Person Re-identification: Past, Present and Future

Liang Zheng, Yi Yang, and Alexander G. Hauptmann

Abstract—Person re-identification (re-ID) has become increasingly popular in the community due to its application and research significance. It aims at spotting a person of interest in other cameras. In the early days, hand-crafted algorithms and small-scale evaluation were predominantly reported. Recent years have witnessed the emergence of large-scale datasets and deep learning systems which make use of large data volumes. Considering different tasks, we classify most current re-ID methods into two classes, *i.e.,* image-based and video-based; in both tasks, hand-crafted and deep learning systems will be reviewed. Moreover, two new re-ID tasks which are much closer to real-world applications are described and discussed, *i.e.,* end-to-end re-ID and fast re-ID in very large galleries. This paper: 1) introduces the history of person re-ID and its relationship with image classification and instance retrieval; 2) surveys a broad selection of the hand-crafted systems and the large-scale methods in both image- and video-based re-ID; 3) describes critical future directions in end-to-end re-ID and fast retrieval in large galleries; and 4) finally briefs some important yet under-developed issues.

Index Terms—Large-scale person re-identification, hand-crafted systems, Convolutional Neural Network, literature survey.

✦

1 INTRODUCTION

ACCORDING to Homer (*Odyssey* iv:412), *Mennelaus* was becalmed on his journey home from the Trojan War; He wanted to propitiate the gods and return safely home. He was told that he should capture *Proteus* and force him to reveal the answer. Although *Proteus* transformed to a lion, a serpent, a leopard, water and also a tree, *Mennelaus* then succeeded in holding him as he emerged from the sea to sleep among the seals. *Proteus* was finally compelled to answer to him truthfully.

Perhaps this is one of the oldest stories about re-identifying a person even after intensive appearance changes. In 1961, when discussing the relationship between mental states and behavior, Alvin Plantinga [1] provided one of the first definitions of re-identification:

> "To re-identify a particular, then, is to identify it as (numerically) the same particular as one encountered on a previous occasion".

Person re-identification had thus been studied in various research and documentation areas such as metaphysics [1], psychology [2], and logic [3]. All these works are grounded on Leibniz's Law which claims that "there cannot be separate objects or entities that have all their properties in common."

In the modern computer vision community, the task of person re-ID shares similar insights with the old times. In video surveillance, when being presented with a person-of-interest (query), person re-ID tells whether this person has been observed in another place (time) by another camera. The emergence of this task can be attributed to 1) the increasing demand of public safety and 2) the widespread large camera networks in theme parks, university campuses and streets, *etc*. Both causes make it extremely expensive to rely solely on brute-force human labor to accurately and efficiently spot a person-of-interest or to track a person across cameras.

Technically speaking, a practical person re-ID system in video surveillance can be broken down into three modules, *i.e.,* person detection, person tracking, and person retrieval. It is generally believed that the first two modules are independent computer vision tasks, so most re-ID works focus on the last module, *i.e.,* person retrieval. In this survey, if not specified, person re-ID refers to the person retrieval module. From the perspective of computer vision, the most challenging problem in re-ID is how to correctly match two images of the same person under intensive appearance changes, such as lighting, pose, and viewpoint, which has important scientific values. Given its research and application significance, the re-ID community is fast growing, evidenced by an increasing number of publications in the top venues (Fig. 1).

1.1 Organization of This Survey

Some person re-ID surveys exist [4], [5], [6], [7]. In this survey, we mainly discuss the vision part of re-ID, which is also a focus in the community, and refer readers to the camera calibration and view topology methods in [5]. Another difference from previous surveys is that we focus on different re-ID subtasks currently available or likely to be visible in the future, instead of very detailed techniques or architectures. Special emphasis is given deep learning methods, end-to-end re-ID and very large scale re-ID, which are currently popular topics or reflect future trends. This survey first introduces a brief history of person re-ID in Section 1.2 and its relationship with classification and retrieval in Section 1.3. We then describe previous literature in image-based and video-based person re-ID in Section 2 and Section 3, respectively. Both sections categorize methods into hand-crafted and deeply-learned systems. In Section 4, since the

