera network consisting of nine non-overlapping cameras. Experimental results using our person re-identification and public datasets show that the proposed methods are promising for both person re-identification and camera topology inference tasks.

**Index Terms**—Person re-identification, Camera network topology inference, Non-overlapping camera network, Person re-identification dataset

## I. INTRODUCTION

Nowadays, a large number of surveillance cameras installed at public places (e.g. offices, stations, airports, and streets) are used to prevent untoward incidents, monitor localities, or to track down evidences. In this context, multi-camera networks have been studied to monitor larger areas or specific individuals. Person re-identification, studied extensively over the past decade, is the task of *automatically* recognizing and identifying a person using multiple views obtained from multi-camera networks. Nevertheless, the re-identification in large-scale multi-camera networks still remains a challenging task because of the large spatio-temporal ambiguity and high complexity due to the large numbers of cameras and people. It becomes even more challenging when camera views do

This work was done when all authors were in Gwangju Institute of Science and Technology. \* These authors contributed equally to this work. Yeong-Jun Cho, Jae-Han Park and Kuk-Jin Yoon are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea. Currently, Su-A Kim is with Intel Visual Computing Institute Saarland Informatics Campus, Germany and Kyuewang Lee is with ASRI, Department of Electrical and Computer Engineering, Seoul National University. Kuk-Jin Yoon is a corresponding author. E-mail: kjyoon@gist.ac.kr

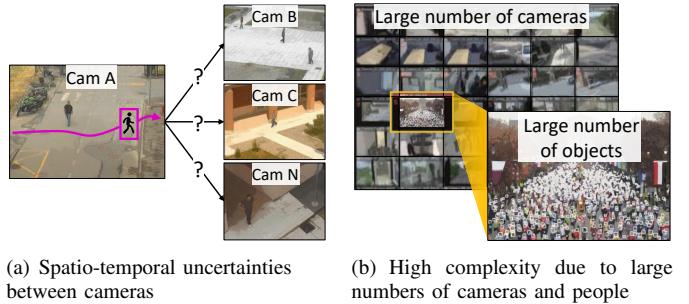


Fig. 1. Challenges of large-scale person re-identification

not overlap each other. As shown in Fig. 1 (a), spatio-temporal uncertainty due to the unknown geometrical relationship between cameras in a multi-camera network makes re-identification difficult. Unless some prior knowledge about the camera network is given, re-identification should be done by thoroughly matching the person of interest with all the other people who appear in the camera network within the given time interval. This exhaustive search method is slow in general and generates unsatisfactory results in many cases because it is hard to find a correct match among a large number of candidates, as there might be many people having appearances similar to the person of interest.

Most of the previous studies [1]–[3] conducted an exhaustive search to re-identify people across multiple cameras, relying solely on the appearance of people. These methods work quite well when the numbers of cameras and people are small, but cannot effectively handle the aforementioned challenges in the large-scale problem illustrated in Fig. 1 (b). To resolve the complexity problem and to improve the accuracy of the large-scale person re-identification, the number of matching candidates should be constrained and reduced by inferring and exploiting the spatio-temporal relation between cameras, referred to as the camera network topology.

In recent years, several camera network topology inference methods [4], [5] have been proposed for large-scale person re-identification. These methods infer the topology of a camera network based on the simple occurrence correlation between the people’s entry and exit events. However, no appearance-based validation is performed for topology inference. Hence, the inferred topology can be inaccurate in crowded scenes. Actually, it is a chicken-and-egg situation: the camera network topology can be more accurately inferred from the re-identification results while person re-identification can be im-

proved by utilizing the camera network topology information.

The main idea of this study is that the camera network topology inference and person re-identification can be achieved jointly while complementing each other. Based on this idea, we propose a unified framework, which automatically solves both the problems together. To the best of our knowledge, this is the first attempt to solve both the problems jointly. In the proposed framework, we first infer the initial camera network topology using only highly reliable re-identification results obtained by the proposed multi-shot matching method. This initial topology is used to improve the person re-identification results, and the improved re-identification results are then used to refine the camera network topology. This procedure is repeated until the estimated camera network topology converges. Once we estimate the reliable camera network topology in the initialization stage, we utilize it for the online person re-identification and update the camera network topology with time.

To sum up, we propose a multi-shot person re-identification method that exploits a time-efficient random forest. (Sec. III-A). We also propose a fast and accurate camera network topology inference method in Sec. III-B. In Sec. III-C, we describe a proposed online person re-identification method. The methods are effectively integrated in our framework. It is important to note that our proposed framework runs automatically with minimal prior knowledge about the environments. Besides the proposed methods, we also provide a new Synchronized Large-scale Person re-identification dataset named SLP (Sec. IV) captured in the multi-camera network consisting of nine non-overlapping cameras. To validate our unified framework, we extensively evaluate the performance and compare it with state-of-the-art methods.

Our previous study is [6]. Differences between the previous study and this study are as follows. In this study, we propose and validate online person re-identification after the camera network topology initialization stage (Sec. III-C). During the online person re-identification, the camera network topology updates itself with time. We describe complete details of the proposed methods and provide more experimental results to explicitly validate our proposed methods. In Sec. VI, we compare our proposed method with state-of-the-art methods in terms of both quantitative and qualitative performances, and show the results using another public dataset.

## II. PREVIOUS WORKS

Person re-identification methods can be categorized into non-contextual and contextual methods as summarized in [7]. In general, non-contextual methods rely only on appearances of people and measure visual similarities between people to establish correspondences, while contextual methods exploit additional contexts such as human pose, camera parameters, geometry, and camera network topology.

### A. Non-contextual Methods

In order to identify people across non-overlapping views, most studies rely on appearances of people by utilizing appearance-based matching methods with feature learning

or metric learning techniques. For feature learning, many studies [1], [8], [9] have tried to design visual descriptors to describe the appearance of people suitably. Regarding the metric learning, several methods such as KISSME [10] and LMNN-R [11] have been proposed and applied to the re-identification problem [2], [12], [13]. Some studies [10], [13] extensively evaluated and compared several metric learning methods and showed the effectiveness of the metric learning for re-identification. In order to utilize more plentiful appearance information, various multi-shot matching methods exploiting multiple appearances were proposed in [1], [14], [15]. In addition, person re-identification methods based on deep learning were proposed in [3], [16], which simultaneously learn both features and distance metrics.

Although many non-contextual methods have improved the performance of person re-identification, the challenges such as spatio-temporal uncertainty between non-overlapping cameras and high computational complexity still remain.

### B. Contextual Methods

Several studies [17]–[19] using human pose priors have been proposed recently. These methods calibrate cameras to obtain camera parameters and estimate human poses. Person re-identification with estimated human poses can then be efficiently carried out. The methods mitigate ambiguities by seeking various poses of people.

To resolve the spatio-temporal ambiguities, many studies have tried to employ camera network topology and camera geometry. Several studies [20]–[22] assume that the camera network topology information is available and show the effectiveness of the topological information. However, the topology is not available in the real-world scenario. Thus, many studies have tried to infer the camera network topology in an unsupervised manner. Makris *et al.* [4] proposed a topology inference method that simply observes the entry and exit events of the targets and measures the correlations between the events to establish the camera network topology. This method was applied and extended in [23]–[25]. Similarly, Loy *et al.* [5], [26] also proposed topology inference methods to understand multi-camera activity by measuring correlation or common information across simple activity patterns.

The aforementioned topology inference methods [4], [5], [24]–[26], so-called event-based approaches are practical since they do not require any appearance matching steps such as re-identification or inter-camera tracking for topology inference. However, the topology inferred by the event-based approach can be inaccurate since the topology may be inferred from false event correlations. Actually, false event correlation may easily occur when people pass through blind regions irregularly. On the other hand, a few studies [27], [28] inferred the topology using person re-identification results. These methods tend to be more robust to noise than the event-based approaches. However, they do not refine or update the initially inferred topology with time while performing re-identification. Thus, these methods may not fully exploit the topology information.

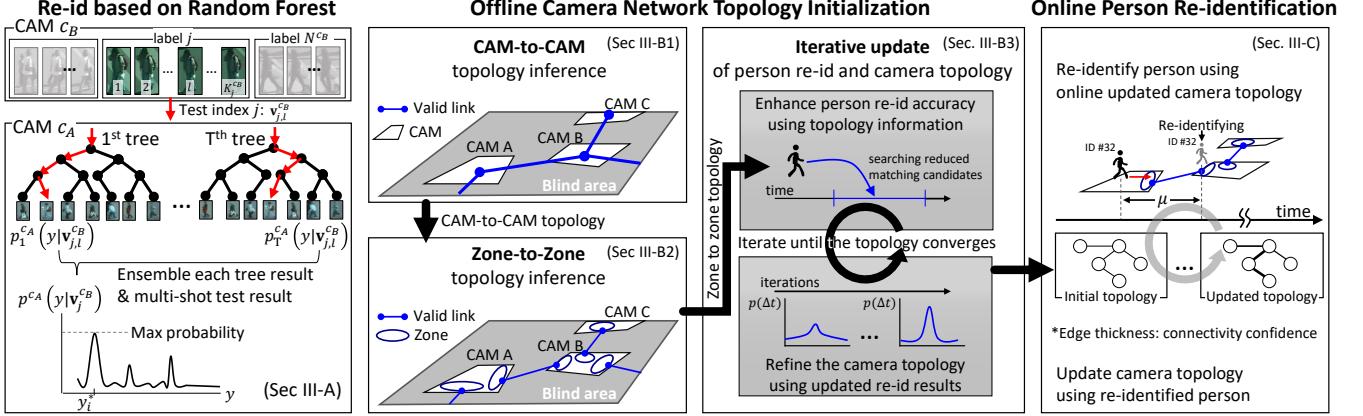


Fig. 2. The proposed unified framework for person re-identification and camera network topology inference.

