This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

A Survey of Open-World Person Re-identification

Qingming Leng, Mang Ye, Qi Tian*, *Fellow, IEEE*

***Abstract*—Person re-identification (re-ID) has been a popular topic in computer vision and pattern recognition communities for a decade. Several important milestones such as metric-based and deeply-learned re-ID in recent years have promoted this topic. However, most of existing re-ID works are designed for closed-world scenarios rather than realistic open-world settings, which limits the practical application of re-ID technique. On one hand, the performance of the latest re-ID methods have surpassed the human-level performance on several commonly used benchmarks (e.g. Market1501 and CUHK03), which are collected from closed-world scenarios. On the other hand, open-world tasks that are less developed and more challenging have received increasing attention in the re-ID community. Therefore, this study starts the first attempt to analyze the trends of open-world re-ID and summarizes them from both narrow and generalized perspectives. In the narrow perspective, open-world re-ID is regarded as person verification (i.e., Open-set re-ID) instead of person identification, that is, the query person may not occur in the gallery set. In the generalized perspective, application-driven methods that are designed for specific applications are defined as generalized open-world re-ID. Their settings are usually close to realistic application requirements. Specifically, this survey mainly includes the following four points for open-world re-ID: (1) analyzing the discrepancies between closed- and open-world scenarios; (2) describing the developments of existing Open-set re-ID works and their limitations; (3) introducing specific application-driven works from three aspects, namely, raw data, practical procedure, and efficiency; (4) summarizing state-of-the-art methods and future directions for open-world re-ID. This survey on open-world re-ID provides a guidance for improving the usability of re-ID technique in practical applications.**

***Index Terms*—Person re-identification, open-world, closed-world, open-set, specific application-driven**

I. Introduction

Person re-identification (re-ID), a popular topic in computer vision and pattern recognition communities, has gained increasing interest in the academia and industry, especially in recent years [1]. The core issue of re-ID is to seek the occurrences of a query person (probe) from a set of person candidates (gallery), where probe and gallery are captured from different non-overlapping camera views. Re-ID is an extremely challenging task because the appearance discrepancies of the same person might be even larger than that of different persons. The variance is usually caused by significant changes in viewpoint, illumination, resolution, occlusion, color, and so on across camera views (see Fig. 1). Biometric cues, such as face and gait, are usually infeasible in practical non-restrictive video surveillance systems due to low-resolution cameras [2]. The re-ID problem in this survey pertains to a task that only relies on the visual appearance of a person [1].

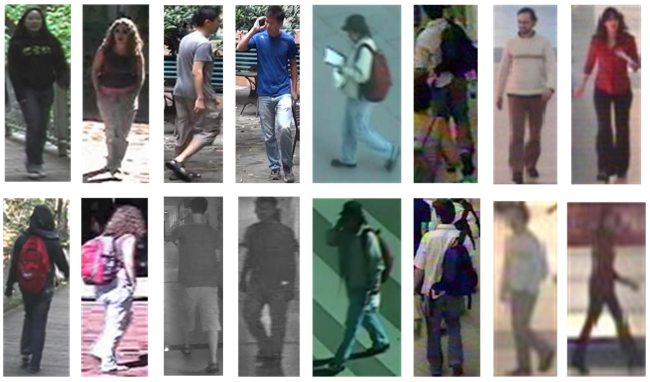

Fig. 1: Person image samples derived from publicly standard VIPeR [3], SYSU-MM01 datasets [4], PRID2011 [5], GRID [6], and CAVIAR4ReID [7]. Each column represents the same person images, and each row represents images observed from the same camera views. As shown, the appearances of same-person images change severely in different camera views due to obvious variations in viewpoint, illumination, color, occlusion, and resolution.

Re-ID was initially studied as a sub-task of multi-camera tracking in early years [8]. *Javed et al.* [9] examined a subspace of inter-camera brightness transfer functions to address the change in observed colors of a pedestrian for multiple non-overlapping camera tracking. In 2006, *Gheissari et al.* [10] firstly defined "person re-identification" as an independent research topic that aims to match persons captured from non-overlapping cameras with their visual appearance. With the development of large-scale datasets and evaluation metrics, re-ID has become a popular research topic in computer vision community. Many studies have been published in top-tier venues [2].

A. Milestones of Existing Re-ID Studies

Through the unremitting efforts of computer vision researchers, re-ID has achieved remarkable success in different aspects. We draw a timeline to introduce important milestones for re-ID and present it in Fig. 2. We also provide a brief introduction of existing re-ID studies from the following methodology-driven perspectives; further details can be found in another survey [1].

Single-shot vs. Multi-shot. Although the visual appearance of a person is captured in surveillance video, most early re-ID studies paid attention to matching single-shot pedestrian

The research was supported by the National Nature Science Foundation of China (61562048).

*Qi Tian is the corresponding author.

Qingming Leng is with the School of Information Science and Technology, Jiujiang University, Jiujiang, China. (e-mail:qingming.leng@gmail.com).

Mang Ye is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. He is the co-first author of this article. (e-mail:mangye@comp.hkbu.edu.hk).

Qi Tian is with the Department of Computer Science, University of Texas at San Antonio, TX, 78256 USA. (e-mail:qitian@cs.utsa.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

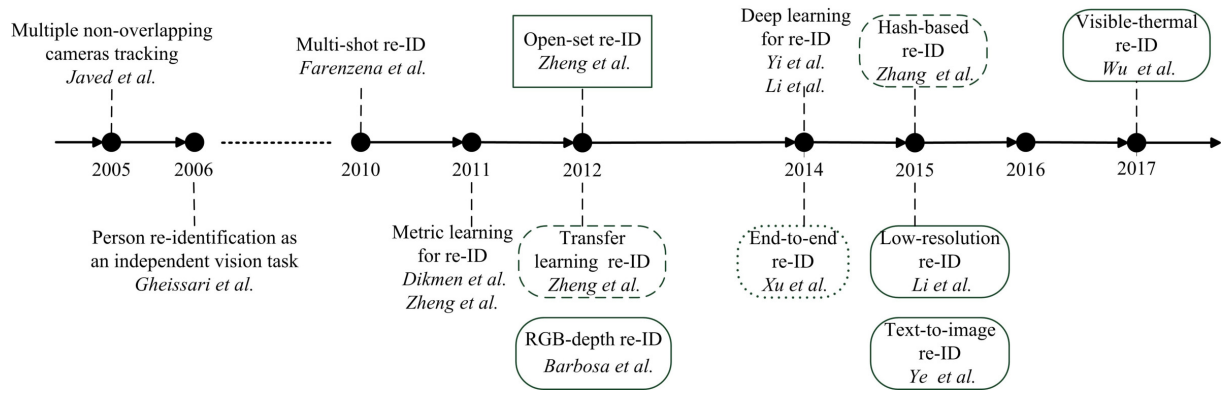

Fig. 2: Milestones of existing re-ID studies and new open-world trends (bounded with boxes). As shown in the figure, traditional re-ID orientations started before 2014. Many new open-world trends have been investigated in recent years due to the increasing interest in the application of re-ID technique. All open-world re-ID trends are highlighted in this figure. In particular, open-set re-ID as the narrow open-world study is marked by solid rectangle, and specific application-driven re-ID as a generalized open-world study is marked by various oval rectangles. Raw data-driven methods such as Visible-thermal re-ID, RGB-depth re-ID, Text-to-image re-ID, and Low-resolution re-ID, are marked by solid lines. Practical procedure-driven methods, such as End-to-end re-ID, are maked by dotted lines. Efficiency-driven methods, such as hash-based re-ID and Transfer learning re-ID, are marked by imaginary lines.

images (also called image-based) rather than multi-shot image sets (also called video-based) [11]. Single-shot means that each associated pair of images includes an instance of an individual. By contrast, multiple images or sequences of the same person can be obtained within each camera to represent this person for multi-shot re-ID. In 2010, *Farenzena et al.* [12] conducted initial multi-shot re-ID work, which involved designing several partial appearance features by exploiting symmetry and asymmetry perceptual principles.The minimum Bhattacharyya distance among multi-shot bounding boxes was calculated for person matching. The researchers found that multiple frames provide more appearance information than the single-shot setting. Therefore multi-shot re-ID has received increasing attention as a candidate for improving re-ID accuracy.

Feature-based vs. Metric-based. Similar to instance retrieval [13], the main processes of re-ID include feature extraction and distance measurement [14, 15]. Early re-ID studies were mainly feature-based and aimed to design robust appearance descriptions for distinguishing different pedestrians across arbitrary cameras [16]. However, due to illumination variation, viewpoint change, and low resolution, discriminative feature construction becomes challenging. Consequently, feature learning methods based on deep neural network [17] have become a popular feature-based paradigm for practicing improved feature representation. Metric-based re-ID emerged in 2011 when *Dikmen et al.* [18] and *Zheng et al.* [16] utilized metric learning to learn an optimal distance metric that maximizes the probability of a true matched pair having a smaller distance than wrong matched pairs. Afterward, a mass of researches focused on learning a discriminative similarity measurement [19, 20].

Hand-crafted vs. Deeply-learned. In the early re-ID system, low-level features or high-level semantic attributes, global representations, and local descriptions were extracted through sophisticated but time-consuming hand-crafted techniques. Performance relied heavily on superior human experiences. However, when the Convolution Neural Network (CNN) was introduced to re-ID in 2014, the methodology of feature extraction completely changed. *Yi et al.* [21] and *Li et al.* [17] employed a Siamese neural network to learn optimal pedestrian image features from training data automatically. They found that the deeply-learned paradigm almost dominates the re-ID feature learning procedure due to its advantages in end-to-end learning [22–25]. Moreover, deep architecture is not specifically designed for feature extraction, and deep metric re-ID approaches have also elicited much attention [21, 26].

B. Re-ID in an Open-world Scenario

To our knowledge, the re-ID task was developed for public safety and security applications [27], such as video investigation that seeks the occurrences and trajectories of a specific person in a practical video surveillance system. Hence, with the increasing interest in the application of re-ID, several new trends those have been developed for practical open-world scenarios in recent years as shown in Fig. 2, such as Open-set re-ID, Visible-thermal re-ID, Low-resolution re-ID. The performance of several recent re-ID methods in the closed-world scenario has surpassed human-level accuracies on several commonly-used benchmarks. For example, the rank-1 accuracy of AlignedReID [28] has reached 94.4% and exceeds the 93.5% accuracy of human beings on the Market-1501 dataset. On another benchmark CUHK03, AlignedReID obtained 97.8% rank-1 accuracy, which is better than the 95.7% accuracy of human performance. Therefore, research efforts must be devoted to the less developed but more challenging open-world task. In this survey, we provide a comprehensive analysis and discussion of open-world re-ID from narrow and generalized perspectives.

