Progressive Unsupervised Person Re-identification by Tracklet Association with Spatio-Temporal Regularization

Qiaokang Xie, Wengang Zhou, Guo-Jun Qi, *Member, IEEE,* Qi Tian, *Fellow, IEEE,* and Houqiang Li, *Senior Member, IEEE*

***Abstract*—Existing methods for person re-identification (Re-ID) are mostly based on supervised learning which requires numerous manually labeled samples across all camera views for training. Such a paradigm suffers the scalability issue since in real-world Re-ID application, it is difficult to exhaustively label abundant identities over multiple disjoint camera views. To this end, we propose a progressive deep learning method for unsupervised person Re-ID in the wild by Tracklet Association with Spatio-Temporal Regularization (TASTR). In our approach, we first collect tracklet data within each camera by automatic person detection and tracking. Then, an initial Re-ID model is trained based on within-camera triplet construction for person representation learning. After that, based on the person visual feature and spatio-temporal constraint, we associate cross-camera tracklets to generate cross-camera triplets and update the Re-ID model. Lastly, with the refined Re-ID model, better visual feature of person can be extracted, which further promote the association of cross-camera tracklets. The last two steps are iterated multiple times to progressively upgrade the Re-ID model. To facilitate the study, we have collected a new 4K UHD video dataset named Campus4K with full frames and full spatio-temporal information. Experimental results show that with the spatio-temporal constraint in the training phase, the proposed approach outperforms the state-of-the-art unsupervised methods by notable margins on DukeMTMC-reID, and achieves competitive performance to fully supervised methods on both DukeMTMC-reID and Campus4K datasets.**

***Index Terms*—Unsupervised person re-identification, Spatio-temporal regularization, Tracklet association.**

I. Introduction

As a hot topic in computer vision, person re-identification (Re-ID) aims at matching pedestrians detected from non-overlapping camera views. Thanks to the potential significance in video surveillance applications, it has attracted wide attention from both academia and industry. Recently, extensive research has been carried out on person Re-ID [1]–[13], and covers a variety of application scenarios such as person search [14], [15], attribute-based Re-ID [16], [17] and Group Re-ID [18], [19], *etc*. In person Re-ID, the key is to learn a good feature representation for pedestrian, which is expected to be invariant to view, pose and the change of cameras, *etc*. In recent years, the performance of person Re-ID has been greatly boosted with the development of deep learning techniques and the release of many large-scale public datasets.

Most existing approaches [1]–[13] for person Re-ID follow the supervised learning paradigm on labeled datasets, where cross-view identity matching image pairs are supposed to be manually labeled for each camera pair. However, we may suffer performance degradation when directly deploying these trained models to a different real-world scenario [20] due to the non-trivial gap between training data and target domain. On the other hand, it is rather strenuous and impractical to annotate the target data, especially for online surveillance videos of large-scale camera network [21]. To directly make full use of the massive and cheap unlabeled video data, person Re-ID by unsupervised learning, where per camera-pair ID labeled training data is no longer required, is gaining increasing popularity [22]–[28].

Spatio-temporal feature [29]–[32] has been widely used in video person Re-ID, which enables the feature extractor to be aware of the current input video sequences. For example, Zhang *et al.* [29] build a spatio-temporal appearance representation for video person re-identification, which exploits the periodicity exhibited by a walking person to generate a spatio-temporal body-action model. Besides, spatio-temporal context in camera network, which means when and where someone appears, also provides a wealth of information to distinguish as well as associate persons of interest [21], [33]–[36]. Since for most people the spatial transfer time between cameras usually follows a similar pattern of walking speed, it helps a lot in short-time person retrieval by eliminating lots of irrelevant images as the camera network is fixed. In the following discussion, the spatio-temporal information refers to spatio-temporal context in camera network.

To learn a person Re-ID model, substantial cross-view training data is needed in order to cope with the significant visual appearance change between different cameras. To this end, cross-camera person tracklet association is an alternative to provide cross-view data for unsupervised Re-ID learning. As spatio-temporal information is beneficial to essentially im-

Manuscript received October 18, 2019; revised February 15, 2020; accepted March 18, 2020. The work of W. Zhou was supported in part by the National Key R&D Program of China under contract 2018YFB1402600, in part by the National Natural Science Foundation of China under contract 61822208 and 61632019, and in part by Youth Innovation Promotion Association CAS 2018497. The work of H. Li was supported by NSFC under contract 61836011.

Qiaokang Xie, Wengang Zhou, and Houqiang Li are with the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230027, China. E-mail: xieqiaok@mail.ustc.edu.cn, {zhwg, lihq}@ustc.edu.cn.

Guo-Jun Qi is with Futurewei Technologies. E-mail: guojunq@gmail.com.

Qi Tian is with Huawei Noah's Ark Laboratory. E-mail: tian.qi1@huawei.com.

Corresponding authors: Wengang Zhou and Houqiang Li.

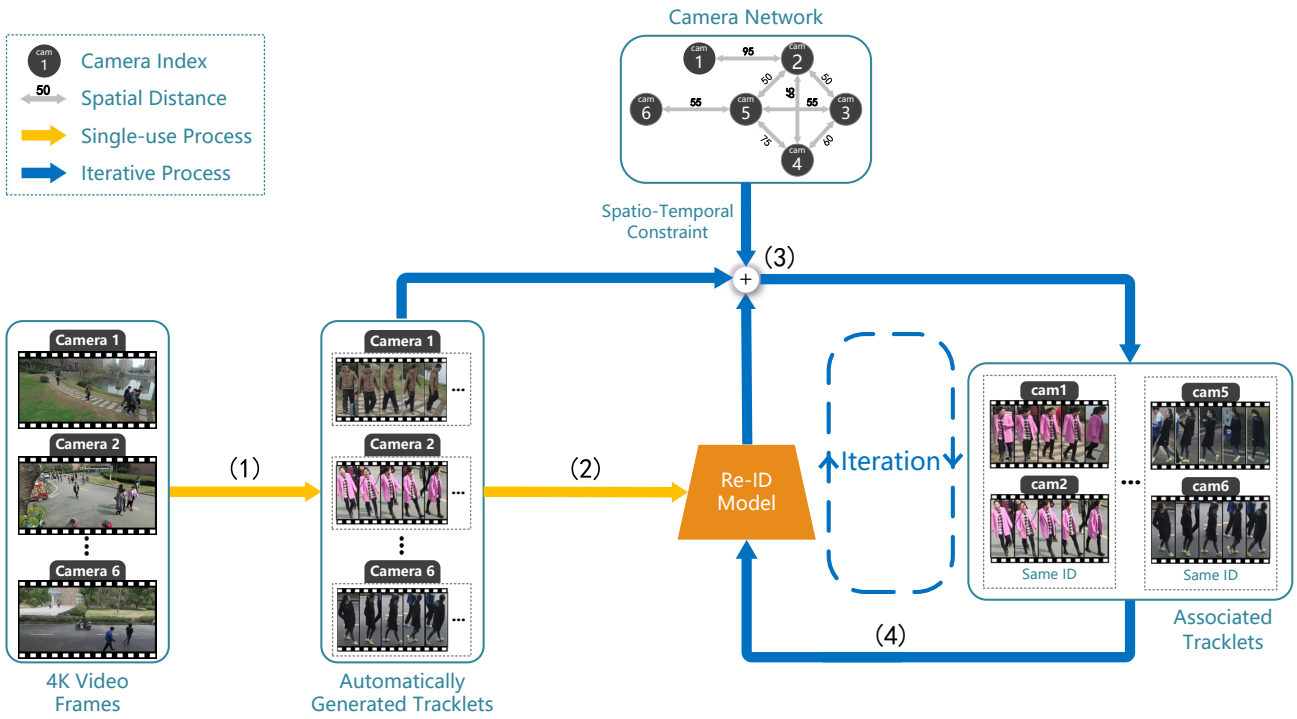


Fig. 1. Our framework consists of 4 steps: (1) Multi-person detection and tracking. (2) Within-camera Re-ID learning. (3) Cross-camera tracklets association with spatio-temporal regularization. (4) Cross-camera Re-ID learning. Step (1) and (2) only conduct once while (3) and (4) are iterated several times for progressive optimization.

prove the performance of short-time person Re-ID, it can also be explored in unsupervised cross-camera tracklet association. Considering the fact that it is more important to improve the precision of tracklet association for unsupervised methods than to sort out all the tracklets belonging to a person, it suffices to design a relatively simple tracklet association based on spatio-temporal constraint without any labeled data, even if we miss some positive tracklet pairs (with the same ID).