### III. PROPOSED UNIFIED FRAMEWORK

Figure 2 illustrates the proposed unified framework for person re-identification and camera network topology inference. In the proposed framework, we first train random forest-based person classifiers (Sec. III-A) for efficient person re-identification. Subsequently, we jointly estimate and refine the camera network topology and person re-identification results (Sec. III-B) using the trained random forests. Finally, we perform an online person re-identification using the inferred camera topology (Sec. III-C).

#### A. Random Forest-based Person Re-identification

Most of the previous studies mainly focused on enhancing the re-identification performance. However, when handling a large number of people, time complexity is also very important for building a practical re-identification framework. To this end, we utilize a random forest algorithm [29] for multi-shot person re-identification and incorporate it into our framework. We denote the  $k$ -th appearance of person  $i$  in the camera  $c_A$  as  $\mathbf{v}_{i,k}^{c_A}$ . A set of the appearances of people in the camera is expressed as,

$$\mathcal{D}^{c_A} = \left\{ \left( \mathbf{v}_{i,k}^{c_A}, y_i \right) | 1 \leq i \leq N^{c_A}, 1 \leq k \leq K_i^{c_A} \right\}, \quad (1)$$

where  $y_i$  is the label of person  $i$  (normally  $y_i$  is set to  $i$ ).  $N^{c_A}$  is the number of people visible in camera  $c_A$  and  $K_i^{c_A}$  is the number of appearances of person  $i$  in camera  $c_A$ . We then train a random forest classifier using the appearance set  $\mathcal{D}^{c_A}$ . After the random forest classifier is trained, we can estimate the probability distribution of classification by aggregating decision tree outputs as:  $p^{c_A}(y|\mathbf{v}) = \frac{1}{T} \sum_{\tau=1}^T p_\tau^{c_A}(y|\mathbf{v})$ , where  $p_\tau^{c_A}$ ,  $T$  denote a trained decision tree and the number of decision trees, respectively. Testing a person  $j$  (i.e. probe) in another camera  $c_B$  to the trained random forest classifier is defined as,

$$p^{c_A} \left( y | \mathbf{v}_{j,l}^{c_B} \right) = \frac{1}{T} \sum_{\tau=1}^T p_\tau^{c_A} \left( y | \mathbf{v}_{j,l}^{c_B} \right). \quad (2)$$

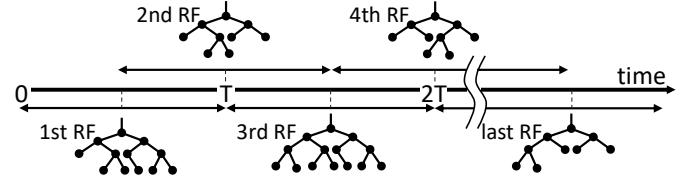


Fig. 3. Training a series of random forest classifiers based on the overlapping time window  $T$ .

To obtain a multi-shot person re-identification result, we test multiple appearances of each person and average the multiple results as

$$p^{c_A} (y | \mathbf{v}_j^{c_B}) = \frac{1}{K_j^{c_B}} \sum_{l=1}^{K_j^{c_B}} p^{c_A} (y | \mathbf{v}_{j,l}^{c_B}), \quad (3)$$

where  $K_j^{c_B}$  is the number of appearances of the probe. From the probability distribution  $p^{c_A} (y | \mathbf{v}_j^{c_B})$ , we choose a final matched label as:

$$y_i^* = \operatorname{argmax}_{y_i \in \{1, 2, \dots, N^{c_A}\}} p^{c_A} (y_i | \mathbf{v}_j^{c_B}). \quad (4)$$

As the result of the multiple appearance matching test, we obtain a corresponding pair  $(\mathbf{v}_{y_i^*}^{c_A}, \mathbf{v}_j^{c_B})$  between the camera  $c_A$  and  $c_B$ . Finally, we calculate a similarity score for the matching pair by selecting the smallest matching score as per the following equation [1],

$$S \left( \mathbf{v}_{y_i^*}^{c_A}, \mathbf{v}_j^{c_B} \right) = e^{-\min_{k,l} \left\| \mathbf{v}_{y_i^*,k}^{c_A} - \mathbf{v}_{j,l}^{c_B} \right\|_2}. \quad (5)$$

The similarity score lies on  $[0, 1]$  and it is used in Sec. III-B for inferring the topology.

The tree structure of the random forest method makes the multi-shot test very fast: its computational complexity is  $O(NK \log N)$ <sup>1</sup>. On the other hand, the conventional exhaustive re-identification approaches comparing every person pair

<sup>1</sup>When analyzing the computational complexity, we assume that the number of objects in  $c_A$  and  $c_B$  are the same ( $i, j = \{1, 2, \dots, N\}$ ) and each object contains the same number of image patches ( $k, l = \{1, 2, \dots, K\}$ ). Details are described in APPENDIX.

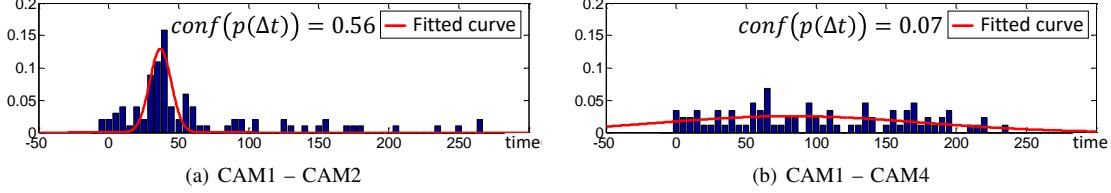


Fig. 4. Examples of estimated transition distributions with connectivity confidences

involve higher computational cost:  $O(N^2K^2)$ . Over and above the superiority in terms of the computational cost, our method also delivers a high person re-identification accuracy as shown in Sec. VI-A1.

### B. Camera Network Topology Initialization

Camera network topology represents spatio-temporal relations and connections between the cameras in the network. It involves the inter-camera transition distributions between two cameras, which denote the transition distributions of objects across cameras according to time, and represent the strength of connectivity between the cameras. In general, the topology is represented as a graph  $G = (V, E)$ , where the vertices  $V$  denote the cameras and the edges  $E$  denote the inter-camera transition distributions, as shown in Fig. 8 (b).

*1) CAM-to-CAM topology inference:* To build the CAM-to-CAM topology, we first estimate the transition distributions between the cameras. In this study, we estimate the transition distributions based on the person re-identification results. Before performing person re-identification, we select key appearances among the multiple appearances of each person based on the method in [30]. By selecting key appearances uncorrelated with each other, we can efficiently reduce the matching complexity and avoid over-fitting problems.

After the key appearance selection, we split a whole group of people into several sub-groups according to their timestamps and train a series of random forest classifiers with overlapping time window  $T$  as shown in Fig. 3. Next, we search correspondences of people who disappeared in a camera using the trained random forest classifiers of other cameras. Initially, we have no transition distributions between cameras to utilize. Hence, we consider every pair of cameras in the camera network and search for correspondences within a wide time interval. When a person disappears at time  $t$  in a certain camera, we search the correspondence of the person from the other cameras within time range  $[t - T, t + T]$ . If there are multiple random forest classifiers overlapping with the time range, we test all random forest classifiers and select the most reliable one among them as a correspondence. Thanks to our overlapping search scheme, we can alleviate the risk of failure in correspondence search.

When these initial correspondences are given, we only use highly reliable correspondences to infer transition distributions between cameras. We regard a correspondence as a reliable one with a high similarity score when  $S(\mathbf{v}_{y_i^*}^{CA}, \mathbf{v}_j^{CB}) > \theta_{sim}$ . The inference procedure of transition distribution is as follows: (1) Calculate the time difference between correspondences and make a histogram of the time difference. (2) Normalize the

histogram using the total number of reliable correspondences. We denote the transition distribution as  $p(\Delta t)$ . Figure 4 shows two distributions: Fig. 4 (a) is obtained from a pair of cameras having a strong connection, and Fig. 4 (b) from a pair of cameras having a weak or no connection.