- *L. Zheng and Y. Yang are with the Centre for Quantum Computation and Intelligent Systems, University of Technology at Sydney, NSW, Australia.*
  E-mail: liangzheng06@gmail.com, yee.i.yang@gmail.com
- *A. Hauptmann is with the School of Computer Science at Carnegie Mellon University, with a joint appointment in the Language Technologies Institute.*
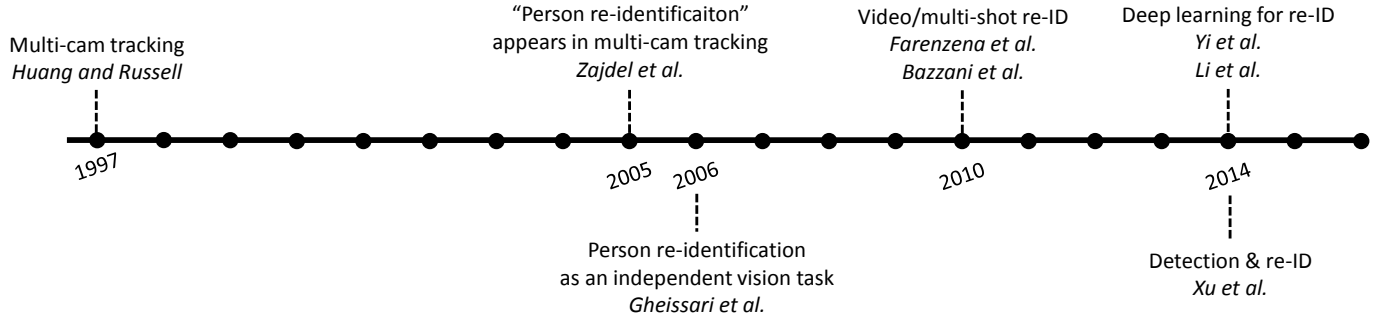 *E-mail: alex@cs.cmu.edu*

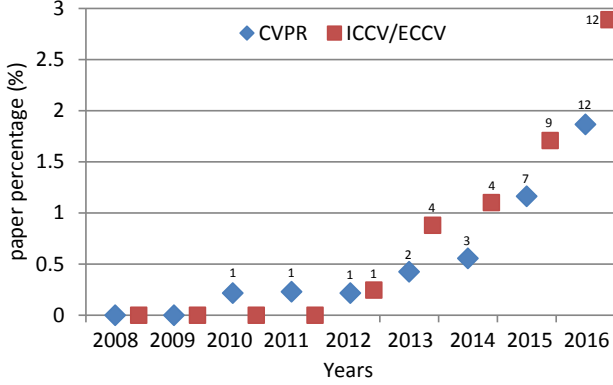Fig. 2: Milestones in the person re-ID history.

Fig. 1: Percentage of person re-ID papers on top conferences over the years. Numbers above the markers indicate the number of re-ID papers.

relationship between detection, tracking, and re-ID has not been extensively studied, we will discuss several previous works and point out future research emphasis. In Section 5, large-scale re-ID which resorts to state-of-the-art retrieval models will be introduced, which is also an important future direction. Some other open issues will be summarized in Section 6, and conclusions will be drawn in Section 7.

1.2 A Brief History of Person Re-ID

Person re-ID research started with multi-camera tracking [8]. Several important re-ID directions have been developed since then. In this survey, we briefly introduce some milestones in person re-ID history (Fig. 2).

Multi-camera tracking. In the early years, person re-ID, as a the term without being formally raised, was tightly twined with multi-camera tracking, in which appearance models were integrated with the geometry calibration among disjoint cameras. In 1997, Huang and Russell [9] proposed a Bayesian formulation to estimate the posterior of predicting the appearance of objects in one camera given evidence observed in other camera views. The appearance model includes multiple spatial-temporal features such as color, vehicle length, height and width, velocity, and time of observation. A comprehensive survey of multi-camera tracking can be accessed in [8].

Multi-camera tracking with explicit "re-identification". To our knowledge, the first work on multi-camera tracking where the term "person re-identification" is proposed, was published in 2005 by Wojciech Zajdel, Zoran Zivkovic and

Ben J. A. Kröse from the University of Amsterdam [10]. In their ICRA'05 paper entitled "Keeping track of humans: Have I seen this person before?", Zajedel *et al.* aims to "re-identify a person when it leaves the field of view and re-enters later". In their method, a unique, latent label is assumed for every person, and a dynamic Bayesian network is defined to encode the probabilistic relationship between the labels and features (color and spatial-temporal cues) from the tracklets. The ID of an incoming person is determined by the posterior label distributions computed by an approximate Bayesian inference algorithm.