Open-world re-ID in a narrow perspective. Open-world re-ID was introduced in [2], and it refers to the scenario where large-scale person identities across cameras may only partially overlap in an unknown spatial environment. In other words, re-ID focuses on verifying whether a probe person is in a gallery or not, and it is likely to be a person verification task that can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

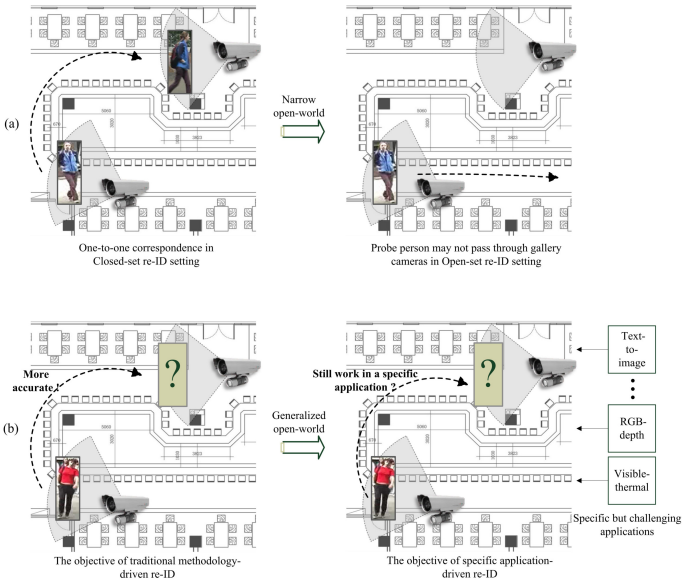

Fig. 3: Illustration of the characteristics of open-world re-ID from narrow and generalized perspectives. (a) From Closed-set re-ID to Open-set re-ID. One-to-one identity correspondence exists in close-set setting, whereas the probe person may not appear in the target cameras in Open-set setting. (b) From methodology-driven re-ID to specific application-driven re-ID. The objective of traditional methodology-driven methods is to pursue highly accurate results. By contrast the primary goal of specific application-driven trends is to make re-ID work in a specific challenging application.

also be called "Open-set re-ID." In comparison, the closed-set scenario [29] usually assumes that the query person exists in the gallery set. Therefore, the focus is to determine which gallery image belongs to the query person, i.e., returning the ID of the query person. This assumption not always holds, i.e., the query person may not appear in the gallery set captured from other camera views. Thus, Open-set re-ID is closer to reality than Closed-set re-ID.

Open-world in a generalized perspective. Many re-ID benchmarks have been constructed based on a limited number of persons, selective image data, and analogic camera settings due to workload, cost, privacy, etc. [29, 30]. These are far away from the realistic requirements in practical applications. As a result, the majority of existing re-ID studies adopted above benchmark configurations and paid more attention to methodology-driven approaches that are likely to improve the former methods rather than develop application-driven techniques (see Fig. 3). That is to say, a re-ID work motivated by specific practical applications could belong to the open-world paradigm in the generalized perspective, even though it does not use the term "open-world" or "open-set."

C. Organization of This Survey

Different from other well-known re-ID surveys [1, 8, 29–32] that focused on introducing methodology-driven features and metrics for closed-world re-ID, this survey presents the first overall review of open-world re-ID that is crucial for practical re-ID applications. In particular, we categorize existing open-world studies into narrow and generalized perspectives, and this categorization provides guidance in understanding the current development of open-world re-ID. Specifically, this survey presents an overview of Open-set re-ID in Section II. Existing generalized open-world studies that were designed for specific applications are discussed in Section III. The generalized open-world re-ID is categorized into three specific application-driven viewpoints, namely, raw data, practical procedure, and efficiency. State-of-the-arts methods and benchmarks for open-world re-ID trends are discussed in Section IV, and future open-world re-ID is presented in Section V. The conclusions are provided in Section VI.

TABLE I: Comparisons of Closed-set and Open-set Re-ID Settings.

Re-ID	Closed-set	Open-set
Property	Identification	Verification
Objective	Seeking the most similar probe person in the gallery	Verifying whether the probe appears in a certain gallery
Data setting	Exactly similar identity correspondence between cameras	Numerous irrelevant persons, and the probe person may not be in the gallery
Evaluation metrics	CMC, mAP	TTR, FTR

II. Open-set Re-ID: a Narrow Open-world Perspective

Open-set re-ID is a task related to Close-set re-ID. As an identification task, Closed-set re-ID aims to determine which gallery image is the probe person based on the precondition that an exactly similar identity correspondence exists between cameras. However, in the open-set setting, numerous irrelevant persons exist, and the probe person may not be in the gallery [33]. In other words, Open-set re-ID can be concluded as person verification [29] that answers the question "Does the probe appear in a certain gallery and in which images?" If we compare Closed-set re-ID with Open-set re-ID from the viewpoint of evaluation, then Cumulative Matching Characteristics (CMC) curve [34], which represents the probability of first match in the top-k gallery ranking list is an accurate and suitable evaluation metric for the closed-set scenario due to the inevitable ground truth. On the contrary, open-set evaluation metrics should be independent of one-to-one identity correspondence, such as high true target recognition (TTR) and low false target recognition (FTR) [35] that focus on calculating the likelihood of the number of query target and non-target images being verified as target identities. To summarize the disparities between Closed-set re-ID and Open-set re-ID, a comprehensive comparison is illustrated in Table I.

In 2012, *Zheng et al.* [36] conducted the first Open-set re-ID work based on a transfer ranking framework for set-based

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

verification. Afterward, *Cancela et al.* [37] addressed the open-set task via online Conditional Random Field (CRF) inference. *Liao et al.* [38] decomposed Open-set re-ID into detection and identification, and two generic evaluation metrics (i.e., identification rate and false acceptance rate) were discussed. *Wang et al.* [39] proposed a regularized kernel subspace learning model for one-shot verification by learning cross-view identity-specific information from unlabeled data alone. *Zheng et al.* [35] presented a group-based setting and a transfer local relative distance comparison model for conquering label scarcity. TTR and FTR were initially utilized for performance evaluation. Compared with previous methods, this work presented clearer descriptions of open-set challenges, standard evaluation metrics, and the integrated framework for individual and group verification; it is highly representative for Open-set re-ID. *Zhu et al.* [40] introduced a large-scale setting characterized by huge size probe images and an open person population. A hashing approach called cross-view identity correlation and verification was adopted to learn cross-view identity representation binarization and discrimination in a joint manner.

Similar to typical open-set works described above, Multi-Target Multi-Camera Tracking (MTMCT) aims to determine the cross-camera trajectories of certain pedestrians captured from multiple cameras [41–43], and it confronts many fundamental challenges that are similar to the major problems of Open-set re-ID. For example, a mass of video data that includes enormous cameras and pedestrians exists, and numerous camera views do not always overlap, leading that spatio-temporal cues being unavailable when people go through disjoint multi-camera views that are often separated in time and space. In addition, the amount of pedestrians is uncertain or unknown in multiple cameras, so MTMCT and Open-set re-ID could also fall under the cross-camera identity verification task. In 1997, *Huang and Russell* [44] presented a Bayesian-based tracking model with color and spatio-temporal features, and it is the first work of multiple non-overlapping camera tracking using appearance features. *Javed et al.* [45] and *Chen et al.* [46] integrated brightness transfer functions with spatio-temporal features for multiple non-overlapping camera tracking. Recently, the deeply-learned paradigm was also introduced to MTMCT, similar to re-ID. *Tesfaye et al.* [47] proposed a unified three-layer hierarchical framework that includes two within-camera layers and one across-camera layer. *Ristani et al.* [48] introduced a CNN with adaptive weighted triplet loss, which was utilized to learn robust features for both MTMCT and re-ID. However, MTMCT still has several differences from Open-set re-ID; for example, the former focuses on reducing classification error rates, whereas the latter pays attention to improving the ranking performance at a certain error-tolerant rate. A generic loss of MTMCT is to minimize the distance between any two co-identical objects less than the smallest distance between any two non co-identical objects. By comparison, the loss of re-ID is only to minimize the distance between different images of any object "a" less than the smallest distance between "a" and other objects.

Although the Open-set re-ID is much closer to practical video surveillance applications than the closed-set setting, the attention devoted to this issue is relatively limited for two possible reasons. Firstly, Closed-set re-ID is a mature technology [30], and it is convenient and fair for conducting research on purposed-based works due to various baselines, datasets, and evaluations. Secondly, the low recognition rates under low false accepted rates of existing results show that Open-set re-ID is challenging [1]. Fortunately, new specific application-driven re-ID directions that pay attention to the realistic environment. Most of these studies are not titled "open-set" or "open-world" but still follow the identity correspondence setting as usual. In a generalized perspective, they could pertain to open-world re-ID because their ideas are motivated by certain demands in practical applications instead of methodology improvements solely. Therefore, we comprehensively summarize and analyze existing specific application-driven re-ID works as generalized open-world re-ID in the following section.

III. Specific Application-driven Re-ID: a Generalized Open-world Perspective

This section mainly talks about generalized open-world re-ID from specific applications rather than existing methodology-driven methods for closed-set scenarios. These specific applications are close to realistic applications in different aspects. In particular, we examine video investigation, which may be the most typical re-ID application, and discuss the main possible aspects of application requirements.

Assuming that images of suspects are recorded by unconstrained cameras [39], and procedure of video investigation application based on re-ID technique can be illustrated as shown in Fig. 4. First, all useful and available person candidates (gallery) captured from visible cameras, infrared cameras, mobile phones, or other photographing equipment are collected. That is, obtaining raw person data is the primary requirement of practical video investigation. Second, the raw person data, especially the principal surveillance videos, are cropped into bounding boxes based on pedestrian detection and tracking for a subsequent re-ID procedure. However, surveillance cameras are almost located in a long-range and busy public space. Thus, they are more likely to produce a large amount of irrelevant or incomplete bounding boxes due to low resolution, partial occlusion, and so on. The effects of person detection and tracking procedures must therefore be considered in a practical re-ID system. Third, if the investigator has obtained suspects' images, or sequences, even eyewitness textual statements (probe), then the objective of video investigation turns into verifying/identifying whether the probe appears in a range of cameras. The gallery would usually generate a large amount of surveillance videos; hence, efficiency may be more important than accuracy for large-scale re-ID.