**Connectivity Check** Based on the estimated transition distributions, we automatically identify whether a pair of cameras is connected or not. We assume that the transition distribution follows a normal distribution if there is a topological connection. Based on this assumption, we fit a Gaussian model  $N(\mu, \sigma^2)$  to the distribution  $p(\Delta t)$  and measure the connectivity of a pair of cameras based on the following observations:

- **Variance of  $p(\Delta t)$ :** In general, most of the people reappear after a certain transition time  $\mu$ ; therefore, the variance in the transition distribution ( $\sigma^2$ ) is not extremely large and the distribution shows a clear peak.
  - **Fitting error:** Even though the distribution is obtained from a pair of cameras with a weak connection, the variance in the distribution can be small and the distribution can have a clear peak due to noise. In order to measure the connectivity robust to noise, we consider the model fitting error  $\mathcal{E}(p(\Delta t)) \in [0, 1]$  calculated by R-squared statistics.

Based on the above observations, we newly define a connectivity confidence between a pair of cameras as

$$conf(p(\Delta t)) = e^{-\sigma} \cdot (1 - \mathcal{E}(p(\Delta t))). \quad (6)$$

The connectivity confidence lies on  $[0, 1]$ . We regard a pair of cameras as a valid link when  $\text{conf}(p(\Delta t)) > \theta_{\text{conf}}$ . Compared to a previous method [4], which only takes the variance of a distribution into account, our method is more robust to noise of distributions. Using the defined confidences, we check every pair of cameras and reject invalid links as shown in Fig. 8. We can see that many camera pairs in the camera network have weak connections. Therefore, we can greatly reduce computation time and save resources. Only the valid pairs of cameras proceed to the next step.

Through this topology inference procedure, CAM-to-CAM topology is represented as

$$\begin{aligned} G_{cam} &= (V_{cam}, E_{cam}), \\ V_{cam} &\in \{c_i | 1 \leq i \leq N_{cam}\}, \\ E_{cam} &\in \{p^{ij}(\Delta t) | 1 \leq i \leq N_{cam}, 1 \leq j \leq N_{cam}, i \neq j\}, \end{aligned} \quad (7)$$

where  $N_{cam}$  is the number of cameras in the camera network,  $c_i$  is a  $i^{th}$  camera, and  $p^{ij}(\Delta t)$  is a transition distribution between  $c_i$  and  $c_j$ .

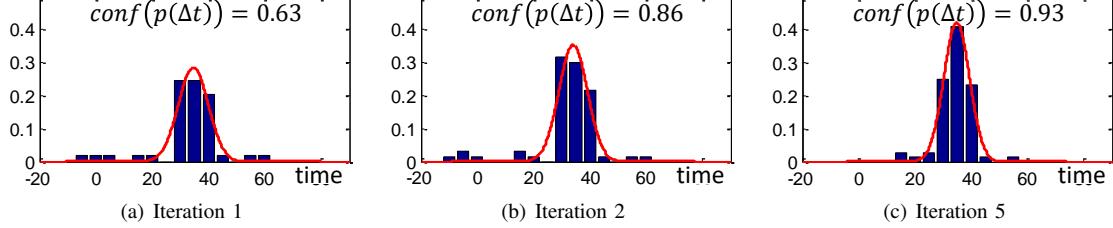


Fig. 5. Iterative update of the topology and re-identification results. Re-id accuracy is improved from 23.5% to 66.31%.

2) *Zone-to-Zone topology inference*: In this step, we estimate transition distributions between zones in cameras and build a Zone-to-Zone topology. For each camera, a set of entry and exit zones is automatically learned by [31]. Note that we only consider exit-to-entry zone pairs when two zones belong to different cameras. Other pairs of zones such as exit-to-exit, entry-to-entry, and entry-to-exit are not considered.

A person who disappeared at an exit zone at time  $t$  is likely to appear at an entry zone in a different camera within a certain time interval  $T$ . Therefore, we search the correspondence of the disappeared person from the entry zones in different cameras within the time range  $[t, t + T]$ . Similarly, as in Sec. III-B1, we train a series of random forest classifiers for each entry zone and measure connectivity confidences of all possible pairs of zones using only reliable correspondences. Through this step, many invalid pairs of zones between the cameras can be ignored, and we can formulate Zone-to-Zone topology as:

$$\begin{aligned} G_{\text{zone}} &= (V_{\text{zone}}, E_{\text{zone}}), \\ V_{\text{zone}} &\in \{c_{i(k)} | 1 \leq i \leq N_{\text{cam}}, 1 \leq k \leq Z_i\}, \\ E_{\text{zone}} &\in \{p_{(kl)}^{ij}(\Delta t) | 1 \leq i \leq N_{\text{cam}}, 1 \leq j \leq N_{\text{cam}}, \\ &\quad i \neq j, 1 \leq k \leq Z_i, 1 \leq l \leq Z_j\} \end{aligned} \quad (8)$$

where  $N_{\text{cam}}$  is the number of cameras in the camera network,  $Z_i$  is the number of zones of  $i^{\text{th}}$  camera.  $c_{i(k)}$  is a  $k^{\text{th}}$  zone of the  $i^{\text{th}}$  camera.  $p_{(kl)}^{ij}(\Delta t)$  is a transition distribution between  $c_{i(k)}$  and  $c_{j(l)}$ . In the next section, we iteratively update the valid links between zones and build a camera topology map of the camera network.

3) *Iterative update of person re-identification and camera network topology*: After conducting the Zone-to-Zone topology inference, we obtain an initial topology map between every pair of zones in the camera network. However, the initial topology map can be inaccurate, because it is inferred from noisy initial re-identification results. Although we consider only highly reliable re-identification results to infer the initial topology, the re-identification results are erroneous because we cannot exploit the topology information initially for re-identification.

As mentioned above, camera network topology information and re-identification results can complement each other. The inferred topology of the camera network can enhance the person re-identification performance, and the person re-identification can assist the topology inference. Therefore, we update person re-identification results and the camera network topology in an iterative manner. Assuming that we have a

---

#### Algorithm 1: Camera network topology initialization

---

**Data:** Image sequences and object identities in each view during initialization stage  
**Result:** Re-identification result (people correspondences), Camera network topology

```

for every camera pair do
     $G_{\text{cam}} = \text{CAM-to-CAM topology inference};$ 
    if  $\text{conf}(p^{ij}(\Delta t)) > \theta_{\text{conf}}$  then
         $G_{\text{zone}} = \text{Zone-to-Zone topology inference};$ 
        if  $\text{conf}(p_{(kl)}^{ij}(\Delta t)) > \theta_{\text{conf}}$  then
            while  $p_{(kl)}^{ij}(\Delta t)$  converging do
                 $T_{(kl)}^{ij} = \frac{1}{1 - \mathcal{E}(p_{(kl)}^{ij}(\Delta t))} (T_{(kl)U}^{ij} - T_{(kl)L}^{ij})$ 
                 $G_{\text{zone}} = \text{Topology update};$ 
            end
        end
    end
end

```

---

transition distribution  $p(\Delta t) \sim N(\mu, \sigma^2)$  between two zones. Each zone belongs to different cameras. Then, the iterative update procedure is as follows:

- **Step 1.** Update the time window  $T$ . The initial time window  $T$  was set quite wide, but now we can narrow it down based on the inferred topology. First, we find the lower and upper time bounds  $(T_L, T_U)$  of the transition distribution  $p(\Delta t)$  with a constant  $R$  as,

$$p(T_L \leq \Delta t \leq T_U) = \frac{R}{100}. \quad (9)$$

We set  $R$  as 95 following 3-sigma rule in order to cover a majority of the distribution (95%) and ignore some outliers (5%). Then, using the time bounds obtained, the time window  $T$  is updated as,

$$T = \frac{1}{1 - \mathcal{E}(p(\Delta t))} (T_U - T_L), \quad (10)$$

where  $\mathcal{E}(p(\Delta t))$  is a Gaussian model fitting error rate. When the fitting error is high, the time window  $T$  becomes large. Thanks to our update strategy, we can avoid the over-fitting of the topology during the update steps.

---

**Algorithm 2:** Online person re-identification and camera network topology update

---

**Data:** Image sequences, Object identities in each view, Camera network topology  
**Result:** Re-identification result (people correspondences), Camera network topology

```

for every zone pair do
    if  $\text{conf}\left(p_{(kl)}^{ij}(\Delta t)\right) > \theta_{\text{conf}}$  then
        Find correspondences;
         $p_{(kl)}^{ij}(\Delta t)$  = Transition distribution update;

        if  $\sum_{\Delta t} \left( |p_{(kl)}^{ij}(\Delta t) - p_{(kl)}^{\prime ij}(\Delta t)| \right) > 0.1$  then
            % Fit Gaussian model to  $p_{(kl)}^{ij}(\Delta t)$ 
             $p_{(kl)}^{ij}(\Delta t) = p_{(kl)}^{\prime ij}(\Delta t) \sim N(\mu, \sigma^2)$ ;
        end
    end
end

```

---

- **Step 2.** Re-train a series of random forests of an entry zone with the updated time window  $T$  as shown in the Fig. 3.
- **Step 3.** Find correspondences of the disappeared people at an exit zone. Based on the topology, a person who disappeared at time  $t$  at the exit zone is expected to appear around the time  $(t + \mu)$  at the entry zone of the other camera. Using the topological information, we search the correspondence of the person from a trained random forest having the center of time slot close to  $(t + \mu)$ .
- **Step 4.** Update the transition distribution using reliable correspondences with a high similarity score  $S(\mathbf{v}_{y_i^*}^{c_A}, \mathbf{v}_j^{c_B}) > \theta_{\text{sim}}$ .