Based on the above motivation, in this paper, we propose a progressive unsupervised learning framework by cross-camera Tracklet Association with Spatio-Temporal Regularization (TASTR). Some previous unsupervised methods [21], [26], [37] either suffer the lack of accurate spatio-temporal information, or directly use off-the-shelf tracklets in training set for model initialization and assume different tracklets contain different person identities. To avoid the above issues and well justify the proposed approach, we have collected a new 4K UHD video dataset for unsupervised person Re-ID, which contains full frames and full spatio-temporal information. The general framework of the proposed approach is illustrated in Fig. 1. In the first stage, we conduct multi-person detection and tracking per camera to get their tracklets and train a within-camera Re-ID model. In the second stage, cross-camera tracklets association with spatio-temporal regularization is proposed to obtain accurate pseudo labels across cameras to further learn cross-view identity-specific discriminative information. Finally, the Re-ID model is progressively optimized by iterating the two steps in the second stage.

Our contributions can be summarized into three aspects:

- We collect a new 4K UHD video dataset, named Campus4K, for unsupervised person Re-ID. Compared with existing Re-ID datasets, Campus4K is of higher quality with full frames and complete spatio-temporal information.
- We propose a progressive unsupervised deep learning framework for person Re-ID by Tracklet Association with Spatio-Temporal Regularization (TASTR), which significantly improves the precision and recall rate of cross-camera tracklet association as well as the performance of the Re-ID model.
- Extensive experiments show that our method notably improves the performance of unsupervised Re-ID. Our unsupervised method with spatio-temporal clues only in the training phase achieves competitive performance (rank-1: 76.4% on Campus4K and 74.1% on DukeMTMC) with fully supervised methods (rank-1: 85.2% on Campus4K and 78.1% on DukeMTMC) using the same batch hard triplet loss [5].

II. Related Work

In this section, we mainly review the unsupervised person Re-ID methods that are most related to the proposed approach, where pairwise ID labeled training data for each pair of camera views is not required in model learning.

Early unsupervised person Re-ID methods mainly focus on feature representation learning [38]–[40]. Dictionary learning [22], [41] and salience learning [25], [42] methods are

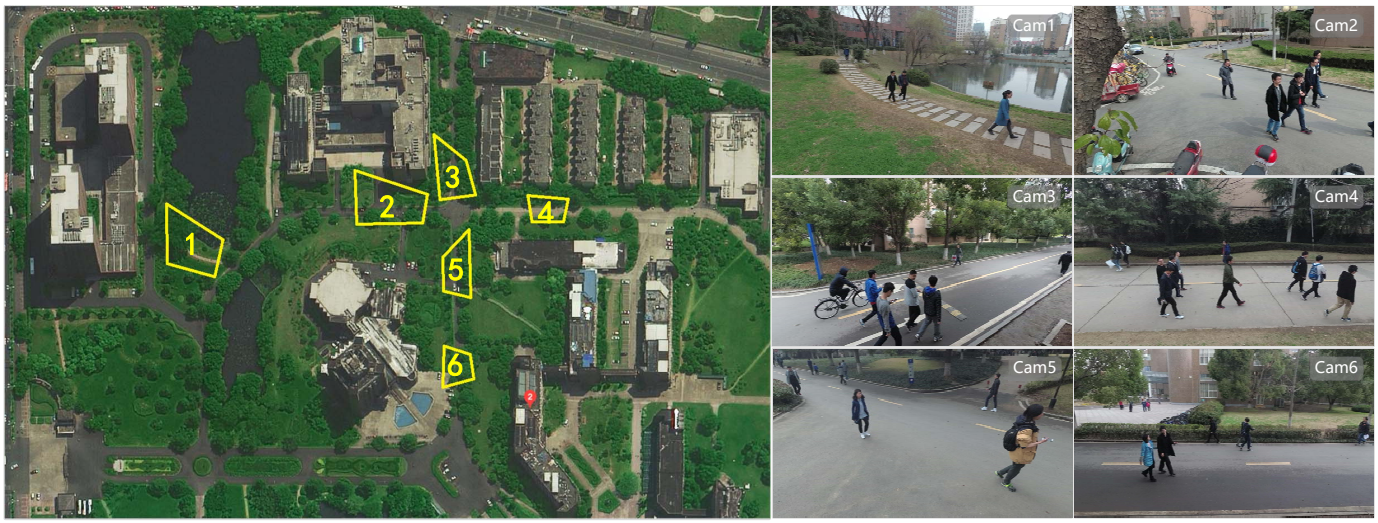


Fig. 2. Camera network of Campus4K: 6 static non-overlapping synchronized 3840×2160 UHD cameras

also proposed to learning salient and view-invariant representations. To balance the scalability and accuracy of the Re-ID model, semi-supervised learning [23], [41] are proposed but they still assume sufficient cross-view labeled data for model training. Recently, some cross-dataset transfer learning methods [20], [43]–[48] have been proposed to leverage the labeled data in other datasets to improve the performance on target dataset. These methods gain much better accuracy than the classical unsupervised methods, but they need the auxiliary source Re-ID dataset and still require latent similarity between the source domain and the unlabeled target domain.

Similar to our work, cross-camera tracklet association (labeling) [26], [37], [49] is more scalable for unsupervised Re-ID with no extra data or assumption on the similarity between source and target domains. Due to the limitation of existing datasets, most of them do not perform Re-ID learning in a pure unsupervised way. Instead, they take for granted the tracklets provided by the dataset, which essentially assumes the perfect tracking in videos. However, such an assumption is somewhat too strong in real-world situations. Li *et al.* [49] propose a Tracklet Association Unsupervised Deep Learning (TAUDL) model to consider a pure unsupervised person Re-ID problem. They perform Sparse Space-Time Tracklet (SSTT) sampling for within-view tracklet labeling and formulate a Cross-Camera Tracklet Association (CCTA) loss for coarse-grained underlying cross-view tracklet association. However, SSTT throws away a lot of tracklets which may be useful for Re-ID learning while CCTA loss does not make explicit cross-view tracklet association. In contrast, our method can make full use of tracklet data and achieve accurate cross-camera tracklet association result, which leads to a considerable improvement in Re-ID performance.

Different from most existing unsupervised person Re-ID approaches, we perform a pure unsupervised person Re-ID at the very beginning (multi-person detection and tracking) without assuming that the tracklets per camera are off-the-shelf and different tracklets indicate different person identities. Actually, the tracklets can be interrupted and the number of tracklets is more than the number of identities per camera. Experiments show that person tracking not only provides within-camera training data for Re-ID but also benefits from within-camera person discriminative learning. Then with the help of spatio-temporal constraint, we considerably promote the precision and recall rate of cross-camera tracklet association. Finally, we iterate cross-camera tracklet association and Re-ID learning to optimize the Re-ID model progressively and achieve competitive performance with supervised methods using the same triplet loss.

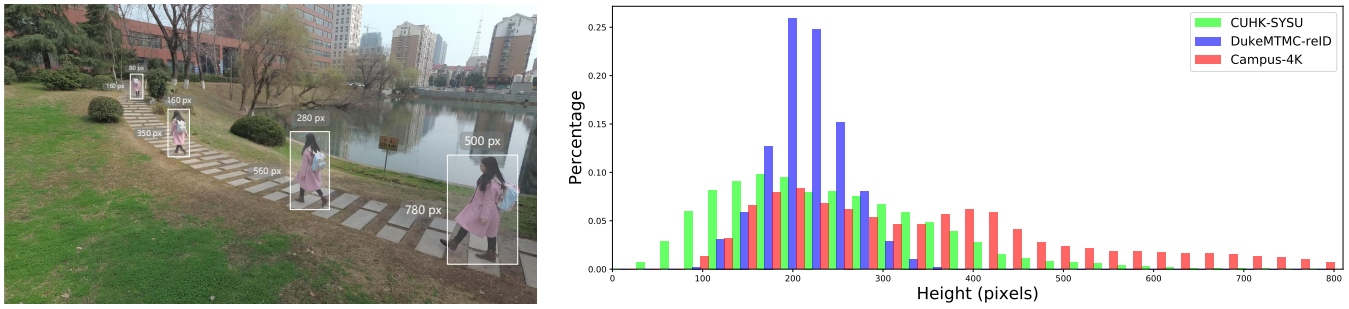
III. Campus4K Dataset

A. Dataset Description

Thanks to many public Re-ID datasets [1], [9], [50]–[55], great progress has been made for person Re-ID in recent years. In most existing Re-ID datasets, the resolution is low and the person images are usually blurred with indistinguishable characteristic detail, which imposes significant difficulty on the algorithm design. To this end, we have collected a new 4K UHD video dataset, *i.e.* Campus4K, with full frames and full spatio-temporal information. It is collected by a non-overlapping outdoor camera network on campus (see Fig. 2), which consists of six synchronized 3840×2160 UHD cameras. It aims to provide a high-quality Re-ID dataset for both supervised and unsupervised Re-ID, and in this paper we mainly focus on unsupervised learning.