Considering the three aspects of video investigation, i.e., raw data, practical procedure, and efficiency, traditional closed-set methodology-driven studies usually focused on single- and multi-shot visible person images and hand-cropped bounding boxes but disregarded the efficiency issue. All three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

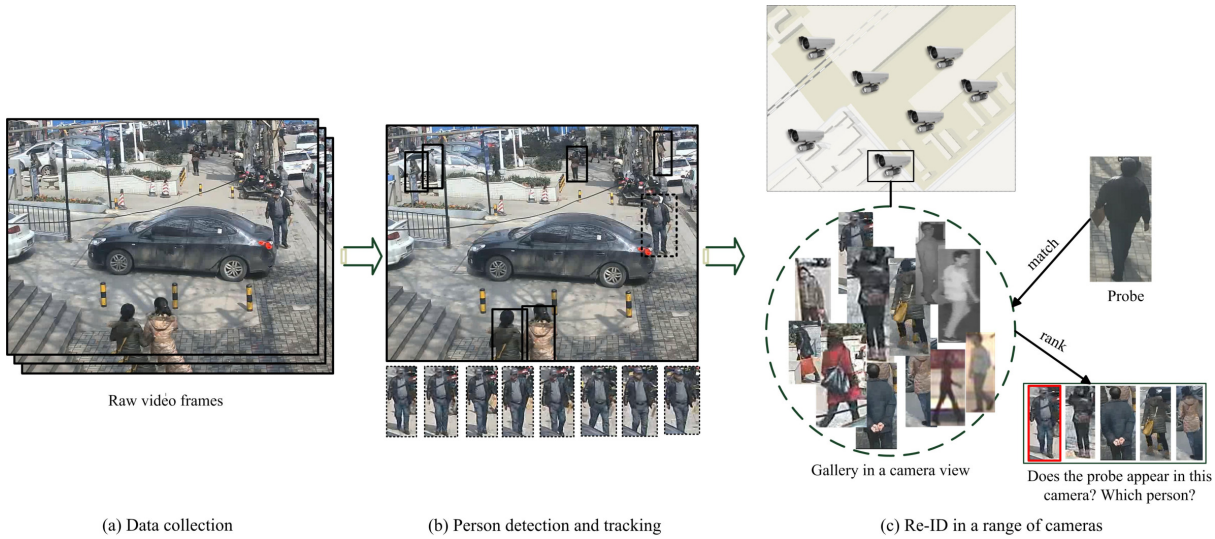

Fig. 4: Illustration of the re-ID technique for video investigation application. (a) Data collection. A larege number of useful and available person candidates (gallery) are collected from various photographing equipment, such as surveillance cameras. (b) Person detection and tracking. Raw video frames are turned into massive pedestrian bounding boxes. (c) Re-ID in a range of cameras. Suspects probably flee across multiple blocks, so if the investigator has obtained images, sequences, or even eyewitness textual statements of suspects (probe), then the objective of video investigation turns into verifying/identifying whether the probe appears in the large-scale gallery under uncertain camera views.

need to be further developed for approximating the realistic scenario. In the meantime, the interest in studying specific application-driven re-ID, which is closer to the practical environment and treated as generalized open-world re-ID in this survey, has increased. For a more clear presentation, existing specific application-driven works are discussed according to the three aspects.

A. Raw Data

What do person data look like in existing re-ID benchmarks? Fig. 1 presents several intuitive examples from public standard datasets. To provide quantitative descriptions, we focus on several commonly-used datasets in the top venues of computer vision in recent years, i.e., VIPeR [3], GRID [6], PRID2011 [5], CAVIAR4ReID [7], CUHK01 [49], CUHK03 [17], Market-1501 [50], MARS [51], iLIDS-VID [52], and DukeMTMC-reID [49]. These datasets are summarized and analyzed from three points, namely, including basic information, camera setting, and typical person appearance variations, as shown in Table II. Several interesting observations are acquired. Positive observations: (1) Half of the prevalent datasets were released within the last four years, and the attention on the re-ID task in recent years has grown rapidly relative to more than a decade of re-ID history. (2) The data volume of re-ID benchmarks has improved greatly. Early VIPeR and GRID had only hundreds of identities and thousands of images, whereas more than thousands of identities and millions of images are produced on the newer MARS dataset. With respect to the growth trend of re-ID study, large size and ease of use are probably the key factors for broadcasting the re-ID community. (3) Manyf existing benchmarks consider the influence of common person appearance changes, such as variations in viewpoint, illumination, resolution, and occlusion, across camera views. Negative observations: (1) Single-modality visible person images (RGB images) always exist on most datasets, whereas multi-modality information, including visible images, thermal images, depth images, and eyewitness textual statements, etc. usually exist in practical application (e.g., video investigation). (2) The majority of re-ID datasets label pedestrian bounding boxes based on a hand-cropped strategy, which is too time-consuming to be feasible in practice. (3) Although an analogic camera setting could be an economical alternative, it is not good that more than half of the datasets adopt an analogic camera setting rather than a realistic surveillance camera network. Market-1501 and MARS, two of the three most frequently used datasets in recent years, were constructed by using several horizontally located cameras, the viewpoints of which are evidently different from those of top-down urban surveillance cameras. (4) The person images on many datasets have the same image resolution, which is inconsistent with significant resolution changes in the practical environment. Although several benchmarks provide original bounding boxes with various resolutions, most existing re-ID studies usually resized the input images into a unified scale, which possibly results in the loss of appearance information (down-sampling) or the addition of more noises (up-sampling).

Notably, data collection and preprocessing are time-consuming and labor-intensive, and we would like to commend those who attempted to develop extensive and meaningful re-ID datasets. Nevertheless, according to our negative observations, the person images in existing benchmarks still cannot be adopted in practical applications. Aside from ongoing efforts to build realistic datasets, several recent studies have put forward the practicability of re-ID by mining the characteristics of raw multi-modality (e.g., thermal, depth, and text) and low-resolution person data in addition to general visible images, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II: Comparisons of Basic Information, Camera Setting, and Typical Person Appearance Variations of Representative Re-ID Datasets.

Datasets	Basic information					Camera Setting	Person Appearance Variations				
	Time	#IDs/-Cameras	#Images	Label	Full frames		Viewpoint	illumination	Resolution	Occlusion	Color
VIPeR	2007	632/2	1,264	hand		analogic	+	+	128×48		
GRID	2009	250/8	1,275	hand		realistic	+	+	vary	+	+
PRID2011	2011	200/2	24,541	hand	+	analogic	+	+	128×64		+
CAVIAR-4ReID	2011	72/2	1,220	hand		realistic	+	+	vary		
CUHK01	2012	971/2	3,884	hand		analogic	+	+	160×60		
CUHK03	2014	1,467/2	13,164	hand/DPM		analogic	+	+	vary	+	
Market-1501[1]	2015	1,501/6	32,668	hand/DPM		analogic[2]	+	+	128×64	+	
MARS[1]	2016	1,261/6	1,191,003	DPM/-GMMCP		analogic[2]	+	+	256×128	+	
iLIDS-VID	2016	300/2	42,495	hand		realistic	+	+	vary	+	
DukeMT-MCreID[1]	2017	1,812/8	36,441	Doppia	+	realistic	+	+	vary	+	

[1] These three datasets are the most frequently used datasets in the last two years.
[2] Cameras in these two datasets are horizontally located rather than placed at a high altitude down.

shown in Fig. 5. We introduce the latest specific data-driven re-ID studies below.

1) Visible-thermal Re-ID: Typical re-ID focuses on the matching procedure of visible person images. However, RGB cameras cannot capture valid person appearance information under low or unavailable lighting scenarios, such as during night time. In comparison, infrared imaging is a widely-used module of most practical surveillance cameras at night/in the dark, such that cameras using RGB and infrared modules alternately can provide 24-hour person image capturing, which is crucial and meaningful for video investigation applications [4]. Therefore, thermal person images can be introduced for robust re-ID under poor lighting environments. Matching visible RGB person images for thermal ones at night or vice versa, can be called Visible-thermal re-ID. The Visible-thermal task is challenging due to the difficulty of cross-modality person matching. RGB images include three channels of visible descriptions (e.g., color cue), whereas thermal images contain only one channel invisible information. Moreover, Visible-thermal re-ID has to deal with the viewpoint change, low resolution, occlusion, etc., similar to traditional RGB-RGB re-ID.

The first thermal-related work is the local feature codebook based thermal re-ID that was proposed by *Kai et al.* [53] in 2010. It developed only thermal-thermal single-modality re-ID rather than the more important Visible-thermal cross-modality one. The importance of infrared cameras was highlighted by very few but representative studies, and the Visible-thermal re-ID issue remains open. *Wu et al.* [4] presented the Visible-thermal re-ID problem originally in 2017. They not only provided a meaningful cross-modality re-ID benchmark called SYSU Multiple Modality Re-ID (SYSUMM01), but also handled Visible-thermal person matching by using deep zero-padding, which was utilized to create a domain-specific one-stream network. Afterward, *Ye et al.* [54, 55] proposed both two-stage and dual-path end-to-end learning frameworks for Visible-thermal re-ID. Different from the former zero-padding method that focuses on identity information, this work utilizes either cross-modality shares or cross-modality and intra-modality variations, which provide guidance for cross-modality learning. However, so many losses of graphic details are in the process of thermal imagery that the re-ID performance is still low. Imaging mechanism studies might be helpful in developing Visible-thermal re-ID.

2) RGB-depth Re-ID: Similar to thermal images, depth images can preserve more invariance in an extremely low illumination and color conditions relative to RGB images. Moreover, depth images provide body shape and skeleton information, which is useful in situation where pedestrians change their clothing across different cameras. In 2012, *Barbosa et al.* [56] performed a re-ID study based on RGB-depth sensors for the first time. To address the clothing change problem for re-ID, they used a set of 3D soft biometrics cues instead of visual appearance for feature extraction, and provided a specific PAVIS dataset that recorded 79 persons with a frontal view and walking slowly in an indoor scenario. Afterward, *Mogelmose et al.* [57] proposed a joint classifier that combines RGB, depth, and thermal data. *Munaro et al.* [58] implemented the re-ID task by using 3D models that were reconstructed based on point cloud tracking for freely moving people and offered a biometric RGB-D dataset BIWI RGBD-ID that includes 50 subjects. *Haque et al.* [59] presented an attention-based re-ID model based on a combination of convolutional and recurrent neural networks that aimed to identify discriminative regions indicative of human identity. 4D spatio-temporal signatures of body shape and motion dynamics were extracted, especially in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 5: Illustration of re-ID using multi-modality and low-resolution person data. The multi-modality person data in sub-figures (a), (b), and (c), including thermal images, depth images, and text, are used to match corresponding RGB images for the re-ID task. Sub-figure (d) presents the objective of low-resolution re-ID, that is, low-resolution (LR) person images are used to match high-resolution (HR) ones or vice versa.

the absence of RGB information. *Wu et al.* [60] put forward a locally rotation invariant depth shape descriptor to depict a pedestrian body. Then kernelized implicit feature transfer was explored to combine the estimated depth features with RGB-based appearance features.

Generally, depth information can benefit re-ID considerably in a low-lighting, clothing-changed setting. However, although RGB-depth is helpful, it has not been sufficiently explored for the practical application of re-ID due to its two evident shortages. First, depth cameras (e.g., Kinect) are sometimes applied in indoor environments, but they rarely used in outdoor scenarios because depth information decreases rapidly with the increase in the distance between a pedestrian and the camera. More powerful depth cameras are crucial for RGB-depth re-ID. Second, the shape and skeleton information acquired from depth images can be indistinguishable, particularly when the viewpoint of a person image changes. Re-ID using depth only may be less helpful than re-ID using multi-modal matching.

3) Text-to-image Re-ID: In addition to person images captured from various cameras, the natural textual statements of eyewitnesses are regarded as the probe in video investigation applications. Thus, how to match corresponding gallery images of natural text is an interesting and meaningful task that is referred to Text-to-image re-ID [61]. An approximate direction to this task is attribute-based re-ID study, the idea of which is to distinguish different persons with discriminative mid-level attributes or high-level semantics rather than low-level features. Compared with generic attribute-based works, re-ID using semantic cues has two advantages. First, pedestrian attributes are related to person image patches (e.g., shirts are dressed on the upper body and pants are worn on the lower body). Second, the number of pedestrian attributes' categories is relatively limited compared with the numerous general attributes in classification tasks. Therefore, many effective attribute-based re-ID methods have been developed [62], and they could be found in other re-ID reviews [1, 29].