This procedure (Step 1 – Step 4) is repeated until the transition distribution converges. We perform the procedure for all Zone-to-Zone pairs in the camera network topology  $G_{\text{zone}}$ . Consequently, the camera topology map in the camera network is refined. If the topology does not change anymore or shows subtle changes during several iterations, we stop the procedure. The above procedure improves the performance of re-identification as well as the accuracy of the topology inference, as shown in Fig. 5. The overall algorithm of the proposed camera network topology initialization process is summarized in Algorithm. 1.

### C. Online Person Re-identification and Camera Network Topology Update

Once we estimate the reliable camera network topology in the initialization stage, we utilize it for the online person re-identification and online camera network topology update. The proposed online process is as follows:

- **Step 1.** Pick out matching candidates of a person based on the initialized camera network topology. When a person disappeared at a time  $t$ , a set of people appearing at a different camera within  $t + T_L \leq t + \mu \leq t + T_U$  are the matching candidates. The lower and upper time bounds

TABLE I  
 LIST OF PUBLIC PERSON RE-IDENTIFICATION DATASETS. WE REFERRED TO THE LIST OF PERSON RE-IDENTIFICATION DATASETS IN [32].

dataset	year	# IDs	# cam	multi-shot	back ground	traj.	sync
VIPeR [33]	2007	632	2				
QMUL iLIDS [34]	2009	119	2	✓			
GRID [5]	2009	250	8				
CAVIAR4ReID [35]	2011	72	2	✓			
3DPeS [36]	2011	192	8	✓	✓		
PRID2011 [37]	2011	934	2	✓	✓	✓	
V4T [38]	2011	47	2	✓			
WARD [39]	2012	70	4	✓	✓	✓	
Softbio [40]	2012	150	8	✓	✓	✓	
CUHK 2 [41]	2013	1816	2	✓			
CUHK 3 [42]	2014	1360	2	✓			
NLPR MCT [43]	2014	235	3	✓	✓	✓	✓
		255	3				
RAID [44]	2014	43	4	✓			
iLIDS-VID [14]	2014	300	2	✓		✓	
Market1501 [45]	2015	1501	6	✓			
PRW [46]	2016	932	6	✓	✓		
MARS [47]	2016	1261	6	✓		✓	
DukeMTMC4ReID [48], [49]	2017	1852	8	✓			✓
Airport [32]	2017	9651	6	✓			
<b>SLP</b>	2017	2632	9	✓	✓	✓	✓

$(T_L, T_U)$  are evaluated using the method described in Sec. III-B3.

- **Step 2.** Find a correspondence among the matching candidates. For efficiency, we only train random forest classifiers where the number of matching candidates is large unless we find the correspondence based on exhaustive search.
- **Step 3.** Update the transition distribution. We use the matching result with a high similarity score ( $> \theta_{\text{sim}}$ ) to update the transition distribution, adding the new matching result to  $p(\Delta t)$  and normalizing the distribution. The updated transition distribution is represented as  $p'(\Delta t)$ .
- **Step 4.** Fit a new Gaussian model to the updated transition distribution  $p'(\Delta t)$ . The cost of model fitting is quite high. Hence, we do not always fit the Gaussian model, but only when the difference between the initial distribution  $p(\Delta t)$  and the updated distribution  $p'(\Delta t)$  is sufficiently large. To this end, we measure the distribution difference as  $\sum_{\Delta t} (|p(\Delta t) - p'(\Delta t)|)$  and fit the model when the value is larger than 0.1 (i.e., 10 % variation). Unless the value is larger than the threshold,  $p'(\Delta t)$  merely cumulates the update.

We perform the aforementioned procedure for all Zone-to-Zone pairs in  $G_{\text{zone}}$ . The overall algorithm of the proposed online process is summarized in Algorithm. 2.

The online topology update step is necessary and useful for the following reasons: First, the pose of a camera can change due to vibration, heavy wind, and so on. Even a slight change in pose can cause a considerable viewpoint change, and it makes the initialized topology inaccurate. Second, people transition time between cameras can change temporarily due to many factors such as sidewalk conditions, weather, and unusual movement of crowd (e.g. parade, march). The online topology update can effectively handle the above issues.

TABLE II  
DETAILS OF OUR NEW DATASET: SLP.

Index	CAM 1	CAM 2	CAM 3	CAM 4	CAM 5	CAM 6	CAM 7	CAM 8	CAM 9	Total
# ID	256	661	1,175	243	817	324	516	711	641	2,632
# frames	19,545	65,518	104,639	41,824	78,917	79,974	93,978	53,621	42,347	580,363
# annotated box	47,870	205,003	310,262	65,732	307,156	160,367	78,259	176,406	117,087	1,468,142
Resolution	720×480	1920×1080	1920×1080	1440×1080	1440×1080	1440×1080	1920×1080	1920×1080	1920×1080	–
Duration	2h 13m	2h 12m	2h 22m	2h	2h 21m	2h	2h 38m	2h 29m	2h 28m	–

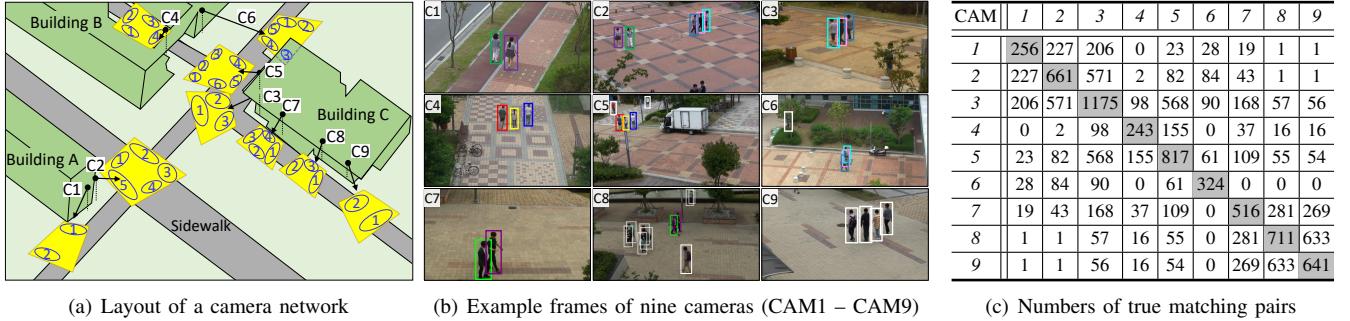


Fig. 6. A new synchronized large-scale person re-identification dataset: SLP.

Thanks to the accurate topology initialization (Sec. III-B) and the online procedure (Sec. III-C), the camera network topology can be maintained automatically with time. When a new camera is added to an existing camera network during the online stage, the camera network topology initialization operates in parallel and the topology is updated efficiently.

#### IV. NEW PERSON RE-IDENTIFICATION DATASET: SLP

To validate the performance of person re-identification methods, numerous datasets have been published. Table. I shows a list of the public person re-identification datasets with their various characteristics, which have been constructed for specific scenarios. For example, VIPeR [33] was constructed by two cameras and contains 632 people. Each camera provides one single image for one person. On the contrary, Softbio [40] includes 150 people captured from eight cameras. In this dataset, each camera provides multiple images for a person (*multi-shot*). However, despite the availability of numerous published datasets, none of them reflect practical large-scale surveillance scenarios, in which (1) multiple observations (*multi-shot*) and moving paths (*traj.*) for each person from each camera are available, and (2) both the number of people (# IDs) and the cameras (# cam) is large.

Most of the public datasets include a small number of people (# IDs < 200) [34]–[36], [38]–[40], [44] or cameras (# cam < 5) [14], [33]–[35], [37], [38], [41]–[44] as shown in Table. I. Some datasets provide just a single shot of each person [5], [33]. Moreover, many datasets do not provide the required annotation with information about people (*traj.*) such as their movement paths [32], [34]–[36], [38], [41], [42], [45], [46] but only give cropped multi-shot appearances. Furthermore, there are only two datasets that provide camera synchronization information or time stamps of all frames (*sync*) [43], [48]. Recently, M. Gou *et al.* [48] provided an re-identification dataset named DukeMTMC4ReID. The dataset covers real

large-scale surveillance scenarios (e.g., synchronized cameras, large number of cameras and people). However, it does not give the trajectory of each person but only gives cropped multi-shot images. If the dataset provides additional information such as *background* and *traj.*, it will be very helpful for other researchers and can be fully exploited.