Our Campus4K dataset carries the spatio-temporal context information about the camera position and recorded videos. In some recent work [21], [36], it has been shown that spatio-temporal context is of great importance to improve the Re-ID accuracy. Such information is also easily accessible in practice.

To our best knowledge, Campus4K is the first public 4K (3840×2160) UHD dataset for person Re-ID. In addition, full frames and full spatio-temporal information are provided. In recent years, there has been a trend of increasing resolutions for surveillance cameras. As video quality improves, some



(a) (b)

Fig. 3. (a) Illustration of person scale variance in our Campus4K dataset. (b) The scale (height) distributions of person bounding boxes in three datasets. It is notable that the person's scale in Campus4K covers a much diverse range than the others.

fine-grained details of people such as textures of clothes, carry-on items and even face become available. They make it easier to recognize many attributes that were previously difficult to classify, and it also brings new challenges such as large scale (resolution) variance of person bounding boxes, and the joint combination of face recognition and person Re-ID in the wild.

As shown in Fig. 4(a), since 4K cameras have a broader view than ordinary cameras and the distance between person and camera is uncontrolled, the scale of people may vary significantly, which introduces a multi-scale matching problem [56]. As shown in Fig. 4(b), in Campus4K, most person bounding boxes are in the range of about 100 to 800 pixels in height, and it covers a more diverse range than other existing datasets. Moreover, the face regions in Campus4K are of relatively high resolution, which makes it possible to perform face recognition tasks. However, since the faces are not always visible, and are taken in different views, lighting conditions and scales, it is still very challenging to link face recognition to person Re-ID, which is out of scope for this work.

Camera	# Identities	# Tracklets	# BBoxes
Cam 1	331	335	59,021
Cam 2	649	660	147,679
Cam 3	645	653	83,569
Cam 4	841	843	62,307
Cam 5	689	713	123,230
Cam 6	643	645	45,503
Total	1,567	3,849	521,309

TABLE I
STATISTICS OF CAMPUS4K.

Dataset	# IDs	# BBoxes	Camera Resolution	Full*
PRID2011 [51]	178	38,466	-	✔
iLIDS-VID [52]	300	43,800	-	
MARS [54]	1,261	1,191,003	1920×1080, 640×480	
Campus4K	1,567	521,309	3840×2160	✔
CUHK03 [1]	1,467	14,097	-	
Market-1501 [53]	1,501	32,668	1280×1080, 720×576	
DukeMTMC-reID [55]	1,812	36,411	1920×1080	✔
MSMT17 [9]	4,101	126,411	-	

TABLE II
COMPARISON BETWEEN CAMPUS4K AND OTHER PERSON RE-ID DATASETS. "-": NO REPORTED CAMERA INFORMATION IS AVAILABLE. "FULL*": FULL FRAMES AVAILABILITY.

B. Evaluation Protocol

The time of all cameras is synchronized, and $77,195$ 4K frames are collected continuously at 30 fps per camera. As ground truth, person bounding boxes with identity information are generated by manual annotation with the help of tracking and temporal smoothing. We split our dataset into training set and testing set according to the moment that the person was first captured by our camera network. For all identities who appear in at least two camera views, we denote by n the index of the earliest frame of his/her appearance in all camera views. Identities with n less than 46,410 would be assigned to the training set, and the rest would be assigned to the testing set. Besides, identities who appear in only one camera would be considered as distractors.

During testing, one tracklet of each person in testing set would be selected to form query set, and the gallery set consists of the rest tracklets in testing set along with all distractors to make the evaluation more challenging. Images of each video are sampled every three frames (0.1 seconds) and there are 135 images per person per video on average. More details and the comparison between Campus4K and other Re-ID datasets can be found in Table I and Table II. It should be noted that if full frames are needed for unsupervised Re-ID, only frames whose indices n are less than $46,410$ can be used for training to avoid the use of identities from the testing set.

Like most existing Person Re-ID datasets, Cumulated Matching Characteristics (CMC) curve is used to evaluate the performance of Re-ID methods. For each query video, multiple ground truths may exist and the CMC curve may be biased since "recall" is not considered [53]. Therefore, mean Average Precision (mAP) is also used for overall performance evaluation.

IV. THE PROPOSED METHOD

A practical person Re-ID system in surveillance usually consists of three modules, *i.e.*, person detection, tracking and re-identification [57]. Although they are generally considered as three independent computer vision tasks, they can complement each other to improve performance. In this paper, we propose a novel progressive unsupervised Re-ID approach

(a) YOLOv3 (b) AlphaPose

Fig. 4. Examples of multi-person detection result *w.r.t.* different detection methods: (a) YOLOv3, (b) AlphaPose. YOLOv3 suffers from various problems such as detection error, location error and miss detection while AlphaPose can provide much more stable and fine-grained results. Better detections are also beneficial for subsequent multi-person tracking. And AlphaPose with PoseFlow can automatically generate high quality data for unsupervised perosn Re-ID learning initialization.

by Tracklet Association with Spatio-Temporal Regularization (TASTR). The proposed framework is shown in Fig. 1, which consists of four key steps as follows. First, for each camera, we obtain tracklets by multi-person detection and tracking for initialization. Then, we conduct within-camera training for person Re-ID based on the initially obtained tracklets. After that, we associate different person tracklets across cameras with Spatio-Temporal Regularization (STR). Finally, based on the cross-camera tracklet association results, we re-train the Re-ID model to generate better feature representation.

The first two steps focus on within camera exploration, and generate within-camera Re-ID model. The remaining two steps are iterated multiple times as the Re-ID model may generate more and better matching pairs, which can be used for more effective Re-ID learning. In other words, we optimize the Re-ID model in a progressive fashion.

A. Multi-person Detection and Tracking

We exploit person detection and tracking to generate initial data for unsupervised person Re-ID learning. To begin with, we detect all persons in each frame by AlphaPose [58]. AlphaPose generates high-quality pose estimation from person bounding boxes obtained by detector, *e.g.*, YOLO [59], and is resilient against imperfect detection. Since the involved detection method AlphaPose is generic and there is no fine-tuning on target dataset, it is reasonable to be used as an off-the-shelf module. Then, we track the pose each person with PoseFlow [60], which involves no learning. Specifically, we use ORB feature [61] for cross-frame matching in PoseFlow. And based on human poses and ORB feature, PoseFlow uses an online optimization framework to build the association of cross-frame poses and form pose flows. Then, pose flow non-maximum suppression (PF-NMS) is used to reduce redundant pose flows and re-link temporal disjointed ones. The bounding box of each person is derived from the pose (see Fig. 4(b)), which is used to crop the image patch for the following tracklet association and Re-ID learning.

It is an alternative to use YOLO (or Faster-RCNN [62]) for detection and IOU for tracking. However, as shown in Fig. 4(a) the general object detection methods such as YOLO and Faster-RCNN are unable to extract fine-grained human bodies and suffer from instability in detecting people of various scales. Thus, an additional learning-based alignment is often required to refine the person detection results to benefit the following person Re-ID task. In contrast, based on human pose estimation with AlphaPose and pose tracking with PoseFlow, we can obtain more stable and fine-grained results, which is free of person alignment for person Re-ID.

B. Within-camera Re-ID Learning

With AlphaPose and PoseFlow, we obtain a series of tracklets each of which contains the temporal patches of a certain person. However, due to imperfect tracking, it is inevitable that the trajectory of a person will be fragmented into different tracklets, especially when people are occluded. Due to the limitation of datasets, many traditional unsupervised methods directly use complete tracklets given by training data and assume different tracklets indicate different identities. Such a setting ignored the tracklet identity duplication problem, *i.e.*, different tracklets may correspond to the same person ID.