However, eyewitness statements are always natural sentence descriptions rather than discrete attributes. Besides, textual statements can serve as available probe, especially when suspect images are lacking. Therefore, cross-modality text-to-image is highly practicable for re-ID in realistic applications. In 2015, *Ye et al.* [63] addressed the specific person retrieval problem via incomplete text description, which can be regarded as the first Text-to-image re-ID work. Specifically, original fragmentary attributes were complemented via an online algorithm of linear sparse reconstruction. Then, attribute classifiers were learned based a restricted latent topic model in the offline process. *Li et al.* [64] addressed the textual-visual matching problem by investigating an identity-aware two-stage framework. Stage-one CNN-LSTM network with a cross-modal cross-entropy loss was learned to embed cross-modal features. Subsequently, stage-two CNN-LSTM network with a latent co-attention mechanism was constructed to refine the matching results. On this basis, *Li et al.* [61] further proposed a recurrent neural network with a gated neural attention mechanism and provided a large-scale person description

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

dataset with language annotations. Detailed information on person images were obtained from various sources.

Text-to-image re-ID receives little attention at present, two possible reasons can be summarized below. On the one hand, it is so challenging that the performance of state-of-the-art is still poor, as shown in Fig. 7 (c). On the other hand, most existing studies implement a textual-visual matching framework for person searching and object searching, and yet the boundary between Text-to-image re-ID and generic cross-modality textual-visual matching is not clear enough. Two significant insights can be found for Text-to-image re-ID in a viewpoint of video investigation application. First, eyewitness textual statements of suspects are always incomplete, suspects may not be noticed until crime happens. Second, these statements are sometimes incorrect. For example, eyewitnesses may mistake the color of a dress due to a specific illumination condition. Thus, although Text-to-image re-ID is less explored, it can be further developed due to its own uniqueness.

4) Low-resolution Re-ID: An urban video surveillance system is extremely costly that only a few cameras on main streets use a high-resolution (HR) configuration. As a result, low-resolution (LR) person images are more common than HR person images in practical applications. For the re-ID task, many datasets provide image samples with various resolutions (see Table II). However, most existing methods directly normalize input images to a uniform scale regardless of the influence of image resolution differences, thereby simply normalizing HR to LR and probably causing the loss of important visual information,; conversely, direct image scaling from LR to HR produces a large amount of noise [65].

To fill the gap between LR and HR person images for re-ID, the generic-purpose super-resolution (SR) technique [66, 67], as a simple pretreatment in re-ID, was introduced to obtain HR mappings of LR person images. However, versatile SR methods aim to enhance image quality instead of re-ID accuracy, even though visual improvement may cause performance degradation due to visual artifacts produced in the SR reconstruction process [68]. To solve these issues, several pioneer works have been investigated in multi-scale learning, and the SR technique and re-ID have been integrated into a joint learning framework.

In 2015, *Li et al.* [65] began the first low-resolution re-ID work, and a multi-scale discriminant distance metric learning model was proposed to optimize cross-scale image domain alignment by minimizing the heterogeneous class mean discrepancy. *Jing et al.* [68] learned a semi-coupled discriminant dictionary for mapping the features of HR gallery images and LR probe images, and low-rank regularization was adopted to learn dictionaries that can well depict the intrinsic feature space of HR and LR images. Different from these two studies that aimed to match single-scale LR person images in the HR gallery, *Wang et al.* [69, 70] implemented a scale-adaptive formulation in which HR probe images are matched against LR gallery images with different scales. A discriminating surface learning model and a cascaded super-resolution GAN framework were put forward successively for scale-adaptive cross-resolution re-ID. *Jiao et al.* [71] addressed low-resolution re-ID by jointly learning SR person images and person identity matching in a hybrid deep CNN.

Low-resolution problem not only exists in LR cameras, but also happens in HR cameras. For instance, if a pedestrian only appears in the distance from a camera, the scale of its imaging can be undersized. Moreover, it is to match single-modality RGB images with various scales, the data of which can be easily obtained. Therefore, Low-resolution re-ID gets more attention than other specific data-driven re-ID trends in recent years. As shown by experiments on these low-resolution re-ID methods, the performance improvements of re-ID after resolution enhancement and multi-resolution fusion exceed those of traditional re-ID methods, indicating that the long-neglected low-resolution problem is not only effective for re-ID accuracy but also substantial for the application of the re-ID technique.

B. Practical Procedure

The goal of practical video investigation application is to seek out a probe person in a gallery of full-frame images, so its procedure is usually two-stage, i.e., person detection and tracking for person image pre-processing and re-ID for person identity matching [72]. In general closed-world re-ID works, person detection and tracking are merely employed to provide desirable person data, and the hand-cropped strategy is usually applied in the data selection procedure. As a result, the person bounding boxes of most existing benchmarks are almost clean and less noisy, which is unavailable in practice. A few datasets have attempted to produce bounding boxes by automatic detectors with minimal manual intervention. For example, Deformable Part Model (DPM) detector was used to construct the Market-1501 dataset [50], and various area ratios of bounding boxes were produced automatically and labeled as "good," "distractor," or "junk." Pedestrian detectors unavoidably generate many "distractor" and "junk" bounding boxes, which can decrease the performance of re-ID significantly. Therefore, to bridge the gap between re-ID and practical procedure, the influence of person detectors (including detection and tracking) on re-ID accuracy must be discussed, and the end-to-end framework from detection to re-ID (see Fig. 6) may be a helpful and practicable open-world re-ID direction.

End-to-end Re-ID: The description of End-to-end re-ID has two versions at present. One is End-to-end re-ID from detection to identity matching, and the other is a combined learning architecture from feature extraction to distance measurement [73]. The latter focuses on jointly learning features and metrics, and it follows that the re-ID procedure begins with hand-cropped bounding boxes. This subsection emphasizes the discussion of the former one, which can also be treated as an End-to-end re-ID system from a practical viewpoint. In 2014, *Xu et al.* [74] posited that error detection is carried into re-ID inevitably. They presented a sliding window searching strategy to jointly model person detection and identity matching. This work was the first step of End-to-end re-ID research, and its experiments showed that the searching results of re-ID combining detection and matching could be better than that of two-stages separately. Afterward, *Xiao et al.* [75]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

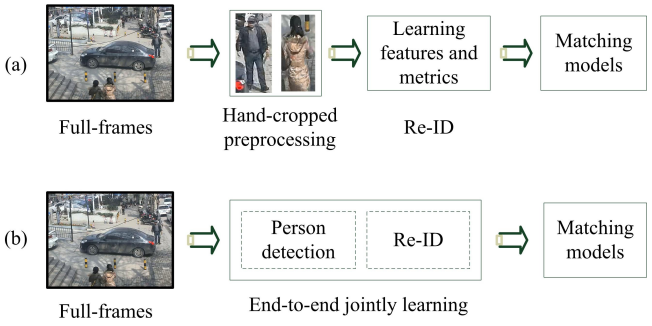

Fig. 6: Illustration of the common two-stage procedure and End-to-end re-ID. (a) Most of existing re-ID works directly employ hand-cropped person bounding boxes to learn robust features and metrics; therefore, person detection is merely the pre-processing stage of re-ID. (b) In the new End-to-end framework, person detection and re-ID are jointly learned to investigate the influence of various area ratios of bounding boxes.

proposed an End-to-end deep learning framework based on an online instance matching loss function to jointly localize and match probe persons from scene images. Furthermore, they constructed a Large-scale Person Search dataset (LSPS) that includes scene-diversified frames from surveillance cameras, hand-held cameras, and movies.

However, *Zheng et al.* [72] presented some concerns about the effectiveness of end-to-end learning. They provided a cascaded fine-tuning to train detection and classification metrics in turn, and devised sophisticated re-weighting schemes that were used to incorporate detection confidence into similarity scores. The experimental results showed that a dataset considering both detection and re-ID is necessary to promote the end-to-end framework; otherwise, the two-stage architecture may demonstrate better performance than end-to-end sometimes. A possible reason is that the intrinsic quality of detection is to distinguish person and non-person bounding boxes, whereas re-ID focuses on identity assignment of person bounding boxes. If there is no proper philosophy or data are adopted for unifying detection and re-ID, End-to-end re-ID still has a long way to go.

C. Efficiency

The MARS dataset contains thousands of identities and millions of images and has the largest number of person images for re-ID at present, but it is still far from a practical scale [1]. Specifically, the number of video surveillance cameras is usually more than 10,000 in a medium-sized city. Once the video investigation procedure is applied in a small city block with only 10 cameras, the time range of surveillance video becomes 24 hours, the frames per second (FPS) is always 25, and $10 \times 24 \times 3600 \times 25 = 2.16 \times 10^7$ full frames are obtained. When we try to automatically detect person bounding boxes, their number would be much more than 10^8 , which is hundreds of times more than the scale of MARS. In this case, what is the impact of data volume on re-ID study? Although no re-ID work has been conducted on such a large-scale dataset, *Zheng et al.* [50] performed a meaningful experiment that tested the same re-ID baseline on the Market-1501 dataset but with two data sizes. The experimental results showed that the evaluation metric mAP in Market-1501 with 500k decreased by 7% compared with the same dataset with 19k. In other words, the growth of data volume would make re-ID accuracy worsens.

Compared with the 25 times of data volume increase, the 7% accuracy decrease may be acceptable. However, if an approximate linear relationship exists between growth of data scale and consumption of running time, the efficiency of re-ID would not be easily approved for practice. For example, deeply-learned methods have become the dominant trend since CNN was introduced to re-ID task in 2014. Afterward, the best accuracies of re-ID benchmarks were nearly achieved by deep-based approaches. Nevertheless, the extraction and measurement stages of hierarchical CNN features comes at the expense of high computational cost; in particular, the computational complexity of the entire input image convolution has a linear relation to the number of pixels at the least [76], and person matching based on simple CNN features may be inapplicable to practical large-scale applications. Efficiency-driven re-ID study has rarely been explored, probably due to the lack of a large-scale dataset approaching realistic data volume so far. Although many re-ID studies have discussed their implementation time, these discussions are almost trivial compared with the large chunks of theoretical algorithms.

From the perspective of methodology, the efficiency issue focuses on the time consumption of re-ID algorithms. However, from the viewpoint of practical re-ID systems, person identity annotation for obtaining sufficient ground-truth data could be one of the most inefficient work. Re-ID labeling work aims to draw identity assignments manually when a person re-enters the Field of View (FOV). However, person appearance may exhibit much discrepancy across multiple cameras,thus making manual large-scale annotations for practical video surveillance applications painstaking and difficult. For the methodology efficiency issue, inverted index and hashing may be two potential directions according to the ideas of large-scale image retrieval [13], which is a similar community as re-ID. In particular, hash-based methods have been applied to re-ID research recently. To overcome the efficiency issue of large-scale identity annotation while balancing efficiency and accuracy, transfer learning/domain adaptation based on unlabeled data and labels from other domains is recommended as a momentous re-ID direction.