In this study, we provide a new Synchronized Large-scale Person re-identification dataset called SLP constructed for practical large-scale surveillance scenarios. The main characteristics of our dataset are as follows:

- **The number of people (# IDs):** The total number of people in the dataset is 2,632. Among them, 1,737 people appear in at least two cameras. The details of the matching pairs are shown in Fig. 6 (c).
- **The number of cameras (# cam):** The dataset was captured from nine non-overlapping cameras. The layout of the camera network and the example frames from nine cameras are shown in Fig. 6 (a-b).
- **Multiple appearances availability (multi-shot):** SLP provides multiple appearances of each person. In this version, we do not provide the original image appearances but provide pre-computed feature descriptors due to legal restrictions. However, we expect that we will provide the original image appearances of people in the future.
- **Scene background availability (background):** SLP provides the scene background. Although it only contains the background of the camera scene, it contains scene structures such as buildings, ground plane, orthogonal lines etc., which are quite useful for understanding the scene.
- **People trajectory availability (traj.):** The ground truth tracking information of every person is available. It includes positions (x,y locations) and sizes (height, width) of people in entire image sequences. To annotate and generate people's trajectories for all video sequences, we

used an annotation tool ‘VATIC’ [50]. The total number of annotated boxes is 1,468,142.

- **Camera synchronization (sync):** Each camera is synchronized with a global timestamp. It provides all timestamps of video frames.
- **Camera calibration:** Each camera is calibrated and camera parameters are given.
- Details of our dataset are summarized in Table II.

As mentioned above, our new dataset enables us to test person re-identification methods under practical large-scale surveillance scenarios. Compared to previous datasets, SLP not only provides competitive number of person *IDs* and cameras but also additional information such as full annotations, camera synchronization and camera parameters. In addition, SLP is also applicable to other computer vision tasks such as multi-object tracking, multi-camera activity recognition, and so on. It is available online: [https://sites.google.com/view/yjcho/project-pages/re-id\\_topology](https://sites.google.com/view/yjcho/project-pages/re-id_topology).

## V. SETTINGS AND METHODOLOGY

### A. Experimental Settings

Since we mainly focus on person re-identification and camera topology inference problems, we assume that person detection and tracking results are given. Most person re-identification researches follow this assumption and setting. In order to evaluate each proposed method, we divided our dataset into two subsets according to time. The first subset contains 1-hour data starting from the global start time (AM 11:20). It is used in an initialization stage. The latter subset including the remaining data is utilized in an online test stage.

In this study, every person image is resized to  $128 \times 48$  pixels. We used the LOMO feature [51] to describe the appearances of people. It first divides the image into six horizontal stripes. Then it extracts an HSV color histogram and builds descriptors based on Scale Invariant Local Ternary Pattern (SILTP) from each stripe. The final descriptor from the  $128 \times 48$  image has 26,960 dimensions.

It is worthy to note that the proposed framework does not utilize pre-matched results between cameras. In general, many conventional person re-identification methods exploit the pre-matched pairs for learning distance metrics such as LMNN [12], ITML [2], KISSME [10] and Mahal [13]. The metric learning methods are effective but not appropriate for practical re-identification systems since the matched results are not given at first. On the other hand, the proposed framework runs fully automatically, providing the person detection/tracking results of each camera.

### B. Evaluation Methodology

In general, many previous works plot a Cumulative Match Curve (CMC) [33] representing true match being found within the first  $n$  ranks for evaluating person re-identification performances. In practice, however, all ranks ( $2, 3, \dots, n$ ) except rank-1 failed to find the correct correspondences. For this reason, rank-1 value is the most important one among all ranks and several studies mainly measure the rank-1 accuracy. In this

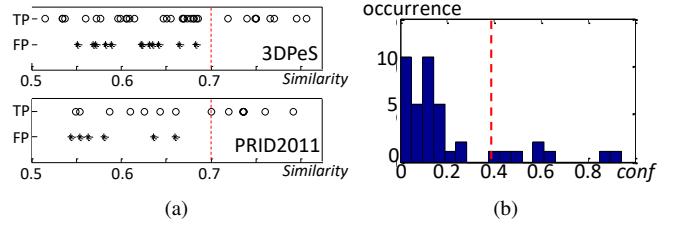


Fig. 7. Observations for selecting reliable parameters

study, we measure the *rank-1 accuracy* by  $\frac{TP}{T_{gt}}$ , where  $TP$  is the number of true matching results and  $T_{gt}$  is the total number of ground truth pairs in the camera network. The evaluation metric *rank-1 accuracy* reflects the overall re-identification performance of the camera network.

To evaluate the accuracy of the camera network topology, we adopt two evaluation metrics: *transition time error* and *topology distance*. When an inferred transition distribution and a ground truth are given as  $(p(\Delta t) \sim N(\mu, \sigma^2), p_{gt}(\Delta t) \sim N(\mu_{gt}, \sigma_{gt}^2))$ , the *transition time error* is simply measured as  $|\mu - \mu_{gt}|$ . We defined the *topology distance* as Bhattacharyya distance that measures the difference between two probability distributions as  $d_B(p, p_{gt}) = -\ln \left( \int \sqrt{p(\Delta t)p_{gt}(\Delta t)} d\Delta t \right)$ . If there are multiple links in the camera network, we measure the evaluation metrics for all links and average them to get the final *transition time error* and *topology distance*.

### C. Parameters Selection

We used fixed parameters for all steps in the Sec. III and applied the same parameters to all datasets SLP, NLPR MCT [43]. We found that the performance of our framework does not depend on each parameter. The parameters are fixed as follows:

- $\theta_{sim}$ : To determine a proper  $\theta_{sim}$ , we tested several other datasets (3DPeS [36], PRID [37]) and analyzed the results. Figure 7 (a) shows re-identification similarity scores of each dataset. Marks  $\circ$ ,  $*$  denote true and false positives. To estimate the accurate topology, we need to exclude false positives. Based on the test results, we strictly set  $\theta_{sim}$  as 0.7 for selecting only true positives.
- $\theta_{conf}$ : Figure 7 (b) shows a histogram of calculated connectivity confidences between nine cameras (36 pairs) in the CAM-to-CAM topology inference step. Weak connections are easily classified, since they have very low confidences (0 – 0.25). Based on this, we set  $\theta_{conf}$  as 0.4 and fixed the value for topology inference steps and all datasets.
- Initial  $T$ : We tested various  $T$  such as 300, 450, 600, which are sufficiently wide to find correspondences under unknown camera networks. We noticed that the results of topology inference are not significantly affected by  $T$ . We set the initial  $T$  as 600.
- \* The proposed framework has only one parameter ( $T$ ) to be adjusted depending on the environments. However,  $T$  does not critically affect the performance of the proposed system.

TABLE III  
INFERRED VALID ZONE-TO-ZONE LINKS AND GROUND-TRUTHS.

Exit	Entry	$\mu$	$\mu_{gt}$	$\sigma$	$\sigma_{gt}$	Exit	Entry	$\mu$	$\mu_{gt}$	$\sigma$	$\sigma_{gt}$
CAM1-ZONE1	CAM2-ZONE5	34.4	34.7	6.25	6.04	CAM2-ZONE5	CAM1-ZONE1	40.4	40.4	7.62	5.93
CAM2-ZONE2	CAM3-ZONE1	36.7	36.3	8.03	5.79	CAM3-ZONE1	CAM2-ZONE2	37.6	37.0	10.3	8.90
CAM3-ZONE2	CAM5-ZONE6	-0.42	-0.57	3.49	3.23	CAM5-ZONE6	CAM3-ZONE2	0.70	1.59	3.43	2.32
CAM3-ZONE3	CAM7-ZONE3	4.8	4.3	4.8	3.5	CAM7-ZONE3	CAM3-ZONE3	3.75	4.68	2.16	3.04
CAM4-ZONE4	CAM5-ZONE2	30.2	30.1	13.4	12.5	CAM5-ZONE2	CAM4-ZONE4	39.5	28.6	3.82	14.8
CAM7-ZONE1	CAM8-ZONE2	28.2	28.4	21.3	6.36	CAM8-ZONE2	CAM7-ZONE1	31.9	30.0	2.41	4.02
CAM8-ZONE1	CAM9-ZONE2	11.6	11.7	4.82	4.24	CAM9-ZONE2	CAM8-ZONE1	10.5	10.5	4.03	4.08

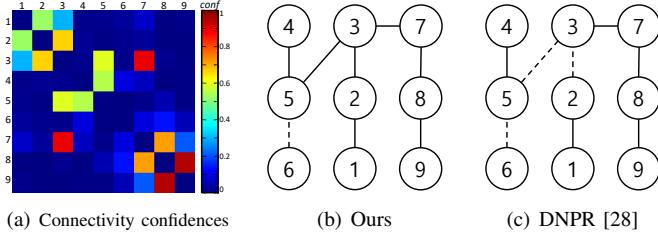


Fig. 8. Results of CAM-to-CAM connectivity check. (a) Connectivity confidence map. (b), (c) Valid camera links of two methods (*solid line*: true link, *dotted line*: missing).