Observing that in surveillance videos such as people re-appearing in a camera view is rare during a short time period and most people travel through a single camera view in a common time period $Q < T$, Li *et al.* [49] propose a Sparse Space-Time Tracklet (SSTT) sampling method and treat each camera view separately as a classification task. However, it risks declining many tracklets that may be helpful for Re-ID learning. Instead of considering it as a classification problem, we adopt triplet loss for Re-ID training. To avoid losing too many tracklets that are potentially useful for Re-ID learning, we use all tracklets to perform within-camera Re-ID training.

Concretely, we adopt triplet loss with hard mining proposed by Hermans *et al.* [5]. For each triplet batch, we randomly sample P tracklets, and then randomly sample K images for each tracklet, resulting in a batch of $P \times K$ images.

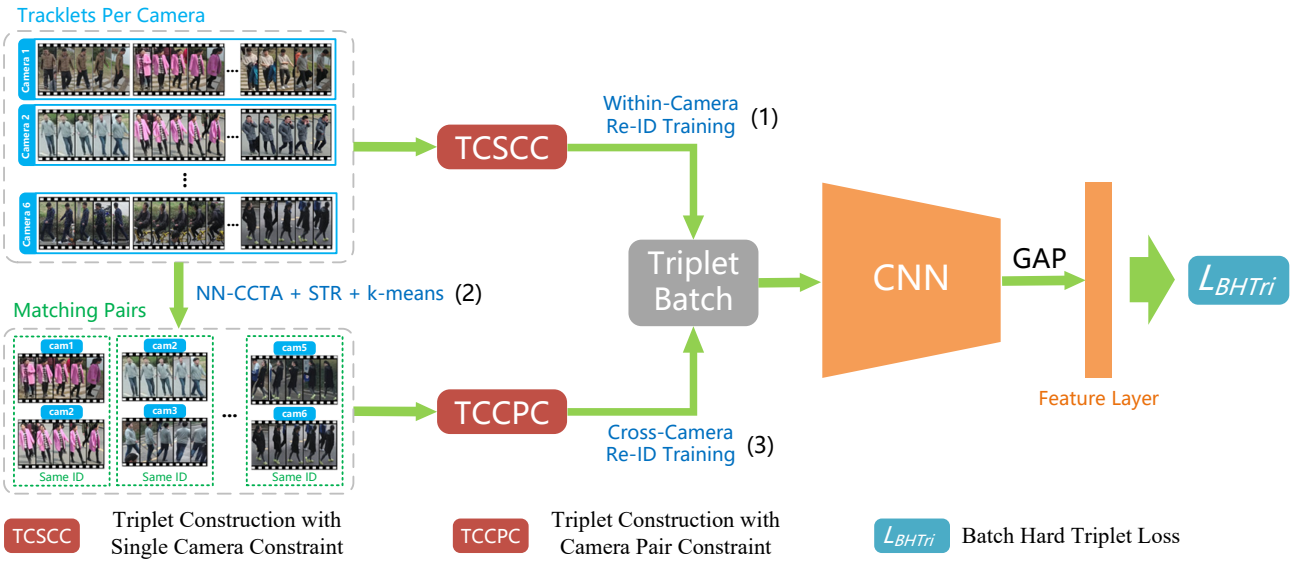


Fig. 5. The pipeline of Re-ID training. (1) Given tracklets per camera, TCSCC sampling strategy is used to form the triplet batch for within-camera Re-ID learning. (2) Based on the learned model we can perform 1-reciprocal Nearest Neighbor based Cross-Camera Tracklet Association (NN-CCTA) with Saptio-Temporal Regularization (STR) and k-means to obtain accurate matching tracklets for each camera pair. (3) TCCPC is used to form triplet batch for cross-camera Re-ID learning. In this figure, process (1) is performed only once while processes (2) and (3) are iterated for progressive optimization.

To deal with false negatives caused by ID duplication as much as possible, we proposed a sampling strategy called *Triplet Construction with Single Camera Constraint* (TCSCC). It requires that all P sampled tracklets within the batch are from the same camera and the time interval between them must be greater than a gap of T. Different from SSTT [49], we do not decline tracklets that do not satisfy the above conditions because they may appear in other triplet batches as well. Finally, for each anchor $\mathbf{a}$ in the batch, only the hardest positive and negative samples will be considered for computing the loss, which is a build-in hard sample mining method called *Batch Hard*. The batch hard triplet loss is computed as:

$$\mathcal{L}_{BHTri} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{\text{all anchors}} \Big[m + \overbrace{\max_{p=1\ldots K} D\left(f(x_a^i), f(x_p^i)\right)}^{\text{hardest positive}} - \overbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D\left(f(x_a^i), f(x_n^j)\right)}^{\text{hardest negative}} \Big]_+, \tag{1}$$

where m is the margin of triplet loss, x_j^i corresponds to the j-th image of the i-th person in the batch, $f(x_j^i)$ denotes its feature, $D(\cdot,\cdot)$ means distance, and $[\cdot]_+ = max(0,\cdot)$.

With the above triplet loss based learning, we can obtain a preliminary Re-ID model, denoted as TASTR-S1, with strong within-camera ID discrimination capability (see Sec. V-C1 and Fig. 6), which, in turn, is helpful for within-camera tracklets association (re-link). In other words, the Re-ID model learns within-view ID discriminative information from tracking and can improve tracking performance in return.

C. Cross-Camera Tracklets Association

There is still a big gap between TASTR-S1 and the supervised model due to the lack of cross-view pairwise data. To address this problem, we need to associate cross-camera tracklets for global ID discriminative learning. It may not be satisfactory to directly use TASTR-S1 for cross-view tracklet association as it may be weak in associating the same person in cross-view circumstances. However, some extra clues like spatio-temporal information which is easily available in real-world practices have great potential for improving the precision and recall of tracklet association.

It is often difficult to model the spatio-temporal pattern of moving people because of the diverse paths and uncontrollable pedestrians among different cameras. However, in unsupervised Re-ID training, it is more important to obtain correctly matched cross-view tracklets that belong to the same ID for cross-view ID discriminative learning than sorting out all the tracklets of a person. We make use of the assumption that most people move with definite purposes at similar speeds, as well as the camera network is fixed. So spatio-temporal constraint can eliminate plenty of irrelevant images and narrow the search space for cross-camera tracklet association. This can considerably improve the precision and quantity of associated tracklets with limited hard samples missing.

Formally, given two tracklets T_i and T_j (i and j denote tracklet indexes), we extract visual features by TASTR-S1 and get two feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively. Then we compute the Euclidean distance between them:

$$D(T_i, T_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \tag{2}$$

On the other hand, we use a simple but effective Gaussian function to reflect the spatio-temporal constraint for a camera

pair:

$$R(\Delta t, c_i, c_j) = exp\Big(- \frac{(\Delta t - \bar{t}_{c_i,c_j})^2}{2\sigma_{c_i,c_j}} \Big), \tag{3}$$

where Δt denotes the time interval between T_i and T_j, c_i means the camera index of the i-th tracklet, $\bar{t}_{c_i,c_j}$ is the average moving time between the camera pair (c_i, c_j), and σ_{c_i,c_j} is its standard deviation. Without labeled training data, $\bar{t}_{c_i,c_j}$ is estimated by the spatial path length between the camera pair and pedestrians' average speed, and we set $\sigma_{c_i,c_j} = \lambda \bar{t}_{c_i,c_j}$. Therefore, farther distance result in larger $\bar{t}_{c_i,c_j}$ and σ_{c_i,c_j}

Finally, we compute the joint distance by visual feature and Spatio-Temporal Regularization (STR) as follows,

$$D_{joint}(T_i, T_j) = \frac{D(T_i, T_j)}{R(\Delta t, c_i, c_j)}. \tag{4}$$

From Eq. (4), we can see that if the time interval of a tracklet pair is far from the average transfer time of the corresponding camera pair, it will lead to a small regularization item R, so the D_{joint} would increase as spatio-temporal prior indicates it is unlikely to have the same identity. In this way, some abnormal true matching pairs may be suppressed. However, much more true matches with common transfer time can be obtained since it eliminates lots of irrelevant matching tracklets and thus narrows the search space of cross-camera tracklet association. It can effectively improve the accuracy and recall rate of matching pairs (*i.e.* more quantity and higher quality), which is quite important for improving the performance of unsupervised Re-ID training.