1) Hash-based Re-ID: The processing time of re-ID heavily relies on feature dimensions and scale of data; correspondingly, the aim of hashing is to cut the cost of searching proximate neighbors, especially when the size of the dataset is far more than large [77]. *Zhang et al.* [78] proposed the first hash-based re-ID framework that learned compact and bit-scalable hashing codes straightforwardly from raw images in 2015. This work built an end-to-end architecture from feature extraction to hashing function learning, the efficiency of which was distinctly dozens of times higher than baselines. *Zhu et al.* [79] presented a part-based deep hashing method that aimed to evaluate accuracy and efficiency by integrating deep learning and hashing into one framework. Batches of triplet samples were utilized as the input of the proposed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

deep hashing architecture, which was learned by a triplet loss function with Hamming intra-person and inter-person distance constraints. The experimental results showed that the accuracy and efficiency of re-ID in large-scale datasets can be balanced by devising sophisticated algorithms. *Wu et al.* [80] employed a deep hashing framework by learning a structured loss function that added positive pairs and hard negatives into mini-batches. Thus, CNNs and the hashing function were simultaneously learned to generate robust deeply-learned features and discriminative similarity-preserving hashing codes.

2) Transfer learning/domain adaptation for Re-ID: When developing a label-insufficient re-ID task, the crucial and perpetual unsupervised paradigm may be the first choice, and early re-ID works with hand-crafted features almost followed the unsupervised style. However, after benefiting from a large amount of labels, supervised re-ID has become a dominant re-ID trend because of its obviously superior performance compared with unsupervised methods. Although several new image-based [81] or video-based [82–84] unsupervised learning methods have emerged, their performance is still inferior to that of state-of-the-art supervised approaches. Different from generic unsupervised approaches, extensively studied transfer learning/domain adaptation not only utilizes unlabeled data of the target domain, but also exploits the characteristics of labeled data in the source domain. Consequently, the accuracy of transfer learning/domain adaptation-based re-ID (abbr. Transfer learning re-ID) is usually better than that of the unsupervised style meanwhile and approximates the best performance of the supervised style [85, 86].

Zheng et al. [36] initially introduced transfer learning to re-ID in 2012, and a set-based transfer learning framework based on bipartite ranking models was proposed for multi-shot verification and one-shot. Afterward, how to promote transfer learning/domain adaptation-driven re-ID has become a research increasing hotspot. *Ma et al.* [87] proposed an adaptive ranking support vector machine model that was used to refine the distance model learnt from the source domain with high confidence in target positive mean and low confidence in target negative image pairs. *Wang et al.* [88] jointly learned similarity measurements for different re-ID scenarios in an asymmetric manner, which was to obtain the shared component of cross-task data and enhance the target inter-class separation during joint learning. Different from previous weakly supervised transfer learning on which target datasets are not exactly unlabeled, *Peng et al.* [89] investigated cross-dataset adaptation based on multi-task dictionary learning with a completely unlabeled target dataset. *Geng et al.* [90] employed a deep transfer learning framework rather than previous hand-crafted features based frameworks, and a two-step fine-tuning strategy was presented to transfer knowledge from auxiliary datasets via co-training. Similarly in recent years, a growing number of re-ID work exhibited a strong interest in the unsupervised deep transfer learning/domain adaptation framework. *Li et al.* [91] proposed an adaptation and re-identification network that focused on leveraging across-dataset information and deriving domain-invariant features. *Wang et al.* [92] put forward a transferable attribute-identity deep learning framework for jointly learning attribute-semantic and identity discriminative feature representation space from a labeled source domain to an unlabeled target domain. *Lv et al.* [93] deployed a learning-to-rank method for incremental transfer learning based on pedestrian spatio-temporal patterns in the target domain. *Deng et al.* [94] introduced a preserving generative adversarial network that includes a Siamese network and a CycleGAN for domain adaptation re-ID. The learning procedure had two constraints, i.e., self-similarity of an image before and after translation and domain dissimilarity of a translated source image and a target image.

Transfer learning re-ID has been developing for a period of time, and supervision knowledge of the source domain can be readily introduced to the target domain in part, in terms of the experimental results of existing studies. However, the mechanism of current Transfer learning re-ID usually assume that the unlabeled target data is known and established, there is seldom studies to deal with a practical situation that the target data is sometimes uncertain. For example, surveillance videos captured from newly-added cameras will often be supplemented according to the dynamic investigation requirements. It is too time-consuming to learn a model after obtaining the target data, but performance of the transferred models can be poor when these models are learned only based on the labeled source data. Moreover, the amount of person data on re-ID datasets is limited compared with massive labeled data on person retrieval. Unsupervised transfer learning from generic person data to specific pedestrian data in the video surveillance environment is an important but still unexplored task.

IV. Datasets and Evaluations

Several surveys [30, 32] and websites[1] have provided holistic descriptions of re-ID datasets, but these introductions are inconvenient and insufficient. First, there lacks of guidance on the suitability of existing benchmarks for particular open-world re-ID applications is lacking. Second, specific open-world datasets are unavailable.This section addresses these limitations and provides instructions of corresponding datasets, evaluations, and re-ID accuracy over the years for each listed open-world trends.

A. For Open-set Re-ID

1) Datasets: In the open-set setting, re-ID is regarded as verification instead of identification which means many irrelevant persons exist, whereas no correct match is present in the gallery. To our knowledge, SAIVT-Softbio [37] is the only dataset designed for Open-set re-ID. This dataset contains 152 persons in eight camera views. Most pedestrians pass through a subset of cameras because it is an uncontrolled collection. A more general strategy is reconstructing numerous closed-set datasets [35, 38–40]. Specifically, person images from the probe and gallery are randomly divided into two parts, namely, one for training and another for testing, with the condition that only some of same-person identities exist in the probe and gallery. For example, VIPeR includes 632 person image

[1]http://robustsystems.coe.neu.edu/sites/robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

pairs obtained from two camera views, but 100 similar people may exist in both the probe and gallery for experiments on Open-set re-ID [39].

2) Evaluations: Liao et al. [38] proposed to calculate the CMC rate at a fixed False Accept Rate (FAR) that indicates the likelihood of misidentify the wrong identity. Their top-1 CMC rates at FAR = 1%, 10% were 3.99%, 14.51% on an unreleased re-ID dataset that contained 119 persons appearing in two to four camera views. *Wang et al.* [39] employed the FAR evaluation on two standard datasets VIPeR [3] and CUHK01[49]. Their top-1 CMC rates at FAR = 1%, 10% were 4.9% and 15.1% for VIPeR, and 7.5% and 20.2% for CUHK01. Neither of these two studies worked well, what's more, CMC is dependent of similar identity correspondence in closed-set scenario. Different from these works, *Zheng et al.* [35] and *Zhu et al.* [40] discarded the CMC rate completely and adopted the true target rate (TTR) and false target rate (FTR) for open-set evaluation.

Supposing that several non-target persons are placed in the probe population, the aim is not only to measure the performance on how well target probe persons are matched, but also how bad non-target ones pass through the verification process. When evaluating different approaches, their TTRs are compared against a series of given FTRs. Therefore, the re-ID performance of various open-set methods can be measured under different verification standards. TTR and FTR are expressed as the following equations [79].

$$TTR = \frac{N_{t2t}}{N_t}, FTR = \frac{N_{nt2t}}{N_{nt}}. \quad (1)$$

where the numbers of probe images from target and non-target persons are indicated by N_t and N_{nt}, respectively. N_{t2t} denotes the number of accurate verifications that target probe images are matched in the gallery. Similarly, N_{nt2t} denotes the number of false verifications that non-target probe images are treated as the target person. In Zheng's work, TTRs at FTR = 1% on VIPeR, iLIDS [52], ETHZ [95], and CAVIAR [7] were 23.47%, 32.03%, 76.29%, and 28.13% for one-shot individual verification, respectively. *Zhu* implemented their work on several larger datasets, and the TTRs at FTR = 1% on CUHK03 [17] and Market-1501 [50] were 49.96% and 66.52% for one-shot individual verification, respectively. By making a comprehensive analysis of these results, two major observations can be obtained.

First, a large performance gap exists between Open-set re-ID and traditional Closed-set re-ID. When using the typical top-1 CMC rate at FAR = 1% metric, Wang's results are merely 4.9% and 7.5% on VIPeR and CUHK01, respectively. By contrast, the best top-1 CMC rate on VIPeR is nearly 70% for Closed-set re-ID. Using the TTR rate at FTR = 1% metric, good reports are obtained, namely, 76.29%, 49.96%, and 66.52% on ETHZ, CUHK03, and Market-1501. In comparison, the performance of traditional Closed-set re-ID on ETHZ has reached its saturation, where the top-1 CMC rate is almost 100% [96]. On CUHK03 and Market-1501, the best closed-set performance has reached 94.4% and 97.8%, both of which are better than human performance [28]. Although TTR is different from CMC rate, as can be seen in Equation. 1, TTR indicates the probability of the correct target similar to CMC, which means they can be comparable with each other to some extent. Nevertheless, the performance of Open-set re-ID can still be improved.

Second, simple and easily-acquired experimental settings should be highlighted for developing Open-set re-ID. Although existing studies do not quite agree with each other's settings, we can provide guidance to uniform these settings according to the characteristic of Open-set re-ID. For the data preparation, many irrelevant probe persons may exist while no correct gallery match is present in the open-set scenarios, so only a few of pedestrians can exist in both the probe and gallery for training and test. For the evaluation, TTRs at certain FTRs are recommended rather than traditional CMC rates. TTRs can measure the performance on verifying the target and non-target persons, and it is independent of the one-to-one identity correspondence, which is a closed-set hypothesis that can employ CMC better.

B. For Specific Data-driven Re-ID

1) Datasets: Benchmarks for Visible-thermal re-ID. Two specialized Visible-thermal datasets, i.e., RegDB and SYSU-MM01, display the obvious distinction between visible RGB images and thermal ones. The RegDB dataset [97] contains 412 persons captured by dual camera systems. For each person, 10 visible images are collected by a RGB camera, and 10 thermal images are acquired by an infrared camera. SYSU-MM01 [4] includes 491 persons captured from six cameras, i.e., four RGB and two infrared cameras. Each person appears in two different cameras at the least. Notably, person images collected in indoor and outdoor scenarios make Visible-thermal re-ID highly challenging.

Benchmarks for RGB-depth re-ID. Four publicly available person depth datasets, i.e., PAVIS [56], BIWI RGBD-ID [58], IAS-Lab RGBD-ID [98], and DPI-T [59], are commonly used for this task. PAVIS contains 79 persons with four states of motion, namely, collaborative, walking, walking2, and backwards. BIWI RGBD-ID includes 50 training and 56 testing sequences of 50 different people. Synchronized RGB images, depth images, segmentation maps, and skeletal data are all provided with a Microsoft Kinect. Compared with the 10 FPS video on the BIWI RGBD-ID dataset, IAS-Lab RGBD-ID provides 30 FPS video that involves 11 training and 22 testing sequences of 11 different people. DPI-T is a newly depth-based re-ID benchmark that employs top-down cameras, and it comprises 24 individuals in 25 videos across several days. However in general, the data volume of existing person depth datasets is too small to be practicable.