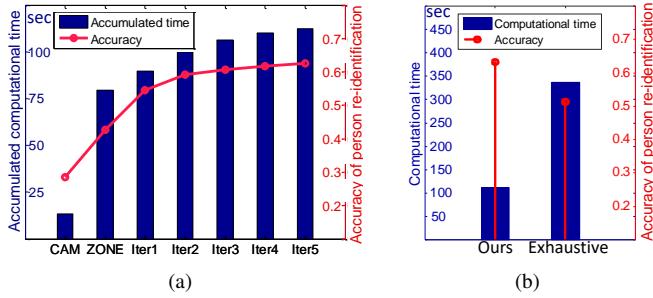


Fig. 9. Results of the person re-identification through the camera topology initialization.

## VI. EXPERIMENTAL RESULTS

### A. Evaluations using Our Dataset: SLP

1) *Camera Network Topology Initialization*: As explained in Sec. III, we first inferred the CAM-to-CAM topology between all cameras. We then found the Zone-to-Zone topology and iteratively updated the topology for the topology initialization. During the topology initialization process, we evaluate both tasks: person re-identification and topology inference.

### CAM-to-CAM topology inference result

In order to check the connectivity between two cameras, person re-identification can be carried out bi-directionally. However, for the sake of efficiency, we only used unidirectional re-identification results to check the connectivities between the cameras. To this end, we assigned a camera containing more people as a training camera (gallery) and a camera containing fewer people as a testing camera (query). That is because the problem of reidentifying a small set of people from a large set is easier than the opposite case. In addition, we can reduce the computational cost for checking the connectivity by half.

For all the camera pairs, we illustrate a color map of estimated CAM-to-CAM connectivity confidence in Fig. 8(a). Each row and column indicates the index of the camera. When the confidence value is greater than  $\theta_{conf}$ , we regard the corresponding camera pair as a valid link. As a result, the valid camera links are drawn as Fig. 8(b). Each vertex indicates the index of the camera and valid links are represented by edges. Unfortunately, CAM6 failed to be linked to CAM5. That is because the size of the person image patches is very small due to the long distance from the camera. Therefore, it is hard to distinguish the appearances of people. In addition, CAM6 is quite isolated from other cameras as we can see from the matching pairs of CAM6 in the Fig. 6(c). Compared to DNPR [28], which also infers the CAM-to-CAM topology, our method provides more accurate results. Our proposed method provides an accuracy of 87.5%, finding 7 links out of 8. The DNPR [28] method provides an accuracy of 62.5%, finding 5 links out of 8.

### Camera topology initialization results

Figure 9 (a) represents the accuracy of person re-identification in each of the proposed initialization steps such as CAM-to-CAM, Zone-to-Zone, and iterative update steps (Sec. III-B1–III-B3). The accuracy of person re-identification is 28.54% at the beginning, but it is consistently improved by using inferred and refined camera topology information. As a result, our method reaches 62.55% accuracy at the last step of the topology initialization. In addition, it took only 112.46 seconds to conduct both person re-identification and camera topology inference tasks with a large number of people in the nine cameras (using Intel i7 CPU in MATLAB). Figure 10 shows the connectivity confidences between the linked camera pairs. The connections become stronger during the iterations. In the final iteration, most connections demonstrate a high connectivity confidence ( $conf(p(\Delta t)) > 0.7$ ). On the other hand, a connection from Exit (CAM4-Zone4) to Entry (CAM5-Zone2) is not a strong connection, since a small number of people moved from CAM4 to CAM5 during the topology initialization period (AM 11:20 – PM 12:20).

Figure 9 (b) shows the comparison of our proposed approach with a conventional approach called an Exhaustive approach, which fully compares the multiple appearances of people and exhaustively searches the correspondences of people between the entry/exit zones without using camera topology information. For the baseline of the exhaustive method, we used a LOMO feature [51] descriptor. It shows a low performance (52.06% person re-id accuracy) compared

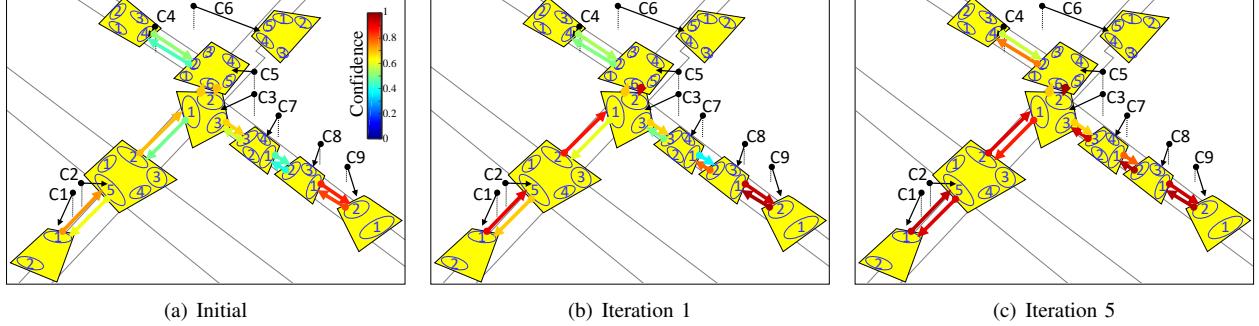


Fig. 10. The iterative update of connectivity confidences. The connectivity of each link has strengthened during the iterations. (best viewed in color).

TABLE IV  
PERFORMANCE COMPARISON WITH OTHER METHODS IN OFFLINE INITIALIZATION STAGE

Performances \ Methods	Makris's [4]	Nui's [23]	Chen's [25]	DNPR [28]	Cai's [27]	LOMO [51] – Exhaustive	Ours
Rank-1 accuracy (%)	49.4	50.0	50.5	40.9	47.0	52.1	<b>62.5</b>
Transition time error (sec)	28.52	23.94	16.29	—	4.84	11.76	<b>1.21</b>

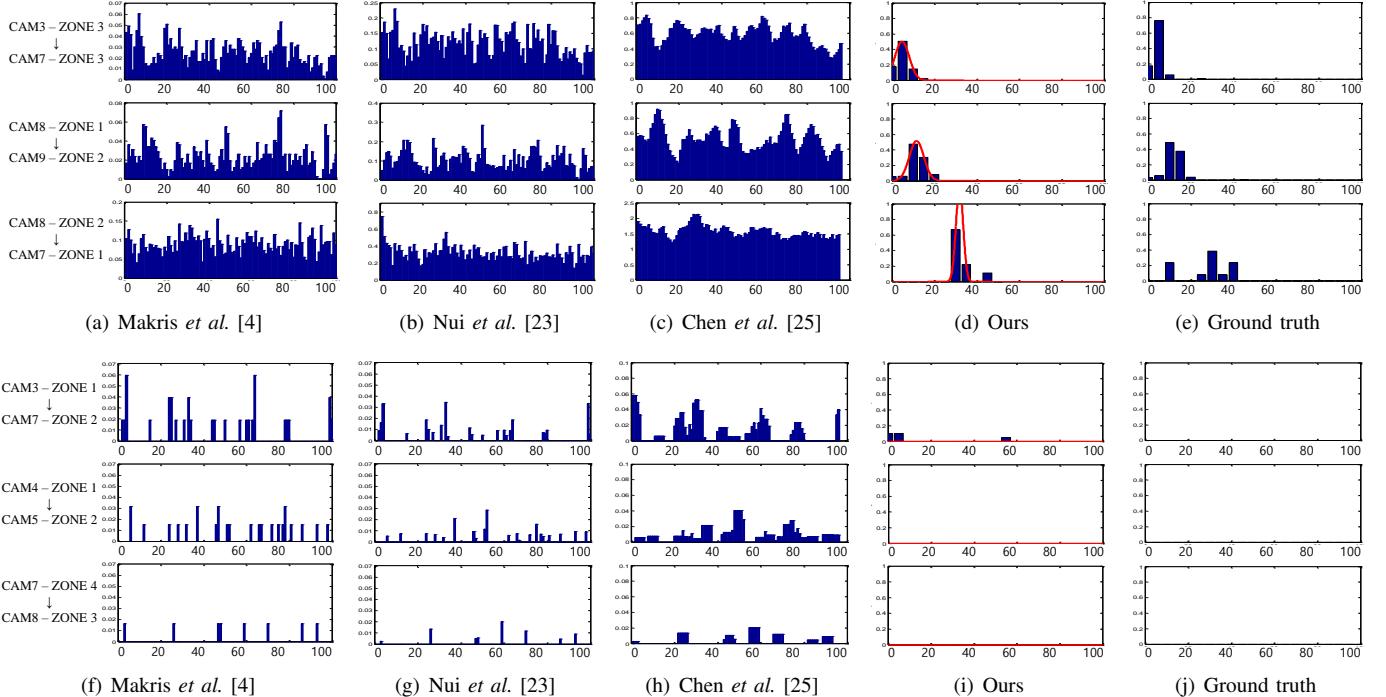


Fig. 11. Comparison of inferred transition distributions and ground truth. First three rows are valid links (a-e) and last three rows are invalid links (f-j).

to the proposed method. Moreover, it takes a much longer time (337.27 seconds) than our proposed method.