The independence of re-ID (image-based). One year later in 2006, Gheissari *et al.* [11] employed only the visual cues of persons after a spatial-temporal segmentation algorithm for foreground detection. Visual matching based on color and salient edgel histograms is performed by either an articulated pedestrian model or the Hessian-Affine interest point operator. Experiments are conducted on a dataset with 44 persons captured by 3 cameras with moderate view overlap. Note that, although Gheissari *et al.* [11] design a spatial-temporal segmentation method using the video frames, neither the feature design nor matching processes use the video information, so we classify [11] into image-based re-ID. This work [11] marks the separation of person re-ID from multi-camera tracking, and its beginning as an independent computer vision task.

Video-based re-ID. Initially intended for tracking in videos, most re-ID works focus on image matching instead. In the year 2010, two works [12], [13] were proposed for multi-shot re-ID, in which frames are randomly selected. Color is a common feature used in both works, and Farenzena *et al.* [13] additionally employ a segmentation model to detect the foreground. For distance measurement, both works calculate the minimum distance among bounding boxes in two image sets, and Bazzani *et al.* further use the Bhattacharyya distance for the color and generic epitome features. It is shown that using multiple frames per person effectively improves over the single-frame version [12], [13] and that re-ID accuracy will saturate as the number of selected frames increases [12].

Deep learning for re-ID. The success of deep learning in image classification [14] spreads to re-ID in 2014, when Yi *et al.* [15] and Li *et al.* [16] both employ a siamese neural network [17] to determine if a pair of input images belong to the same ID. The reason for choosing the siamese model is probably that the number of training samples for each identity is limited (usually two). Aside from some variations in parameter settings, the main differences are that [15] adds

Task	Train Class	Test Class	Advantage
Classification	available	seen	discri. learning
Retrieval	not available	unseen	efficiency
Person re-ID	available	unseen	discri. + efficiency?

TABLE 1: Comparing re-ID with classification and retrieval

an additional cost function in the network, while [16] uses a finer body partitioning. The experimental datasets do not overlap in [15] and [16], so the two methods are not directly comparable. Although its performance is not stable yet on the small datasets, deep learning methods has since become a popular option in re-ID

End-to-end image-based re-ID. While a majority of works use hand-cropped boxes or boxes produced by a fixed detector in their experiments, it is necessary to study the impact of pedestrian detectors on re-ID accuracy. In 2014, Xu *et al.* [18] addressed this topic by combining the detection (commonness) and re-ID (uniqueness) scores. It is shown that on the CAMPUS dataset, jointly considering detection and re-ID confidence leads to higher person retrieval accuracy than using them separately.

1.3 Relationship with Classification and Retrieval

Person re-ID lies inbetween image classification [14] and instance retrieval [19] in terms of the relationship between training and testing classes (Table 1). For image classification, training images are available for each class, and testing images fall into these predefined classes, denoted as previously "seen" in Table 1. For instance retrieval, usually there is no training data because one does not know the content of the query in advance and the gallery may contain various types of objects. So the training classes are "not available" and the testing classes (queries) are denoted as previously "unseen".

Compared to image classification, person re-ID is similar in that the training classes are available, which includes images of different identities. Person re-ID is also similar to instance retrieval in that the testing identities are unseen: they do not have overlap with the training identities, except that both training and testing images are of pedestrians.

As a consequence, person re-ID can be positioned to take advantage of both classification and retrieval. On the one hand, using training classes, discriminative distance metrics [20] or feature embeddings [16], [21] can be learned in the person space. On the other hand, when it comes to retrieval, efficient indexing structures [22] and hashing techniques [23] can be beneficial for re-ID in a large gallery. In this survey, both effective learning and efficient retrieval approaches will be introduced or pointed out as important future directions.