Benchmarks for Text-to-image re-ID. No unified standard dataset consists of person images with natural language; therefore, existing re-ID datasets are often labeled by crowd workers for Text-to-image re-ID. CUHK-PEDES [64] contains 13,003 person identities with 40,206 images from CUHK03 [17], Market-1501 [50], VIPER [3], CUHK01 [49], etc. Each person image is described by two sentences via Amazon Mechanical Turk (AMT), and a total of 80,412 sentences are collected.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

Benchmarks for Low-resolution re-ID. Several low-resolution studies produced multi-resolution person images by down-sampling [49,52-54] from several public datasets, such as VIPeR, CUHK03, PRID450s [99], and 3DPES [100]. Take the SALR-VIPeR dataset [69] as an example. The widely-used VIPeR dataset contains 1264 outdoor images captured from two views of 632 persons. In the experimental setting of low-resolution re-ID, person images from view A are set as the high-resolution probe, and images from view B are treated as the low-resolution gallery whose resolutions are down-sampled randomly to different scales. However, the mechanism of resolution degradation is uncertain in realistic video surveillance environments, and re-ID based on such simulative person data may be not robust enough for practical applications. Fortunately, several re-ID datasets have weighted the impact of practical resolution changes and provided various person images with various resolutions. Typically, the CAVIAR4ReID dataset [7] includes 72 person identities from two camera views in a shopping mall, and the image resolution of the second camera is much lower than that of the first one, which is extraordinarily fitted for cross-resolution re-ID. Market-1501 and PRW [72] simulate the video surveillance scenario, where person images are collected from six horizontal cameras in front of a supermarket. Five of the six cameras are high-resolution ones, and the other is a low-resolution camera.

Benchmarks for End-to-End re-ID. The core data requirement of End-to-end re-ID from detection to person matching is raw surveillance video; therefore, if a dataset possesses full-frame images, it can be used for End-to-end re-ID [74]. Many datasets have provided full-frame images, such as PRID2011, PRW, and DukeMTMC4ReID. More details can be found in a previous re-ID dataset review [32].

Benchmarks for Hash-based re-ID. Many re-ID datasets claim that the their data volume is large. For example, CUHK03, Market-1501, LSPS [75], MSMT17 [101], and airport [32] have thousands of person identities. However, relative to practical data volume (refer to Section III.C), the size of existing datasets is actually small [102]. Therefore, existing hash-based methods have to be tested on relatively large datasets. To acquire useful data, two latent strategies can be adopted. First, combinations of person data from heterologous datasets are usually utilized for transfer learning re-ID. Second, GANs are designed to generate sophisticated samples [39,54,72], which can be used to expand the training set with labeled and unlabeled data; then, the data volume of person images can be developed rapidly and annotated automatically.

Benchmarks for Transfer learning/domain adaptation re-ID. Transfer learning and domain adaptation generally convert supervision knowledge from the source domain to the target domain. More than 30 publicly standard re-ID datasets are available, so the regular idea of data preparation is splitting several of these datasets into two parts. The one with labeled data is regarded as the source domain, and the other without identity annotations is treated as the target domain. Market-1501 and DukeMTMC-reID [49] are frequently treated as the source-target in recent years.

2) Evaluations: Cumulative matching characteristics (CMC) curve is the most well-known metric for the evaluation of specific application-driven re-ID algorithms. It is a plot of identification performance versus the ranking score that represents the probability of finding the correct probe person inside top k matches of the gallery. In other words, CMC depicts the ranking score of the best match. Although specific application-driven re-ID methods belong to generalized open-world orientations, their setting of identity correspondence across camera views is similar to that of closed-set works. Thus, most specific application-driven studies still employ CMC curve as the chief evaluation metric. For the CMC evaluation, the first match is calculated for counting the ranking score regardless of how many ground truths match in the gallery. Therefore, CMC is a unilateral metric for evaluating a re-ID method when more than one ground truth exists in commonly practical gallery. In the case of multiple ground truths, the mean average precision (mAP) [50], which is a conventional metric in image retrieval community, could be used to comprehensively appraise the performance of multi-matching. mAP represents the retrieval recall capability of re-ID methods, and it provides an overall metric by measuring the quality of all rank lists.

3) Re-ID Accuracy Over the Years: In this section, re-ID accuracies for each listed specific application-driven trends are summarized over the years. Visible-thermal re-ID, RGB-depth re-ID, Text-to-image re-ID, and Low-resolution re-ID usually employ image-based ground truths. The typical top-1 CMC rates for each trend are reported over the years (see Fig. 7). For procedure-driven and efficiency-driven works, such as End-to-end re-ID, Hash-based re-ID, and Transfer learning re-ID, some of their datasets are in a situation of multi-shot person images or video frames. Hence, mAP is also adopted for comparisons in addition to the top-1 CMC rate (see Fig. 8). We list the representative re-ID performance with the commonly-used implementation settings here. By summarizing the two figures above, three major insights can be obtained .

First, the performance of all re-ID trends is improving rapidly. For Visible-thermal re-ID on the RegDB dataset [97], RGB-depth re-ID on the PAVIS dataset [56], Text-to-image re-ID on the CUHK-PEDES dataset [64], Low-resolution re-ID on the SALR-CUHK03 dataset [71], Hash-based re-ID on the CUHK03 dataset [17], and Transfer learning re-ID on the DukeMTMC-reID [49] and Market-1501 datasets [50], the top-1 CMC rates have a salient increase of 15.7%, 42%, 17.9%, 45.5%, 52.2%, and 27.9%, respectively. Typically, on the classic dataset PAVIS, several representative RGB-depth works [56–60] from 2012 to 2017 have achieved an accuracy increase from 15% to 57% and an improvement of +42%. On the newly released DukeMTMC-reID and Market-1501 datasets, the state-of-the-art results of Transfer learning re-ID [89, 91, 92, 94] have been promoted from 18.5% to 46.4% since 2016, which is equivalent to an improvement of +27.9%.

Second, performance can still be improved significantly. Several reports have shown that the best score of Low-resolution re-ID on the CUHK03 dataset is 67.7%, and Transfer learning re-ID on the Market-1501 is 46.4%. However, the top-1 accuracies of traditional re-ID works on Market-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

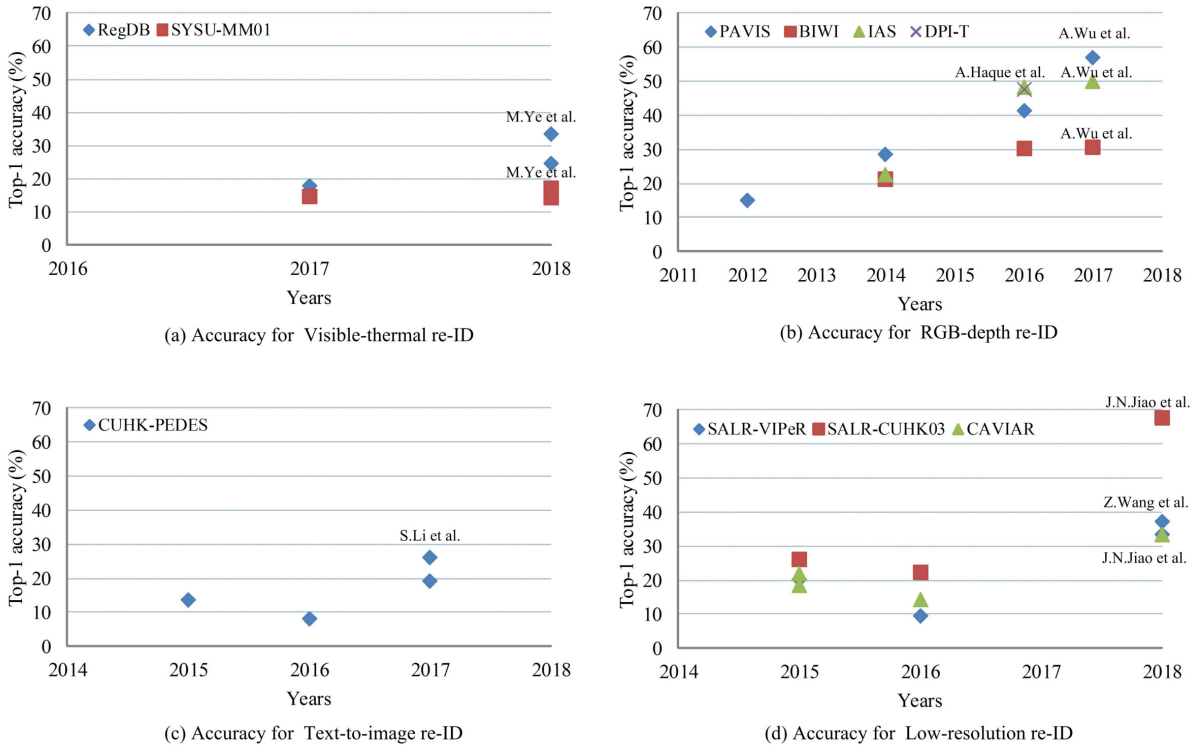

Fig. 7: Re-ID accuracy for specific data-driven trends. (a) Visible-thermal re-ID, (b) RGB-depth re-ID, (c) Text-to-image re-ID, (d) Low-resolution re-ID. Top-1 CMC rates on representative datasets are plotted according to their publication (or ArXiv) time, and the best performance is highlighted on the corresponding dataset for each re-ID trend.

1501 and on CUHK03 datasets have marvelously increased to 94.4% and 97.8% [28], both of which exceed the 93.5% and 95.7% accuracies of human beings. Furthermore, the performance of transfer learning re-ID with multi-shot ground truths has surpassed 70% on the DukeMTMC-reID dataset, but its corresponding mAP is less than 40%, as shown in Fig. 8(d). This implies that most ground truths are still ranked at the bottom of the re-ID ranking list, although the best match is probably correct.

Third, available standard datasets need to be further explored for most specific application-driven re-ID trends. Compared with traditional Closed-set re-ID works that have dozens of various baselines on more than 30 datasets [1], the majority of specific application-driven re-ID trends are still in an embryonic stage. Specifically, the data volume of the largest re-ID dataset MARS [51] is more than one million, but it is far less than the real data size in practical applications for verifying the efficiency of hash-based re-ID. End-to-end re-ID has no specialized benchmarks, and only one or two datasets are available for Visible-thermal re-ID and Text-to-image re-ID. RGB-depth re-ID has four specialized datasets, i.e., PAVIS, BIWI RGBD-ID, IAS-Lab RGBD-ID, and DPI-T, but the number of pedestrians is less than one hundred on all of these datasets. Fortunately, Low-resolution re-ID and Transfer learning re-ID benefit from the accessibility of data that can be obtained by reconstructing numerous traditional re-ID benchmarks, and they both developed rapidly in recent years.

V. Future Open-world: Long-term Re-ID

Re-ID has just entered its second decade, and discussions of its future issues always arise. Representative surveys [1, 8, 29–32] have summarized various challenges and perspectives, most of which have been promoted in some degree, such as various hand-crafted and deeply-learned methods for overcoming inter- and intra-class variations, unsupervised learning and transfer learning approaches for coping with data labelling requirement, and other application-driven works mentioned in this survey. However, most existing trends generally belong to short-term re-ID in which pedestrians always move in a small space for a short period. Take the representative re-ID benchmarks in Table. II for example. The number of camera views is only two for 60% of the benchmarks, and the maximum is eight. By contrast, practical applications usually employ a long-term scenario in which people probably go through numerous camera views (or other photographing equipment) in a large video surveillance network for a long period. In the case of a long-term scenario, three questions are raised for re-ID.