A list of valid Zone-to-Zone links inferred by the proposed methods is summarized in Table. III. Note that the pair of Exit (CAM3-ZONE2) – Entry (CAM5-ZONE6) has the negative value of the transition time because these zones are overlapped. The overall results are close to the ground-truth  $N(\mu_{gt}, \sigma_{gt}^2)$ . We expect that the inferred transition distributions of valid links show Gaussian-like distributions and those of invalid links show uniform or arbitrary distributions. However, the previous methods [4], [23], [25] showed unclear and noisy distributions for both valid and invalid links as

shown in Fig. 11 (a-c,f-h). On the other hand, our results are very similar to the ground truth (Fig. 11 (d,e,i,j)).

In Table. IV, we compared several methods in terms of re-identification and topology accuracies. To ensure a fair comparison of the previous methods with our method, we applied the same baseline (e.g. feature descriptor, parameters, experimental settings) to the previous methods except topology inference methods. The inferred topologies of previous methods [4], [23], [25] are used for person re-identification tasks by following the method in [5]. Note that we cannot estimate the *transition time error* of DNPR [28] since the method only infers CAM-to-CAM connections. As we can

see, our method shows superior performances for both re-identification and topology inference than other methods.

2) *Online Person Re-identification Results*: Based on the inferred camera topology in the initialization stage, we conducted online person re-identification and compared it with two different approaches. The first approach estimates the camera topology based on *Exhaustive search* in the offline initialization stage and uses the inferred topology for the online re-identification test. Note that this approach exploits the inferred topology but still fully compares multiple appearances of people to find correspondences in the online test stage. The second approach based on *True matching* estimates the camera topology using ground truth re-identification pairs in the offline initialization stage and performs the online re-identification test in the same way as our method.

As shown in Table V, the performance of our online re-identification test is comparable to that of *True matching*. Our method also outperforms *Exhaustive search* in terms of both re-identification accuracy and topology distance. Thanks to the online topology update in the online re-identification test stage, the topology distance of our method decreased during the online test stage ( $0.076 \rightarrow 0.023$ ). As we can see in the table, our method with the online topology update (*i.e.*, *ours*) performs better than our method without online topology update (*i.e.*, *Ours– no update*) in terms of both re-identification accuracy and topology distance in the online test stage. It demonstrates that our online topology update and re-identification methods are effective and complement each other.

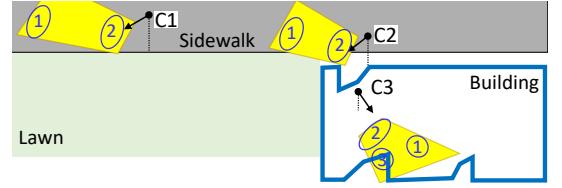
#### B. Evaluations using Public Dataset: NLPR MCT

In order to validate the proposed method using public datasets, we tested the proposed method using the NLPR MCT [43] dataset. The dataset contains four subsets with different non-overlapping multi-camera networks and different numbers of people. Among them, we used two datasets (*i.e.*, DATA1, DATA2) containing larger numbers of people than others (the number of people – DATA1: 235, DATA2: 255, DATA3: 14, DATA4: 49). DATA1 and DATA2 have the same layout and camera field of views as shown in Fig. 12 (a, b-d). Duration of each dataset is 20 minutes. Since the datasets are not large-scale datasets (the number of people is small), we used whole videos (20min) for our proposed camera network topology initialization method and did not perform the online person re-identification.

#### CAM-to-CAM topology inference

Based on our proposed method, we inferred the CAM-to-CAM topologies of both datasets (DATA1, 2). We found strong connections between CAM1 and CAM2 in both datasets. Unfortunately, we could not find the connections from CAM2 to CAM3 in both datasets, although there is a uni-directional link from Exit (CAM2-Zone2) to Entry (CAM3-Zone2) because of the severe illumination difference (indoor and outdoor) and different radiometric parameters between the two cameras.

In this case, person re-identification methods based on matching should handle these photometric issues, which is



(a) A layout of the dataset



(b) CAM1



(c) CAM2



(d) CAM3

Fig. 12. A public multi-camera person re-identification dataset: NLPR MCT [43]

beyond the scope of this study. In this case, even the exhaustive search approaches failed to find the connection between CAM2 and CAM3 because of the same reasons. We expect that by adopting some color correction methods such as [25] before re-identification or exploiting illumination invariant feature descriptors for re-identification, we can find the connection between CAM2 and CAM3. We will try this in future studies.

#### Camera topology initialization results

After the CAM-to-CAM topology inference, we infer the topology of entry and exit zones. We found two valid links between CAM1 and CAM2 in both the datasets. The first link is Exit (CAM1-Zone2) – Entry (CAM2-Zone1) and the second link is Exit (CAM2-Zone1) – Entry (CAM1-Zone2). Through this step, invalid pairs of cameras and zones are detected and ignored. In the next step, we iteratively update the valid links between zones and build a final camera topology map of the camera network.

After evaluating the Zone-to-Zone topology inference, we iteratively update the valid links between zones. In the case of NLPR MCT dataset [43], valid Zone-to-Zone links converged within the second iteration. Table VI shows the person re-identification accuracy according to each proposed topology initialization step. Our method shows a superior re-identification performances: 93.2% and 91.5%, for the camera pair CAM1–CAM2.

We compared our method to several previous topology inference methods [4], [23], [25]. The previous methods showed unclear and noisy transition distributions as shown in Fig. 13 (a-c). Even though the transition distribution inferred by Chen’s method [25] shows a peak of the distribution, the overall distribution is too noisy to utilize. On the other hand, our result is very close to the ground truth and shows a very clear distribution as shown in Fig. 13 (d,e).

## VII. DISCUSSIONS

The proposed unified framework can infer camera network topology while it performs person re-identification. In this study, we assume that the transition distribution between the two links follows a single-mode Gaussian model since the

TABLE V  
PERFORMANCE COMPARISON IN ONLINE TEST STAGE.

	Offline initialization stage		Online test stage	
	Rank-1 accuracy	Topology dist	Rank-1 accuracy	Topology dist
LOMO [51] – Exhaustive	52.1%	5.620	65.6%	0.046
Ours – no update	62.5%	0.076	72.3%	0.093
Ours			<b>74.8%</b>	<b>0.023</b>
True matching	100%	0	75.6%	0.011

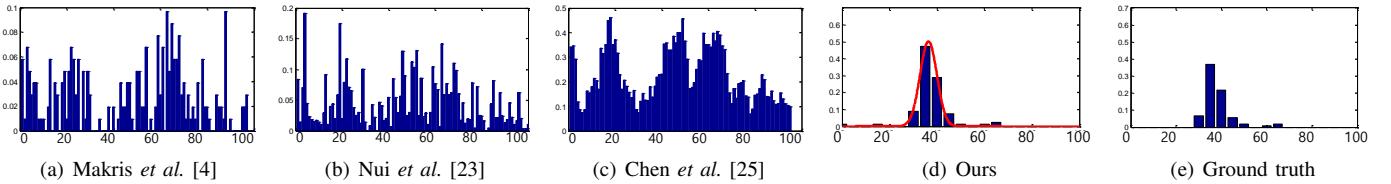


Fig. 13. Comparison of inferred transition distributions and ground truth (DATA2 - Exit: CAM1-ZONE2 , Entry: CAM2-ZONE1).

TABLE VI  
RANK-1 ACCURACIES AT DIFFERENT STAGES OF CAMERA TOPOLOGY INITIALIZATION.

	CAM-to-CAM	Zone-to-Zone	Iteration 1
DATA1 (CAM1 – CAM2)	72.6%	76.7%	93.2%
DATA2 (CAM1 – CAM2)	69.8%	70.8%	91.5%

model works well in most cases. In general, most previous study [4], [23] also follow the same assumption. Although it works well in most cases, it may fail in more complex situations, e.g., multiple paths between two zones. We will consider this issue in further studies.

## VIII. CONCLUSIONS

In this study, we proposed a unified framework to automatically solve both person re-identification and camera network topology inference problems. In addition, to validate the performance of person re-identification in the practical large-scale surveillance scenarios, we provided a new person re-identification dataset called SLP. We qualitatively and quantitatively evaluated and compared the performance of the proposed framework with state-of-the-art methods using public and our own datasets. The results show that the proposed framework is promising for both person re-identification and camera topology inference and superior to other frameworks in terms of both speed and accuracy.