Then we perform cross-camera tracklet association based on D_{joint}. Specifically, for each camera pair, given a tracklet in one camera as probe and all tracklets in another camera as gallery, the ranking list can be computed by their joint distance. If we directly use the top match as the association result, it can lead to many false matches. As k-reciprocal nearest neighbors are more likely to be relevant to the probe [63], we adopt 1-reciprocal Nearest Neighbor based Cross-Camera Tracklet Association (NN-CCTA) as a strong constraint to identify possible matched candidates, *i.e.*, both of them are the top one match of each other. Formally, for each camera pair, let $\boldsymbol{s}_i$ be the i-th tracklet in one camera, and $\mathcal{N}^1(\boldsymbol{s}_i)$ be the nearest neighbour (NN) of $\boldsymbol{s}_i$ in another camera. The 1-reciprocal NN $\mathcal{R}(\boldsymbol{s}_i)$ for $\boldsymbol{s}_i$ is defined as:

$$\mathcal{R}(\boldsymbol{s}_i) = \{\boldsymbol{s} | \boldsymbol{s} \in N^1(\boldsymbol{s}_i) \ and \ \boldsymbol{s}_i \in N^1(\boldsymbol{s})\}. \tag{5}$$

After all the cross-camera tracklet pair candidates are obtained by NN-CCTA, we perform k-means for further refining to improve the quality of matching pairs, which is important for progressive improvements. Specifically, it is 1-D k-means on the Euclidean distances of all tracklet pair candidates for each camera pair, and the k initial points are evenly located between the minimum and maximum distance values. Finally, the tracklet pairs that belong to the cluster with minimum average distance are selected as matching pairs for cross-view Re-ID training.

D. Cross-camera Re-ID Training

So far we have got many cross-view tracklet pairs with high confidence via strict matching. Similar to Sec. IV-B, we use batch hard triplet loss for cross-camera Re-ID training. All matching tracklets of all camera pairs are adopted for cross-view discriminative learning, but in each triplet batch only those tracklets from the same camera pair are sampled to deal with ID duplication problem between different camera pairs. This sampling strategy is denoted as *Triplet Construction with Camera Pair Constraint* (TCCPC).

The pipeline of Re-ID learning is shown in Fig. 5. With the help of cross-camera tracklet association, the Re-ID model can achieve stronger cross-view ID discrimination capability. In other words, based on the refined model, we extract more discriminative visual feature. With the better visual feature and the spatio-temporal clues, we can derive higher quality cross-view matching pairs by cross-camera tracklet association. Based on such motivation, we repeat cross-camera tracklet association and Re-ID training for several times. As a result, the Re-ID model can get progressive improvements in a mutual promotion manner.

V. Experiments

A. Setup

1) Datasets: Most existing Re-ID datasets lack both synchronized time-stamp and spatial distribution information of the camera network. Market1501 [53] and GRID [64] provide the frame numbers in video sequences, which can be used as time-stamps, but it is unknown whether the time is synchronized and the location and coverage *w.r.t.* each camera is not given as well. DukeMTMC-reID [55] also provides the frame numbers, and it is a subset of the multi-target multi-camera tracking dataset DukeMTMC [65]. The cameras of DukeMTMC are synchronized and the spatial distribution of the cameras is also provided. So besides our Campus4K dataset, we choose DukeMTMC-reID to evaluate the proposed unsupervised TASTR model. No spatio-temporal information is used in testing phase on both datasets.

For DukeMTMC-reID, as the cropped person images are off-the-shelf, we consider the image patches of a person in one camera as a tracklet and ignore its label for unsupervised learning as the previous tracklet based unsupervised methods do [49], [66]. Notably, this is not a pure unsupervised setting as there is no within-camera ID duplication problem so different tracklets per camera contain different person identities. Campus4K is a new Re-ID dataset with full frames and full spatio-temporal information. It is readily ready for the evaluation of the unsupervised Re-ID learning from multi-person detection and tracking. We conduct unsupervised person Re-ID using automatically generated tracklets (obtained by detection and tracking), which inevitably involves errors (*e.g.* unknown trajectory fragmentation due to no manual verification), to test the realistic model performances in the wild.

2) Implementation Details: We use the ResNet-50 [67] model pre-trained on ImageNet as the backbone, and train the model with batch hard triplet loss. Images are resized to 256×128, and we augment the training images online with

Methods	Reference	rank-1	rank-5	mAP
LOMO [69]	CVPR'15	12.3	21.3	4.8
BOW [53]	ICCV'15	17.1	28.8	8.3
UDML [43]	CVPR'16	18.5	31.4	7.3
PUL† [20]	TOMM'18	30.4	44.5	16.4
CycleGAN† [44]	ICCV'17	38.5	54.6	19.9
SPGAN† [47]	CVPR'18	41.1	56.6	22.3
TJ-AIDL† [46]	CVPR'18	44.3	59.6	23.0
SPGAN+LMP† [47]	CVPR'18	46.9	62.6	26.4
HHL† [70]	ECCV'18	46.9	61.0	27.2
TAUDL [49]	ECCV'18	61.7	-	43.5
UTAL [66]	TPAMI'19	62.3	-	44.6
MAR† [48]	CVPR'19	67.1	79.8	48.0
TASTR	This work	**74.1**	**85.5**	**54.9**

TABLE III
COMPARISON OF THE PROPOSED TASTR METHOD WITH STATE-OF-THE-ART UNSUPERVISED METHODS ON DUKEMTMC-REID. "-" : NO REPORTED RESULT OR IMPLEMENTATION CODE IS AVAILABLE. †: TRANSFER BASED METHOD USING EXTRA RE-ID DATASET.

Methods	Source Dataset	rank-1	rank-5	mAP
LOMO [69]	-	8.5	21.1	11.3
PUL† [20]	CUHK03	15.1	30.5	17.9
PUL† [20]	Market1501	26.1	44.5	29.3
PUL† [20]	DukeMTMC-reID	38.8	59.3	40.3
TAUDL [49]	-	56.2	78.4	58.2
MAR† [48]	MSMT17	66.1	83.0	67.3
TASTR	-	**76.4**	**91.6**	**78.3**

TABLE IV
BENCHMARKING RESULTS OF SOME EXISTING METHODS ON CAMPUS4K. "-" : NO SOURCE RE-ID DATASET REQUIRED. †: TRANSFER BASED METHOD USING EXTRA RE-ID DATASET.

cropping, horizontal flip and normalization. To balance the number of images of different tracklets as some tracklets may contain hundreds of images, at most 60 images of each tracklet would be randomly selected for training. We use Adam [68] as the optimizer and the initial learning rate is 0.0003. We set $k = 3$ for k-means, $\lambda = 0.7$ for STR, and the number of iterations is fixed to 5. We use a server with 2 1080Ti GPU cards and single CPU (Intel Xeon CPU E5-2640) for model training. The number of training epochs of TASTR is 100. The first half epochs are used for within-camera Re-ID training, while the remaining epochs are used for cross-camera Re-ID training. We deploy cross-camera tracklet association (CCTA) at the beginning of each iteration, which requires little extra time (less than 3 minutes). The total training time is about 3 hours on DukeMTMC-reID and 4 hours on Campus4K. Our code and models are available at https://github.com/xieqk/TASTR.

B. Comparison to the State-of-the-Art Methods

We compare our TASTR model with some existing unsupervised state-of-the-art methods on DukeMTMC-reID as shown in Table III. Among the comparison methods, BOW [53], LOMO [69] and UDML [43] are based on hand-crafted feature representation. PUL [20], CycleGAN [44], SPGAN [47], TJ-AIDL [46], HHL [70] and MAR [48] are methods which transfer the knowledge of labeled data in source Re-ID dataset, *e.g.* Market1501, through model adaptation to the unlabeled target dataset. Besides, TAUDL [49], UTAL [66] and the proposed method are unsupervised tracklet based methods, and all of them assume person images per ID per camera are drawn from a single person tracklet on target dataset DukeMTMC-reID, without involving any additional labeled source Re-ID dataset.

From Table III, it is observed that BOW [53], LOMO [69] and UDML [43] achieve very limited performance, while transfer based model obtain higher performance than hand-crafted features due to the use of additional label information, for instance, HHL [70] achieves 46.9% rank-1 accuracy. Comparing with unsupervised transfer-based methods, tracklet association based unsupervised methods such as TAUDL [49], UTAL [66] and the proposed TASTR are much superior, and TASTR outperforms all the competing methods. Notably, no extra labeled data is used for training in TASTR, therefore it is more scalable than those approaches that require extra Re-ID dataset or attribute labels.