2 IMAGE-BASED PERSON RE-ID

Since the work by Gheissari *et al.* in 2006 [11], person re-ID has mostly been explored using single images. Let us consider a closed-world toy model, in which $\mathcal{G}$ is a gallery (database) composed of N images, denoted as $\{g_i\}_{i=1}^N$. They belong to N different identities $1, 2, ..., N$. Given a probe (query) image q, its identity is determined by:

$$i^* = \arg\max_{i \in 1,2,...,N} \text{sim}(q, g_i), \tag{1}$$

where i^* is the identity of probe q, and $\text{sim}(\cdot, \cdot)$ is some kind of similarity function.

2.1 Hand-crafted Systems

It is apparent from Eq. 1 that two components are necessary for a toy re-ID system, *i.e.,* image description and distance metrics.

2.1.1 Pedestrian Description

In pedestrian descriptions, the most commonly used feature is color, while texture features are less frequent. In [13], the pedestrian foreground is segmented from the background, and a symmetrical axis is computed for each body part. Based on body configuration, the weighted color histogram (WH), the maximally stable color regions (MSCR), and the recurrent high-structured patches (RHSP) are computed. WH assigns larger weights to pixels near the symmetrical axis and forms a color histogram for each part. MSCR detects stable color regions and extracts features such as color, area, and centroid. RHSP instead is a texture feature capturing recurrent texture patches. Gheissari *et al.* [11] propose a spatial-temporal segmentation method to detect stable foreground regions. For a local region, an HS histogram and an edgel histogram are computed. The latter encodes the dominant local boundary orientation and the RGB ratios on either sides of the edgel. Gray and Tao [24] use 8 color channels (RGB, HS, and YCbCr) and 21 texture filters on the luminance channel, and the pedestrian is partitioned into horizontal stripes. A number of later works [25], [26], [27] employ the same set of features as [24]. Similarly, Mignon *et al.* [28] build the feature vector from RGB, YUV and HSV channels and the LBP texture histograms in horizontal stripes.

Compared to the earlier works described above, hand-crafted features have remained more or less the same in recent years [20], [29], [30], [31], [32]. In a series of works by Zhao *et al.* [30], [33], [34], the 32-dim LAB color histogram and the 128-dim SIFT descriptor are extracted from each 10×10 patch densely sampled with a step size of 5 pixels; this feature is also used in [35]. Adjacency constrained search is employed to find the best match for a query patch in horizontal stripes with similar latitudes in a gallery image. Das *et al.* [36] apply HSV histograms on the head, torso and legs from the silhouette proposed in [12]. Li *et al.* [31] also extract local color descriptors from patches but aggregate them using hierarchical Gaussianization [37] to capture spatial information, a procedure followed by [38]. Pedagadi *et al.* [39] extract color histograms and moments from HSV and YUV spaces before dimension reduction using PCA. Liu *et al.* [40] extract the HSV histogram, gradient histogram and the LBP histogram for each local patch. To improve the robustness of the RGB values against photometric variance, Yang *et al.* [41] introduce the salient color names based color descriptor (SCNCD) for global pedestrian color descriptions. The influence of the background and different color spaces are also analysed. In [20], Liao *et al.* propose the local maximal occurrence (LOMO) descriptor, which includes the color and SILTP histograms. Bins in the same horizontal stripe undergo max pooling and a three-scale pyramid model is built before a log transformation. LOMO is later employed by [42], [43] and a similar set of features is used by Chen *et al.* [32]. In [44], Zheng *et al.* propose extracting the 11-dim color names descriptor [45] for each local patch, and aggregating them into a global vector through a Bag-of-Words (BoW) model. In

[46], a hierarchical Gaussian feature is proposed to describe color and texture cues, which models each region by multiple Gaussian distributions. Each distribution represents a patch inside the region.