Generalization capability in large-scale camera views. One-to-one corresponding identities always exist in pair cameras on early re-ID datasets, such as VIPeR, PRID, iLIDS, and CUHK01. An intuitional re-ID idea is to train robust models for a specific pair of cameras. In the condition of multiple cameras, re-ID models learnt from one camera pair may be not sufficiently generic for others, and studying models for every pair of cameras is infeasible because it needs N×(N-1) models when the number of cameras is N. Two choices

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2898940, IEEE Transactions on Circuits and Systems for Video Technology

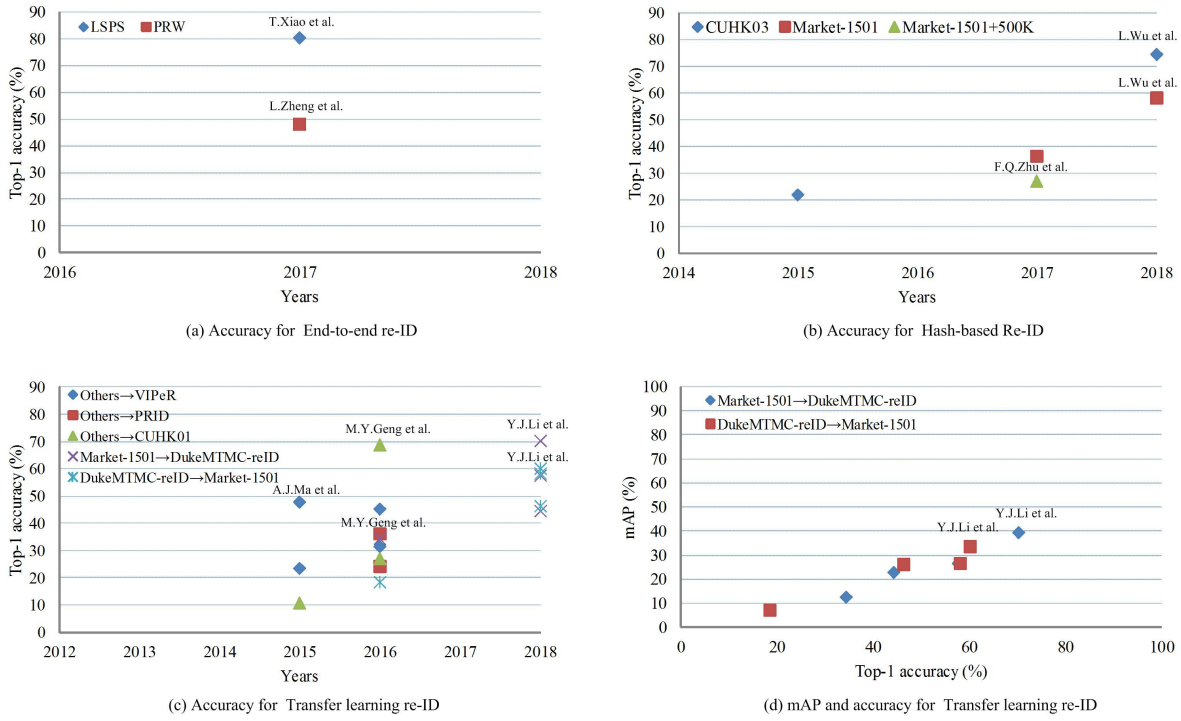

Fig. 8: Re-ID accuracy for procedure-driven and efficiency-driven trends. (a) End-to-end re-ID, (b) Hash-based re-ID, (c) Transfer learning re-ID, and mAP for (d) Transfer learning re-ID. All of these re-ID results on representative datasets are plotted according to their publication (or ArXiv) time, and the best performance is highlighted on the corresponding dataset for each re-ID trend. Notably, in subfigure (b), all of the results are tested with 128-bit hashing codes. In subfigure (c), '→' presents the source-to-target relationship, and 'others' means that the source datasets include several or all of VIPeR, PRID, CUHK01, CAVIAR, and i-LIDS benchmarks.

can be adopted. First option is to treat person images from multiple cameras as a whole, which is a prevailing practice nowadays. This practice is suitable for datasets that have no obvious differences in camera settings, such as CUHK03 and Market-1501, but it may be unsuitable for scenarios with evident inter-camera differences, such as the PRID dataset that has serious chromatic aberration between cameras. The second option is to seek for a model that can apply different camera views to a unified feature space [103]. However, covering diversified cross-view variations is difficult when only a single model is used. For example, re-ID in daylight is probably different from that at night. Other meaningful efforts have been exerted to study a dataset-guided model [23], which jointly trains domain-shared and domain-specific features. The underlying assumption of the domain-specific feature is that intra-domain can be regarded as an entity, but it should be noted that intra-domain differences may be larger than inter-domain differences in practical applications. Moreover, researchers have attempted to generalize their re-ID models, but the generalization capability in large-scale camera views remains an open issue.

Scalability in a dynamic camera network. The influence of large-scale cameras on re-ID is not only statically quantitative but also dynamically spatio-temporal [104]. Although the locations of cameras are normally fixed, the searching area and time slot can be altered or expanded according to certain clues in practical applications. Training all prior data or upcoming ones aforehand is difficult, and the key question of re-ID becomes "how could we incorporate new data (e.g., person images captured from a new camera) into trained models better?" On the one hand, the probe may pass the new camera view or not; hence, it is a open-set problem. On the other hand, retraining all the data after adding new cameras is unreasonable, so test-time scalability is a significant issue. *Panda et al.* [105] employed an unsupervised geodesic flow kernel to determine the best-matched source camera of the newly introduced target camera then exploited the source-target correlation for other cameras pairs. Although this unsupervised style may avoid time-consuming data relabeling, performance and the practical on-the-fly real-time requirement can still be improved.

Clothing change in a long period. Appearance invariance is a basic hypothesis of the generic re-ID task, and existing studies focused on obtaining improved appearance descriptions or matching the person figure with many discriminative metrics while the availability of appearance is overblown currently. Ordinarily, pedestrians do not change their clothing within several minutes when they pass through a few cameras. However, crime suspects probably change their clothes in a long period to evade the video surveillance. In this case, the core point of re-ID becomes "what features would still operate if appearance is out of work?" Three potential options are available. First, although spatio-temporal clues may be not robust in non-overlapping camera views, they can still be useful supplementary information especially for large-scale trajectory inference. However, existing spatio-temporal re-ID

studies have paid attention to time-sequential appearance for video-based re-ID [73, 106, 107]. Studies that used the spatio-temporal correlation of pedestrians in a large-scale camera network are rare. Second, the 3D body model presents certain invariance even when clothing is changed; it is also available for object classification and recognition [108]. Effective 3D body models are built on powerful depth cameras, which may exist in several indoor environments but rarely appear in practical outdoor video surveillance systems. Pose-driven re-ID is a good attempt to utilize the information on body structure [109, 110], but this technology merely focuses on pose-normalized person matching. Third, classic biometrics, such as face and gait, are feebly affected by clothing changes. Meanwhile, biometrics are usually unavailable or unobtainable due to low-resolution, viewpoint change, etc. Therefore, current biometrics-based re-ID is almost always performed in an indoor simulation environment [111]. The use and popularization of high-resolution and intelligent cameras are good news for biometric-based re-ID, and several high-resolution re-ID datasets with 4K (3840×2160 resolution) cameras are being developed. These would promote appearance- and biometric-based re-ID in the near future.

VI. Conclusions

This survey presents the first overall review of open-world person re-identification and categorizes existing works from narrow and generalized perspectives. In this regard, we first introduce the development of Open-set re-ID, which is treated as the narrow definition of open-world re-ID, and analyze the core differences between closed- and open-set scenarios. Second, specific application-driven re-ID, which is regarded as generalized open-world re-ID, is summarized from three aspects of application requirements, namely, raw data, practical procedure, and efficiency. Third, suitable datasets for each open-world directions are described, and the availabilities of evaluations and re-ID accuracies over the years are summarized and discussed. Finally, long-term re-ID, an important future open-world topic that is completely different from previous short-term re-ID, is discussed in three aspects: generalization capability, scalability, and clothing change. This survey highlights open-world re-ID, which is a crucial but long-ignored issue that is and helpful for developing less explored re-ID orientations toward practical applications.

References

[1] Liang Zheng, Yi Yang, and Alexander G Hauptmann, "Person re-identification: Past, present and future," *Arxiv*, 2016.

[2] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Change Loy Chen, "Person re-identification," *Advances in Computer Vision and Pattern Recognition*, vol. 42, no. 7, pp. 301–313, 2013.

[3] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision - ECCV 2008, European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings*, 2008, pp. 262–275.

[4] Ancong Wu, Wei Shi Zheng, Hong Xing Yu, Shaogang Gong, and Jianhuang Lai, "Rgb-infrared cross-modality person re-identification," in *IEEE International Conference on Computer Vision*, 2017, pp. 5390–5399.

[5] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian Conference on Image Analysis*, 2011, pp. 91–102.

[6] Change Loy Chen, Tao Xiang, and Shaogang Gong, "Multi-camera activity correlation analysis," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1988–1995.

[7] Seon Cheng Dong, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference*, 2011, pp. 68.1–68.11.

[8] Xiaogang Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013.

[9] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 26–33 vol. 2.

[10] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley, "Person reidentification using spatiotemporal appearance," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 1528–1535.

[11] Loris Bazzani, Marco Cristani, Alessandro Perina, Michela Farenzena, and Vittorio Murino, "Multiple-shot person re-identification by hpe signature," in *International Conference on Pattern Recognition*, 2010, pp. 1413–1416.

[12] 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, pp. 2360–2367.

[13] Liang Zheng, Yi Yang, and Qi Tian, "Sift meets cnn: A decade survey of instance retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1224–1244, 2018.

[14] Qingming Leng, Ruimin Hu, Chao Liang, Yimin Wang, and Jun Chen, "Person re-identification with content and context re-ranking," *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 6989–7014, 2015.

[15] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2553–2566, 2016.

[16] Wei Shi Zheng, Shaogang Gong, and Tao Xiang, "Person re-identification by probabilistic relative distance comparison," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 649–656.

[17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[18] Mert Dikmen, Emre Akbas, Thomas S. Huang, and Narendra Ahuja, "Pedestrian recognition with a learned metric.," *Lecture Notes in Computer Science*, vol. 6495, pp. 501–512, 2010.

[19] Yimin Wang, Ruimin Hu, Chao Liang, Chunjie Zhang, and Qingming Leng, "Camera compensation using feature projection matrix for person re-identification," in *IEEE International Conference on Multimedia and Expo*, 2013, pp. 1–6.

[20] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[21] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Deep metric learning for person re-identification," in *International Conference on Pattern Recognition*, 2014, pp. 34–39.

[22] Ejaz Ahmed, Michael Jones, and Tim K. Marks, "An improved deep learning architecture for person re-identification," in

IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.

[23] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1249–1258.

[24] Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, and Nanning Zheng, "Point to set similarity based deep feature learning for person reidentification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 6.

[25] Sanping Zhou, Jinjun Wang, Deyu Meng, Xiaomeng Xin, Yubing Li, Yihong Gong, and Nanning Zheng, "Deep self-paced learning for person re-identification," *Pattern Recognition*, vol. 76, pp. 739–751, 2018.

[26] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *European Conference on Computer Vision*, 2016, pp. 732–748.

[27] Shaogang Gong, Change Loy Chen, and Tao Xiang, *Security and Surveillance*, Springer London, 2011.

[28] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *Arxiv*, 2017.