As shown in Table IV, we also provided some benchmarking results of existing methods on Campus4K using officially released code, and all of them use the same automatically generated tracklets as target training data. The performance of LOMO [69] is very limited. PUL [20] uses other Re-ID datasets as the source training data and need to manually set the number of clusters K based on the number of IDs in target training set. We use different source Re-ID datasets and set K to be equal to the number of IDs in training set (*i.e.* 695) to get the best results. The results show that PUL achieves better performance but is greatly affected by using different source Re-ID datasets, and it benefits more from DukeMTMC-reID. This may be because the data distributions and viewing conditions of DukeMTMC-reID are more similar to the unlabeled target dataset Campus4K. MAR [48] adopt the MSMT17 as an auxiliary Re-ID dataset, which has more identities and is collected along several days. It enhances the validity and capacity of the soft multilabel learning of MAR, and the performance is much better than PUL. TAUDL [49] and the proposed TASTR are tracklet based methods, which do not involve any additional labeled source Re-ID training data. The results show TASTR works the best among these approaches.

C. Ablation Study

We evaluate the effectiveness of different components as shown in Table. V, which are categorized into three groups, *i.e.*, 1) batch hard triplet loss based model trained on Market1501 (make prediction directly) and target dataset, 2) within-camera Re-ID training with or without the TCSCC, and 3) cross-camera Re-ID training by 1-reciprocal Nearest Neighbor based Cross-Camera Tracklets Association (NN-CCTA) with different refinements.

1) Effectiveness of the Training Data in Target Domain: From group 1 we can see that directly deploy the model pre-trained on another dataset to a new dataset will lead to poor performance. And the performance of TASTR-S1 in group 2, which only uses within-camera tracking data for training, is significantly superior to the pre-trained model on Market1501 by a large margin. On the one hand, it is due to the fact that the target domain can be significantly different from the

Group	Dataset / Methods			Campus4K					DukeMTMC-reID				
				rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20	mAP
(1)	BHTri (Market1501)			37.3	57.6	65.8	74.1	39.3	24.1	39.4	46.6	54.2	12.1
	BHTri (Target Dataset)			85.2	96.1	97.8	98.8	87.5	78.1	88.5	91.4	93.8	60.9
(2)	TASTR-S1 w/o TCSCC			5.2	15.5	23.6	33.4	8.6	13.0	20.0	24.9	30.6	6.5
	TASTR-S1			53.0	74.7	83.2	88.3	53.9	56.4	71.5	77.0	82.6	38.4
(3)	STR	k-means	iter*	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20	mAP
				63.4	82.3	88.1	92.9	65.3	67.6	81.3	86.1	88.5	47.2
			✔	61.5	81.6	87.4	91.4	63.2	67.5	80.7	84.6	88.4	46.3
	✔			63.9	82.9	88.2	93.1	65.3	69.4	82.0	85.9	89.4	48.3
	✔		✔	64.0	82.5	88.0	92.0	65.1	67.4	81.0	85.5	88.7	44.6
		✔		53.9	73.0	79.4	84.9	53.7	57.4	73.2	78.9	83.1	37.4
		✔	✔	56.2	78.1	82.7	87.8	59.3	61.1	75.5	80.8	84.9	40.1
	✔	✔		65.3	83.6	88.9	92.4	65.8	69.3	81.8	86.1	89.1	48.1
	✔	✔	✔	**76.4**	**91.6**	**94.6**	**96.4**	**78.3**	**74.1**	**85.5**	**89.0**	**91.8**	**54.4**

TABLE V

COMPARISON OF THE EFFECTIVENESS OF DIFFERENT COMPONENTS. THEY ARE CATEGORIZED INTO THREE GROUPS. GROUP (1): SUPERVISED MODELS TRAINED ON EXTRA OR TARGET DATASET. GROUP (2): WITHIN-CAMERA RE-ID TRAINING. GROUP (3): CROSS-CAMERA RE-ID TRAINING. STR: SPATIO-TEMPORAL REGULARIZATION, "ITER*": PROGRESSIVE OPTIMIZATION (ITERATION).

source domain. On the other hand, limited scale of current Re-ID datasets restricts the generalization capability of the model. However, as mentioned at the beginning, annotating the target data from online surveillance videos of large-scale camera network is strenuous and impractical. It reveals the importance of unsupervised methods that can take advantage of the data obtained in the target environment.

2) Effectiveness of Within-Camera Re-ID Training: Within-camera person tracklets can provide some ID discriminative information for Re-ID learning. However, there are two problems for a pure unsupervised method, *i.e.*, within-camera ID duplication and cross-camera ID duplication. The results in group 2 of Table V show that without TCSCC sampling strategy, Triplet based Re-ID model achieves very poor performance. It is caused by enormous ID duplication problems which make the model training unstable, while hard mining may degrade it since tracklets with different ID labels of the same person are more likely to be hard negatives and contribute to the loss, which may guide the Re-ID learning in a wrong way.

Besides, we conduct an additional experiment to evaluate the within-view ID discrimination capability of TASTR-S1 on Campus4K dataset. For all tracklets in testing set, we randomly remove several frames from the middle third of them to simulate tracking fragmentation. So each tracklet is divided into two sub-tracklets, then we put one into query set and another into gallery set to verify whether the model can rematch them. The rank-1 performances of different models are shown in Fig. 6, and TASTR-S1 is much stronger than other models pre-trained on ImageNet or Market1501. Although the actual situation in real-world application is more complicated, it is worth pointing out that all tracklet data for TASTR-S1 training come from single-camera multi-person tracking, and TASTR-S1 demonstrates great potential for promoting the performance of single-camera multi-person tracking in return.

3) Effectiveness of Cross-Camera Re-ID Training: For reciprocal nearest neighbour cross-camera tracklet association (NN-CCTA), we evaluate the effects and differences of different combinations of STR, k-means and iteration. As shown in

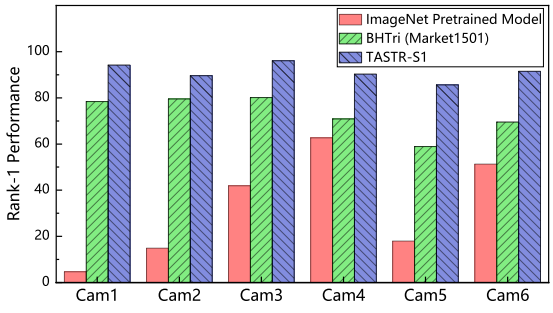
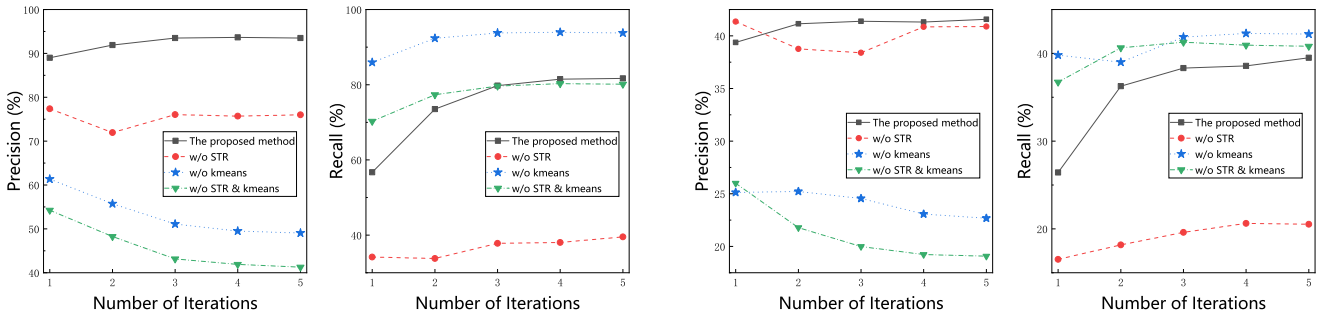