Apart from directly using low-level color and texture features, another good choice is the attribute-based features which can be viewed as mid-level representations. It is believed that attributes are more robust to image translations compared to low-level descriptors. In [47], Layne *et al.* annotate 15 binary attributes on the VIPeR dataset related to attire and soft biometrics. The low-level color and texture features are used to train the attribute classifiers. After attribute weighting, the resulting vector is integrated in the SDALF [13] framework to fuse with other visual features. Liu *et al.* [48] improve the latent Dirichlet allocation (LDA) model using annotated attributes to filter out noisy LDA topics. Liu *et al.* [49] propose discovering some pedestrian prototypes with common attributes in an unsupervised manner and adaptively determine the feature weights of different query person according to the prototypes. Some recent works borrow external data for attribute learning. In [50], Su *et al.* embed the binary semantic attributes of the same person but different cameras into a continuous low-rank attribute space, so that the attribute vector is more discriminative for matching. Shi *et al.* [51] propose learning a number of attributes including color, texture, and category labels from existing fashion photography datasets. These attributes are directly transferred to re-ID under surveillance videos and achieve competitive results. Recently, Li *et al.* [52] collected a large-scale dataset with richly annotated pedestrian attributes to facilitate attribute-based re-ID methods.

2.1.2 Distance Metric Learning

In hand-crafted re-ID systems, a good distance metric is critical for its success, because the high-dimensional visual features typically do not capture the invariant factors under sample variances. A comprehensive survey of the metric learning methods can be accessed in [53]. These metric learning methods are categorized w.r.t supervised learning versus unsupervised learning, global learning versus local learning, *etc*. In person re-ID, the majority of works fall into the scope of supervised global distance metric learning.

The general idea of global metric learning is to keep all the vectors of the same class closer while pushing vectors of different classes further apart. The most commonly used formulation is based on the class of Mahalanobis distance functions, which generalizes Euclidean distance using linear scalings and rotations of the feature space. The squared distance between two vectors x_i and x_j can be written as,

$$d(x_i, x_j) = (x_i - x_j)^{\mathrm{T}} \mathbf{M}(x_i - x_j), \qquad (2)$$

where $\mathbf{M}$ is a positive semidefinite matrix. Equation 2 can be formulated into the convex programming problem suggested by Xing *et al.* [54].

In person re-ID, currently the most popular metric learning method, *i.e.*, KISSME [55] is based on Eq. 2. In this method [55], the decision on whether a pair (i, j) is similar or not is formulated as a likelihood ratio test. The pairwise difference ($x_{i,j} = x_i - x_j$) is employed and the difference space is assumed to be a Gaussian distribution with a zero mean. It

is shown in [55] that the Mahalanobis distance metric can be naturally derived from the log-likelihood ratio test and in practice, the principle component analysis (PCA) is applied to the data points to eliminate dimension correlations.

Based on Eq. 2, a number of other metric learning methods have been introduced. In the early days, some classic metric learning methods target at nearest neighbor classification. Weinberger *et al.* [56] propose the large margin nearest neighbor Learning (LMNN) method which sets up a perimeter for the target neighbors (matched pairs) and punishes those invading the perimeter (imposters). This method belongs to the supervised local distance metric learning category [53]. To avoid the overfitting problems encountered in LMNN, Davis *et al.* [57] propose the information-theoretic metric learning (ITML) as a trade-off between satisfying the given similarity constraints and ensuring that the learned metric is close to the initial distance function.

In recent years, Hirzer *et al.* [58] proposed relaxing the positivity constraint which provides a sufficient approximation for the matrix $\mathbf{M}$ with a much lower computational cost. Chen *et al.* [38] add a bilinear similarity in addition to the Mahalanobis distance, so that cross-patch similarities can be modeled. In [31], the global distance metric is coupled with the local adaptive threshold rule which additionally contains the orthogonal information of (x_i, x_j). In [59], Liao *et al.* suggest perserving with a positive semidefinite constraint and propose weighting the positive and negative samples differently. Yang *et al.* [60] consider both the differences and commonness between image pairs and show that the covariance matrices of dissimilar pairs can be inferred from those of the similar pairs, which makes the learning process scalable to large datasets.