[29] Apurva Bedagkar-Gala and Shishir K Shah, "Editor's choice article: A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[30] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara, "People reidentification in surveillance and forensics:a survey," *Acm Computing Surveys*, vol. 46, no. 2, pp. 1–37, 2013.

[31] Riccardo Satta, "Appearance descriptors for person re-identification: a comprehensive review," *Eprint Arxiv*, 2013.

[32] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2016.

[33] Xiang Li, Ancong Wu, and Wei-Shi Zheng, "Adversarial open-world person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 280–296.

[34] Xiaogang Wang, G Doretto, T Sebastian, and J Rittscher, "Shape and appearance context modeling," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[35] W. S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 591, 2016.

[36] Wei Shi Zheng, "Transfer re-identification: From person to set-based verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2650–2657.

[37] Brais Cancela, Timothy M Hospedales, and Shaogang Gong, "Open-world person re-identification by multi-label assignment inference," in *British Machine Vision Conference*, 2014.

[38] Shengcai Liao, Zhipeng Mo, Jianqing Zhu, Yang Hu, and Stan Z. Li, "Open-set person re-identification," *Computer Science*, 2014.

[39] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong, "Towards unsupervised open-set person re-identification," in *IEEE International Conference on Image Processing*, 2016, pp. 769–773.

[40] X. Zhu, B. Wu, D. Huang, and W. S. Zheng, "Fast open-world person re-identification.," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. PP, no. 99, pp. 1–1, 2017.

[41] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 267–282, 2008.

[42] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang, "An equalized global graph model-based approach for multi-camera object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 11, pp. 2367–2381, 2017.

[43] Andrii Maksai, Xinchao Wang, François Fleuret, and Pascal Fua, "Non-markovian globally consistent multi-object tracking," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2563–2573.

[44] Timothy Huang and Stuart Russell, "Object identification in a bayesian context," in *IJCAI*, 1997, vol. 97, pp. 1276–1282.

[45] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 146–162, 2008.

[46] Kuan-Wen Chen, Chih-Chuan Lai, Pei-Jyun Lee, Chu-Song Chen, and Yi-Ping Hung, "Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 625–638, 2011.

[47] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah, "Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets," *arXiv preprint arXiv:1706.06196*, 2017.

[48] Ergys Ristani and Carlo Tomasi, "Features for multi-target multi-camera tracking and re-identification," *arXiv preprint arXiv:1803.10859*, 2018.

[49] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*, 2016, pp. 17–35.

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2016, pp. 1116–1124.

[51] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*, 2016, pp. 868–884.

[52] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, 2014, pp. 688–703.

[53] Jungling Kai and Michael Arens, "Local feature based person reidentification in infrared image sequences," in *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 448–455.

[54] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *the Association for the Advance of Artificial Intelligence*, 2018.

[55] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *International Joint Conference on Artificial Intelligence*, 2018.

[56] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino, "Re-identification with rgb-d sensors," in *International Conference on Computer Vision*, 2012, pp. 433–442.

[57] Andreas Mogelmose, Chris Bahnsen, Thomas Moeslund, Albert Clapes, and Sergio Escalera, "Tri-modal person re-identification with rgb, depth and thermal features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 301–307.

[58] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool,

and Emanuele Menegatti, "3d reconstruction of freely moving persons for re-identification with a depth sensor," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 4512 – 4519.

[59] Albert Haque, Alexandre Alahi, and Fei Fei Li, "Recurrent attention models for depth-based person identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1229–1238.

[60] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai, "Robust depth-based person re-identification.," *IEEE Trans. Image Processing*, vol. 26, no. 6, pp. 2588–2603, 2017.

[61] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, "Person search with natural language description," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5187–5196.

[62] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai, "Adversarial attribute-image person re-identification.," in *IJCAI*, 2018, pp. 1100–1106.

[63] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu, "Specific person retrieval via incomplete text description," in *ACM on International Conference on Multimedia Retrieval*, 2015, pp. 547–550.

[64] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang, "Identity-aware textual-visual matching with latent co-attention," in *IEEE International Conference on Computer Vision*, 2017, pp. 1908–1917.

[65] Xiang Li, Wei Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong, "Multi-scale learning for low-resolution person re-identification," in *IEEE International Conference on Computer Vision*, 2015, pp. 3765–3773.

[66] Junjun Jiang, Xiang Ma, Chen Chen, Tao Lu, Zhongyuan Wang, and Jiayi Ma, "Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 15–26, 2016.

[67] Junjun Jiang, Chen Chen, Jiayi Ma, Zheng Wang, Zhongyuan Wang, and Ruimin Hu, "Srlsp: A face image super-resolution algorithm using smooth regression with local structure prior," *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 27–40, 2017.

[68] X. Y. Jing, X. Zhu, F. Wu, R. Hu, X. You, Y. Wang, H. Feng, and J. Y. Yang, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1363–1378, 2017.

[69] Zheng Wang, Ruimin Hu, Junjun Jiang, Junjun Jiang, Chao Liang, and Jinqiao Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 2669–2675.

[70] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shinichi Satoh, "Cascaded sr-gan for scale-adaptive low resolution person re-identification," in *International Joint Conference on Artificial Intelligence*, 2018.

[71] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang. Gong, "Deep low-resolution person reidentification," in *the Association for the Advance of Artificial Intelligence*, 2018.

[72] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian, "Person re-identification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[73] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person reidentification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6776–6785.

[74] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *ACM International Conference on Multimedia*, 2014, pp. 937–940.

[75] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang, "Joint detection and identification feature learning for person search," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[76] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3270–3278.

[77] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1556–1564.

[78] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification.," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4766–4779, 2015.

[79] Fuqing Zhu, Xiangwei Kong, Zheng Liang, Haiyan Fu, and Qi Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4806–4817, 2017.

[80] Lin Wu, Yang Wang, Zongyuan Ge, Qichang Hu, and Xue Li, "Structured deep hashing with convolutional neural networks for fast person re-identification," *Computer Vision and Image Understanding*, vol. 167, pp. 63–73, 2018.

[81] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593.

[82] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *IEEE International Conference on Computer Vision*, 2017.

[83] Mang Ye, Xiangyuan Lan, and Pong C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *ECCV*, 2018.

[84] Mang Ye, Jiawei Li, Andy J Ma, Liang Zheng, and Pong C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Transactions on Image Processing (TIP)*, 2019.

[85] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *IEEE International Conference on Computer Vision*, 2017.

[86] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[87] A. Ma, J. Li, P. C. Yuen, and P. Li, "Cross-domain person re-identification using domain adaptation ranking svms," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1599–1613, 2015.

[88] Xiaojuan Wang, Wei Shi Zheng, Xiang Li, and Jianguo Zhang, "Cross-scenario transfer person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 8, pp. 1447–1460, 2016.

[89] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1306–1315.

[90] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian, "Deep transfer learning for person re-identification," *Arxiv*, 2016.

[91] Yu Jhe Li, Fu En Yang, Yen Cheng Liu, Yu Ying Yeh, Xiaofei Du, and Yu Chiang Frank Wang, "Adaptation and re-

identification network: An unsupervised deep transfer learning approach to person re-identification," *Arxiv*, 2018.

[92] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," *arXiv preprint arXiv:1803.09786*, 2018.

[93] Jianming Lv, Weihang Chen, Qing Li, and Can Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," *Arxiv*, 2018.

[94] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," *Arxiv*, 2017.

[95] William Robson Schwartz and Larry S Davis, "Learning discriminative appearance-based models using partial least squares," in *Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on*. IEEE, 2009, pp. 322–329.

[96] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury, "Re-identification in the function space of feature warps," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1656–1669, 2015.

[97] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras.," *Sensors*, vol. 17, no. 3, pp. 605, 2017.

[98] M Munaro, S Ghidoni, D. T Dizmen, and E Menegatti, "A feature-based approach to people re-identification using skeleton keypoints," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 5644–5651.

[99] Peter M. Roth, Martin Hirzer, Martin K?stinger, Csaba Beleznai, and Horst Bischof, *Mahalanobis Distance Learning for Person Re-identification*, Springer London, 2014.

[100] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Joint ACM Workshop on Human Gesture and Behavior Understanding*, 2011, pp. 59–64.

[101] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, "Person transfer gan to bridge domain gap for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[102] Jia Deng, Wei Dong, R. Socher, Li Jia Li, Kai Li, and Fei Fei Li, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248–255.

[103] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617.

[104] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 392–408, 2018.

[105] Rameswar Panda, Amran Bhuiyan, Vittorio Murino, and Amit K. Roy-Chowdhury, "Unsupervised adaptive re-identification in open world dynamic camera networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1377–1386.

[106] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3810–3818.

[107] Apurva Bedagkar-Gala and Shishir K Shah, "Part-based spatio-temporal model for multi-person re-identification," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1908–1915, 2012.

[108] Jake K Aggarwal and Lu Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[109] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue, "Pose-normalized image generation for person re-identification," *arXiv preprint arXiv:1712.02225*, 2017.

[110] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian, "Pose-driven deep convolutional model for person re-identification," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3980–3989.

[111] Aditi Roy, Shamik Sural, and Jayanta Mukherjee, "A hierarchical method combining gait and phase of motion with spatiotemporal model for person re-identification," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1891–1901, 2012.

Qingming Leng received the B.S degree in life science from Nanchang University, Nanchang, China, in 2007, M.S degree in International School of Software from Wuhan University, Wuhan, China, in 2009, and the Ph.D degree in National Engineering Research Center for Multimedia Software from Wuhan University, Wuhan, China, in 2014. He is currently working as a lecturer at School of Information Science and Technology, Jiujiang University, China. His research interests include person re-identification, image retrieval and machine learning.

Mang Ye received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2013 and in 2016. He is currently a Ph.D student at Department of Computer Science, Hong Kong Baptist University. His research interests focus on multimedia content analysis and retrieval, computer vision and pattern recognition.

Qi Tian (M'96-SM'03-F'16) received the B.E. degree in electronic engineering from Tsinghua University, China, the M.S. degree in electrical and computer engineering from Drexel University, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, in 1992, 1996, and 2002, respectively. He is currently a Professor with the Department of Computer Science, University of Texas at San Antonio (UTSA). He took a one-year faculty leave at Microsoft Research Asia from 2008 to 2009. His research interests include multimedia information retrieval and computer vision. He has authored more than 230 refereed journal and conference papers. His research projects were funded by NSF, ARO, DHS, SALSI, CIAS, and UTSA, and he also received faculty research awards from Google, NEC Laboratories of America, FXPAL, Akiira Media Systems, and HP Laboratories. He received the Best Paper Awards in PCM 2013, MMM 2013, and ICIMCS 2012, the Top 10% Paper Award in MMSP 2011, the Best Student Paper in ICASSP 2006, and the Best Paper Candidate in PCM 2007. He received 2010 ACM Service Award. He is the Guest Editor of the IEEE Transactions on Multi-media, Journal of Computer Vision and Image Understanding, Pattern Recognition Letter, EURASIP Journal on Advances in Signal Processing, Journal of Visual Communication and Image Representation, and is on the Editorial Board of the IEEE Transactions on Multimedia, IEEE Transactions on Circuit and Systems for Video Technology, Multimedia Systems Journal, Journal of Multimedia, and Journal of Machine Visions and Applications.