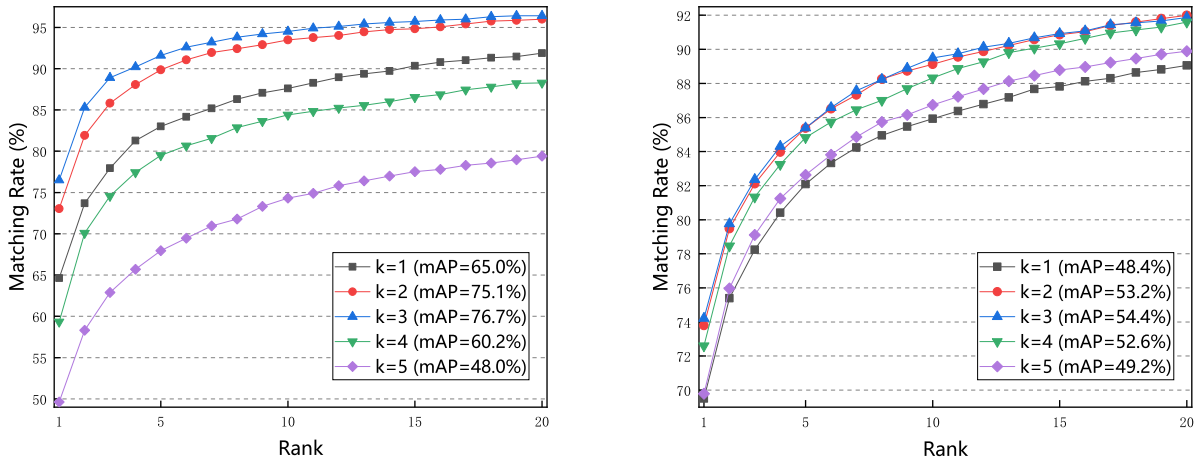
Fig. 6. Illustration of the within-view ID discrimination capability (rank-1 performance) of different models on different cameras on Campus4K dataset.

group 3 of Table V. Based on TASTR-S1, even without Spatio-Temporal Regularization (STR), k-means and iteration, NN-CCTA can get a big improvement over TASTR-S1 because of the importance of cross-view ID discriminative learning. However, when performing it in a progressive optimization way (*i.e.* iteration), the performance gain is limited or even worse. This is caused by the poor precision and quantity of cross-camera matching pairs, which makes the model unstable. If we make some refining such as STR or k-means, the performances may be lower than NN-CCTA. The main reason is that a large number of matching tracklets with potential information useful for more effective learning are temporarily abandoned. However, these operations have greater potential for progressive optimization and TASTR (with STR & k-means) achieves the best performance. It indicates that further refining the cross-camera tracklet association results is significant for progressive optimization as errors may be propagated from wrong matching pairs. More detailed information about how STR and k-means affect the precision and recall of cross-camera tracklet association can be found in Section V-E. Besides, Fig. 10 shows some examples of cross-camera matching tracklet pairs obtained by the final TASTR model on Campus4K and DukeMTMC-reID.



(a) Campus4K (b) DukeMTMC-reID

Fig. 7. The precision and recall of the cross-camera associated tracklet pairs *w.r.t.* different iterations of progressive optimization on (a) Campus4K and (b) DukeMTMC-reID

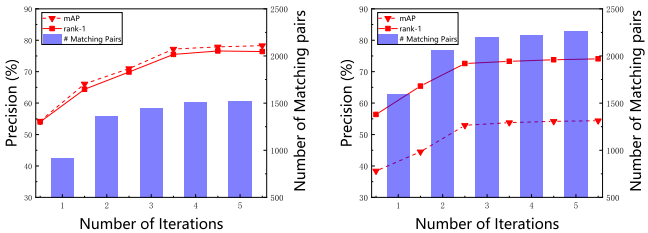


(a) Campus4K (b) DukeMTMC-reID

Fig. 8. CMC curves and mAP corresponding to different k in k-means on (a) Campus4K and (b) DukeMTMC-reID

Models	rank-1	rank-5	rank-10	mAP
TASTR-S1 (unsupervised)	53.0	74.7	83.2	53.9
TASTR-S1 (weakly supervised)	58.6	78.6	84.4	57.8
TASTR (unsupervised)	76.4	91.6	94.6	78.3
TASTR (weakly supervised)	**82.7**	**94.6**	**96.8**	**84.4**

TABLE VI
EVALUATION OF WEAKLY SUPERVISED LEARNING ON CAMPUS4K



(a) Campus4K (b) DukeMTMC-reID

Fig. 9. The performance and the number of associated tracklet pairs *w.r.t.* different iterations (0 means TASTR-S1) of progressive optimization on (a) Campus4K and (b) DukeMTMC-reID

D. Weakly Supervised Learning

In person Re-ID training data annotation, the process can be divided into two steps: 1) per-camera person annotation, 2) cross-camera person matching. The first step is to track pedestrians manually, which is quite easy for human. The second step requires exhaustive manual search of cross-camera person matching, which is the most costly procedure as it is hard to know when and where a specific person will appear given complex camera network topology and unconstrained people's moving in the public spaces. Thus, per-camera ID labeling is much cheaper, and such labeled data are much weaker due to the lack of cross-camera ID pairs information. This setting in person Re-ID can be called weakly supervised learning.

The proposed TASTR can be easily applied in the weakly supervised setting. In this setting, per-camera person ID information is given, so there is almost no within-camera ID duplication problem due to no trajectory fragmentation. Furthermore, it allows to test how much performance gain the Re-ID model can get from perfect within-camera multi-person detection and tracking. Table VI shows the results of

TASTR-S1 and TASTR under weakly supervised setting on Campus4K. This demonstrates the suitability of our method for different labeling settings in real-world scenarios.

E. Error Analysis of Tracklet Association

We evaluate different components in cross-camera tracklet association. Since no label information is available on Campus4K when using automatically generated person tracklet data, it is difficult to obtain the accuracy and recall of matching pairs. Therefore, the above results are obtained under weakly supervised setting defined in Section V-D. The precision and recall of the associated tracklet pairs over iterations are shown in Fig. 7. Without k-means and STR, the model would get relatively high recall but the lowest precision. STR (*i.e.* w/o k-means) can improve both the precision and recall rate of NN-CCTA, while k-means removes a large number of matching tracklet pairs with large distances, which results in higher precision but lowest recall. The proposed method TASTR (NN-CCTA with STR and k-means) achieves the highest precision as well as high recall rate, and the recall grows steadily while the precision remains at a high level in the process of progressive optimization, which leads to the best Re-ID performance. This observation demonstrates the effectiveness of our method, which can acquire more training data while achieving higher precision.

F. Parameter Sensitivity Study

In our TASTR, there are two important parameters, *i.e.*, k in k-means clustering and n which denotes the iteration number of progressive optimization. In the following, we study the sensitivity of our approach to the setting of the two parameters.

As shown in Fig. 8, when $k = 1$, which means all matching tracklet pair candidates are used for progressive training, the performance of Re-ID model is poor as it introduces more incorrectly matching pairs. With k increases, it becomes more and more strict for cross-camera tracklet association, thus leads to fewer and fewer training data for cross-view discriminative learning. The best performance is obtained when $k = 3$, which indicates a balance between the accuracy of cross-camera tracklet matching and the sufficiency of cross-view training data.

To evaluate the effectiveness of progressive optimization, we train our model with the best setting and observe its performance and the number of associated tracklet pairs in different iterations. The mAP and rank-1 performance before the first iteration means the performance of TASTR-S1 model. As shown in Fig. 9, both the performance and the number of associated tracklet pairs gain considerable improvement in the first three iterations of cross-view optimization, and then keeps relatively stable. In order to get the best results, we set $n = 5$ in our approach.

VI. Conclusion

In this work, we are dedicated to unsupervised person Re-ID and contribute a new high-quality Campus4K dataset with full frames and full spatio-temporal information. We propose a new progressive learning method by Tracklet Association with Spatio-Temporal Regularization (TASTR). Instead of assuming perfect person tracking in each camera view, we automatically explore the identity discriminative information from imperfect tracklets with spatio-temporal context. We initially learn a Re-ID model with triplets constructed with spatio-temporal constraint. Then, we associate the tracklets across each camera pair which is further used to fine-tune and update the initial Re-ID model. The above processes are iterated to progressively improve the discriminative capability of the Re-ID model. We make evaluations on two datasets and extensive experiments demonstrate the effectiveness of our approach.

References

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 152–159.

[2] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3908–3916.

[3] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia (TMM)*, vol. 18, no. 12, pp. 2553–2566, 2016.

[4] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3219–3228.

[5] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[6] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 7, pp. 3492–3506, 2017.

[7] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong, and N. Zheng, "Large margin learning in set-to-set similarity comparison for person reidentification," *IEEE Transactions on Multimedia (TMM)*, vol. 20, no. 3, pp. 593–604, 2017.

[8] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2285–2294.

[9] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 79–88.

[10] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1179–1188.

[11] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia (TMM)*, 2019.

[12] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *IEEE Transactions on Image Processing (TIP)*, 2019.

[13] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[14] 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 (CVPR)*, 2017, pp. 3415–3424.

[15] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2158–2167.

[16] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proceedings of the British Machine Vision Conference (BMVC)*, vol. 2, no. 3, 2012, p. 8.