## APPENDIX

As we mentioned in the paper, the traditional exhaustive re-identification approaches examine every person pair between two different cameras ( $c_A, c_B$ ). Assuming that the numbers of objects in  $c_A$  and  $c_B$  are the same as  $N$  and each object contains the same number of appearances as  $K$ , it is obvious that the exhaustive approaches need  $(N^2 K^2)$  time for appearance comparisons for re-identification. Using the Big-O notation, the computational complexity of the exhaustive approach is defined as  $O(N^2 K^2)$ .

On the other hand, in the proposed method, the tree structure of the random forest method makes the multi-shot person re-identification very fast. In return, the random forest containing

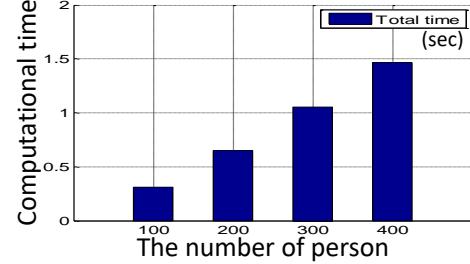


Fig. 14. Computation time of the proposed method based on the number of people in each camera We fixed the number of appearances of each person used for multi-shot matching ( $K = 30$ ). The computation time includes random forest training and testing time.

multiple decision trees needs training time. If we denote the number of decision trees and the random forest training time as  $M$  and  $T$  respectively, we need  $(MNK \log N)$  time for the appearance comparisons and  $T$  for forest training. Here, note that the number of decision trees is a constant value. We fixed it as 10 in our experiments. Therefore, we can ignore the number of the decision trees  $M$  in the Big-O notation since  $M \ll N, K$ . In addition, the random forest training time in the training stage can be also be ignored because the training time  $T$  linearly increases when the number of cameras in a camera network increases whereas the test time exponentially increases when the number of cameras increases. Actually, the computational complexity of the test stage is proportional to the number of possible camera pairs:  $LC_2$ , where  $L$  is the number of cameras in a camera network. For that reason, the training time  $T$  can be ignored when the camera network is large. As a result, the complexity of proposed method is approximated as  $O(NK \log N)$ .

Fig. 14 shows the computation time according to the number of people in each camera in the two-camera re-identification case. As we can see, the computation time of the proposed method increases linearly, not exponentially, as the number of people increases.

## REFERENCES

- [1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [2] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *IEEE International Conference on Machine Learning*, 2007.
- [3] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," *Differences*, 2015.
- [4] D. Makris, T. Ellis, and J. Black, "Bridging the gaps between cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
- [5] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *International Journal of Computer Vision*, 2010.
- [6] Y.-J. Cho, J.-H. Park, S.-A. Kim, K. Lee, and K.-J. Yoon, "Unified framework for automated person re-identification and camera network topology inference in camera networks," *arXiv preprint arXiv:1704.07085*, 2017.
- [7] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, 2014.
- [8] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *European Conference on Computer Vision*, 2012.
- [9] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [10] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [11] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in *Asian Conference on Computer Vision*, 2011.
- [12] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in Neural Information Processing Systems*, 2005.
- [13] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, 2014.
- [14] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, 2014.
- [15] Y. Li, Z. Wu, and R. J. Radke, "Multi-shot re-identification with random-projection-based random forests," in *IEEE Winter Conference on Applications of Computer Vision*, 2015.
- [16] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *IEEE International Conference on Pattern Recognition*, 2014.
- [17] S. Bak, F. Martins, and F. Bremond, "Person re-identification by pose priors," in *IS&T/SPIE Electronic Imaging*, 2015.
- [18] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," 2015.
- [19] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [20] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," *IEEE International Conference on Computer Vision*, 2003.
- [21] Y. Cai and G. Medioni, "Exploring context information for inter-camera multiple target tracking," in *IEEE Winter Conference on Applications of Computer Vision*, 2014.
- [22] A. Rahimi, B. Dunagan, and T. Darrell, "Simultaneous calibration and tracking with a network of non-overlapping sensors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
- [23] C. Niu and E. Grimson, "Recovering non-overlapping network topology using far-field vehicle tracking data," in *IEEE International Conference on Pattern Recognition*, 2006.
- [24] C. Stauffer, "Learning to track objects through unobserved regions," 2005.
- [25] X. Chen, K. Huang, and T. Tan, "Object tracking across non-overlapping views by learning inter-camera transfer models," *Pattern Recognition*, 2014.
- [26] C. C. Loy, T. Xiang, and S. Gong, "Incremental activity modeling in multiple disjoint cameras," 2012.
- [27] Y. Cai, K. Huang, T. Tan, and M. Pietikainen, "Recovering the topology of multiple cameras by finding continuous paths in a trellis," in *IEEE International Conference on Pattern Recognition*, 2010.
- [28] N. Martinel, G. L. Foresti, and C. Micheloni, "Person reidentification in a distributed camera network framework," *IEEE transactions on cybernetics*, 2016.
- [29] L. Breiman, "Random forests," *Machine learning*, 2001.
- [30] S. H. Han, K. J. Yoon, and I. S. Kweon, "A new technique for shot detection and key frames selection in histogram space," 2000.
- [31] D. Makris and T. Ellis, "Automatic learning of an activity-based semantic scene model," in *IEEE Advanced Video and Signal Based Surveillance*, 2003.
- [32] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," *arXiv:1605.09653*, 2016.
- [33] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.
- [34] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *British Machine Vision Conference*, 2009.
- [35] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference*, 2011.
- [36] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, 2011.
- [37] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *IEEE Scandinavian conference on Image analysis*, 2011.
- [38] S. Wang, M. Lewandowski, J. Annesley, and J. Orwell, "Re-identification of pedestrians with variable occlusion and scale," in *IEEE International Conference on Computer Vision Workshops*, 2011.
- [39] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012.
- [40] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey, "A database for person re-identification in multi-camera surveillance networks," in *IEEE Digital Image Computing: Techniques and Applications*, 2012.
- [41] W. Li and X. Wang, "Locally aligned feature transforms across views," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
- [42] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [43] W. Chen, L. Cao, X. Chen, and K. Huang, "An equalised global graphical model-based approach for multi-camera object tracking," *arXiv:1502.03532*, 2015.
- [44] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *European Conference on Computer Vision*, 2014.
- [45] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015.
- [46] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," *arXiv preprint arXiv:1604.02531*, 2016.
- [47] 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 *European Conference on Computer Vision*, 2016.
- [48] M. Gou, S. Karanam, W. Liu, O. Camps, and R. J. Radke, "Dukemtmc4reid: A large-scale multi-camera person re-identification dataset," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, July 2017.
- [49] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision Workshops on Benchmarking Multi-Target Tracking*, 2016.
- [50] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*.
- [51] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

PLACE  
PHOTO  
HERE

**Yeong-Jun Cho** received the B.S. degree in information and communication engineering from Korea Aerospace University and the M.S. degree in information and communications from Gwangju Institute of Science and Technology (GIST), in 2012 and 2014, respectively. He is currently pursuing the ph.D. degree as a member of the Computer Vision Laboratory in GIST. His research interests include computer vision topics such as person re-identification, multi-object tracking, object detection and medical image analysis.

PLACE  
PHOTO  
HERE

**Su-A Kim** received the B.S. degree from the Department of Electronic and IT Media Engineering from Seoul National University of Science and Technology in 2013 and the M.S. degree in School of Information and Communications from Gwangju Institute of Science and Technology (GIST) in 2015. She is currently a scientific researcher at Intel Visual Computing Institute and pursuing the Ph.D. degree in Saarland University. Her research interests include key research topics in computer vision such as person re-identification, 3D object detection and pose estimation, and 4D light field segmentation.

PLACE  
PHOTO  
HERE

**Jae-Han Park** received the B.S. degree in electrical engineering and computer science from Gwangju Institute of Science and Technology (GIST) in 2015. He is currently pursuing M.S degree as a member of Computer Vision Laboratory in GIST. His research interests include person re-identification, video summarization and age estimation using deep learning.

PLACE  
PHOTO  
HERE

**Kyuewang Lee** received the B.S. degree in Electrical Engineering and Computer Science from Gwangju Institute of Science and Technology (GIST) in 2016. He is currently pursuing the integrated M.S./Ph.D. degree as a member of the Perception and Intelligence Laboratory in Seoul National University. His research interests include key research topics in computer vision, such as visual object tracking, multi-object tracking, and person re-identification.

PLACE  
PHOTO  
HERE

**Kuk-Jin Yoon** received the B.S., M.S., and Ph.D. degrees in Electrical Engineering and Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 1998, 2000, 2006, respectively. He was a post-doctoral fellow in the PERCEPTION team in INRIAGrenoble, France, for two years from 2006 and 2008 and joined the School of Information and Communications in Gwangju Institute of Science and Technology (GIST), Korea, as an assistant professor in 2008. He is currently an associate professor and a director of the Computer Vision Laboratory in GIST.