(a) Campus4K

(b) DukeMTMC-reID

Fig. 10. Example cross-camera matching tracklet pairs obtained by the proposed method on (a) Campus4K, (b) DukeMTMC-reID. "C1-C2": Matching pairs between Camera1 and Camera2.

[17] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Satoh, "Learning sparse and identity-preserved hidden attributes for person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 29, no. 1, pp. 2013–2025, 2019.

[18] G. Lisanti, N. Martinel, A. Del Bimbo, and G. Luca Foresti, "Group re-identification via unsupervised transfer of sparse features encoding," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2449–2458.

[19] Z. Huang, Z. Wang, W. Hu, C.-W. Lin, and S. Satoh, "Dot-gnn: Domain-transferred graph neural network for group re-identification," in *Proceedings of the ACM International Conference on Multimedia (MM)*, 2019, pp. 1888–1896.

[20] 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 (TOMM)*, vol. 14, no. 4, p. 83, 2018.

[21] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7948–7956.

[22] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *Proceedings of the British Machine Vision Conference (BMVC)*, vol. 3, 2015, p. 8.

[23] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 769–773.

[24] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification

by unsupervised l_1 graph learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 178–195.

[25] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 2, pp. 356–370, 2017.

[26] 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 (ICCV)*, 2017, pp. 2429–2438.

[27] 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 (ICCV)*, 2017, pp. 5142–5150.

[28] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong, "Person re-identification by unsupervised video matching," *Pattern Recognition (PR)*, vol. 65, pp. 197–210, 2017.

[29] W. Zhang, B. Ma, K. Liu, and R. Huang, "Video-based pedestrian re-identification by adaptive spatio-temporal appearance model," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 4, pp. 2042–2054, 2017.

[30] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4733–4742.

[31] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[32] Y. Liu, Z. Yuan, W. Zhou, and H. Li, "Spatial and temporal mutual promotion for video-based person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 8786–8793.

[33] W. Huang, R. Hu, C. Liang, Y. Yu, Z. Wang, X. Zhong, and C. Zhang, "Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations," in *International Conference on Multimedia Modeling (MMM)*, 2016, pp. 174–186.

[34] N. Martinel, G. L. Foresti, and C. Micheloni, "Person reidentification in a distributed camera network framework," *IEEE transactions on cybernetics*, vol. 47, no. 11, pp. 3530–3541, 2017.

[35] Y.-J. Cho, S.-A. Kim, J.-H. Park, K. Lee, and K.-J. Yoon, "Joint person re-identification and camera network topology inference in multiple cameras," *Computer Vision and Image Understanding*, vol. 180, pp. 34–46, 2019.

[36] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[37] 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 (CVPR)*, 2018, pp. 5177–5186.

[38] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2360–2367.

[39] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2012.

[40] ——, "Local descriptors encoded by fisher vectors for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012, pp. 413–422.

[41] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3550–3557.

[42] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3586–3593.

[43] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1306–1315.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[45] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 475–491.

[46] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2275–2284.

[47] 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 (CVPR)*, 2018, pp. 994–1003.

[48] 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 (CVPR)*, 2019, pp. 2148–2157.

[49] 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 (ECCV)*, 2018, pp. 737–753.

[50] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008, pp. 262–275.

[51] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*, 2011, pp. 91–102.

[52] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 688–703.

[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[54] 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*. Springer, 2016, pp. 868–884.

[55] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3754–3762.

[56] X. Lan, X. Zhu, and S. Gong, "Person search by multi-scale matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 536–552.

[57] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[58] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2334–2343.

[59] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[60] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[61] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.

[62] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[63] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1318–1327.

[64] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1988–1995.

[65] 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 the European Conference on Computer Vision (ECCV)*, 2016, pp. 17–35.

[66] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

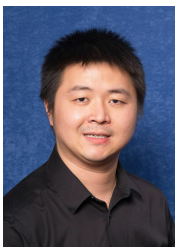
[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[69] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2197–2206.
[70] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–188.

Qiaokang Xie received the B.E. degree in electronic information engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2017. He is currently pursuing the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering and Information Science, USTC.

His research interests include person re-identification and computer vision.

Wengang Zhou received the B.E. degree in electronic information engineering from Wuhan University, China, in 2006, and the Ph.D. degree in electronic engineering and information science from University of Science and Technology of China (USTC), China, in 2011. From September 2011 to 2013, he worked as a post-doc researcher in Computer Science Department at the University of Texas at San Antonio. He is currently a Professor at the EEIS Department, USTC.

His research interests include multimedia information retrieval and computer vision.



Guo-Jun Qi (M14-SM18) is the Chief Scientist leading and overseeing an international R&D team in the domain of multiple intelligent cloud services, including smart cities, visual computing service, medical intelligent service, and connected vehicle service at Futurewei, since August 2018. He was a faculty member in the Department of Computer Science and the director of MAchine Perception and LEarning (MAPLE) Lab at the University of Central Florida since August 2014. Prior to that, he was also a Research Staff Member at IBM T.J. Watson Research Center, Yorktown Heights, NY.

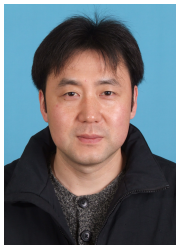
His research interests include machine learning and knowledge discovery from multi-modal data sources (e.g., images, videos, texts, and sensors) in order to build smart and reliable information and decision-making systems. His research has been sponsored by grants and projects from government agencies and industry collaborators, including NSF, IARPA, Microsoft, IBM, and Adobe.

Dr. Qi has published over 150 papers in a broad range of venues, such as Proceedings of IEEE, IEEE T PAMI, IEEE T KDE, IEEE T Image Processing, ICML, NIPS, CVPR, ECCV, ACM MM, SIGKDD, WWW, ICDM, SDM, ICDE and AAAI. Among them are the best student paper of ICDM 2014, "the best ICDE 2013 paper" by IEEE Transactions on Knowledge and Data Engineering, as well as the best paper (finalist) of ACM Multimedia 2007 (2015).

Qi Tian is currently a Chief Scientist in Artificial Intelligence at Cloud BU, Huawei. From 2018-2020, he was the Chief Scientist in Computer Vision at Huawei Noah's Ark Lab. He was aslo a Full Professor in the Department of Computer Science, the University of Texas at San Antonio (UTSA) from 2002 to 2019. During 2008-2009, he took one-year Faculty Leave at Microsoft Research Asia (MSRA).

Dr. Tian received his Ph.D. in ECE from University of Illinois at Urbana-Champaign (UIUC) and received his B.E. in Electronic Engineering from Tsinghua University and M.S. in ECE from Drexel University, respectively. Dr. Tian's research interests include computer vision, multimedia information retrieval and machine learning and published over 530 refereed journal and conference papers. His Google citation is over 19300+ with H-index 69. He was the co-author of best papers including IEEE ICME 2019, ACM CIKM 2018, ACM ICMR 2015, PCM 2013, MMM 2013, ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, a Student Contest Paper in ICASSP 2006, and co-author of a Best Paper/Student Paper Candidate in ACM Multimedia 2019, ICME 2015 and PCM 2007.

Dr. Tian research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar and UTSA. He received 2017 UTSA President's Distinguished Award for Research Achievement, 2016 UTSA Innovation Award, 2014 Research Achievement Awards from College of Science, UTSA, 2010 Google Faculty Award, and 2010 ACM Service Award. He is the associate editor of IEEE TMM, IEEE TCSVT, ACM TOMM, MMSJ, and in the Editorial Board of Journal of Multimedia (JMM) and Journal of MVA. Dr. Tian is the Guest Editor of IEEE TMM, Journal of CVIU, etc. Dr. Tian is a Fellow of IEEE.



Houqiang Li (S'12) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively, where he is currently a Professor with the Department of Electronic Engineering and Information Science.

His research interests include multimedia search, image/video analysis, video coding and communication. He has authored and co-authored over 200 papers in journals and conferences. He is the winner of National Science Funds (NSFC) for Distinguished Young Scientists, the Distinguished Professor of Changjiang Scholars Program of China, and the Leading Scientist of Ten Thousand Talent Program of China. He served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2013. He served as the TPC Co-Chair of VCIP 2010, and he will serve as the General Co-Chair of ICME 2021. He is the recipient of National Technological Invention Award of China (second class) in 2019 and the recipient of National Natural Science Award of China (second class) in 2015. He was the recipient of the Best Paper Award for VCIP 2012, the recipient of the Best Paper Award for ICIMCS 2012, and the recipient of the Best Paper Award for ACM MUM in 2011.