Other than learning distance metrics, some works focus on learning discriminative subspaces. Liao *et al.* [20] propose learning the projection $\boldsymbol{w}$ to a low-dimensional subspace with cross-view data solved in a similar manner to linear discriminant analysis (LDA) [61],

$$\mathcal{J}(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_w \boldsymbol{w}}, \qquad (3)$$

where $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are the between-class and within-class scatter matrices, respectively. Then, a distance function is learned in the resulting subspace using KISSME. To learn $\boldsymbol{w}$, Zhang *et al.* [42] further employ the null Foley-Sammon transform to learn a discriminative null space which satisfies a zero within-class scatter and a positive between-class scatter. For dimension reduction, Pedagadi *et al.* [39] sequentially combine the unsupervised PCA (principle component analysis) and supervised local Fisher discriminative analysis which preserves the local neighborhood structure. In [28], the pairwise constrained component analysis (PCCA) is proposed which learns a linear mapping function to be able to work directly on high-dimensional data, while ITML and KISSME should be preceded by a step of dimension reduction. In [62], Xiong *et al.* further propose improved versions of two existing subspace projection methods, *i.e.*, regularized PCCA [28] and kernel LFDA [39].

Aside from the methods that use Mahalanobis distance (Eq. 2), some use other learning tools such as support vector machine (SVM) or boosting. Prosser *et al.* [25] propose learning a set of weak RankSVMs which are subsequently

Model	Identification		Verification	
	rank-1 (%)	mAP (%)	rank-1 (%)	mAP (%)
AlexNet [14]	56.03	32.38	41.24	22.47
VGG-16 [67]	64.34	40.77	42.99	24.29
Residual-50 [68]	72.54	46.00	60.12	40.54

TABLE 2: Comparison of the identification and verification (siamese) models on the Market-1501 dataset (single query).

assembled into a stronger ranker. In [63], a structural SVM is employed to combine different color descriptors at decision level. In [43], Zhang *et al.* learn a specific SVM for each training identity and map each testing image to a weight vector inferred from its visual features. Gray and Tao [24] propose using the AdaBoost algorithm to select and combine many different kinds of simple features into a single similarity function.

2.2 Deeply-learned Systems

CNN-based deep learning models have been popular since Krizhevsky *et al.* [14] won ILSVRC'12 by a large margin. The first two works in re-ID to use deep learning were [15], [16] as mentioned in Section 1.2 and Fig. 2. Generally speaking, two types of CNN models are commonly employed in the community. The first type is the classification model as used in image classification [14] and object detection [64]. The second is the siamese model using image pairs [65] or triplets [66] as input. The major bottleneck of deep learning in re-ID is the lack of training data. Most re-ID datasets provide only two images for each identity such as VIPeR [24], so currently most CNN-based re-ID methods focus on the siamese model. In [15], an input image is partitioned into three overlapping horizontal parts, and the parts go through two convolutional layers plus a fully connected layer which fuses them and outputs a vector for this image. The similarity of the two output vectors are computed using the cosine distance. The architecture designed by Li *et al.* [16] is different in that a patch matching layer is added which multiplies the convolution responses of two images in different horizontal stripes, similar to ACS [30] in spirit. Later, Ahmed *et al.* [69] improved the siamese model by computing the cross-input neighborhood difference features, which compares the features from one input image to features in neighboring locations of the other image. While [16] uses product to compute patch similarity in similar latitude, Ahmed *et al.* [69] use subtraction. Wu *et al.* [70] deepen the networks using convolutional filters of smaller sizes, called "PersonNet". In [71], Varior *et al.* incorporate long short-term memory (LSTM) modules into a siamese network. LSTMs process image parts sequentially so that the spatial connections can be memorized to enhance the discriminative ability of the deep features. Varior *et al.* [72] propose inserting a gating function after each convolutional layer to capture effective subtle patterns when a pair of testing images are fed into the network. This method achieves state-of-the-art accuracy on several benchmarks, but its disadvantage is also obvious. The query has to pair with each gallery image before being sent into the network - a time inefficient process in large datasets. Similar to [72], Liu *et al.* [73] propose integrating a soft attention based model in a siamese network to adaptively focus on the important local

parts of an input image pair; however, this method is also limited by computational inefficiency. While these works use image pairs, Cheng *et al.* [74] design a triplet loss function that takes three images as input. After the first convolutional layer, four overlapping body parts are partitioned for each image and fused with a global one in the FC layer. Su *et al.* [75] propose a three-stage learning process which includes attribute prediction using an independent dataset and an attributes triplet loss trained on datasets with ID labels.

A drawback of the siamese model is that it does not make full use of re-ID annotations. In fact, the siamese model only needs to consider pairwise (or triplet) labels. Telling whether an image pair is similar (belong to the same identity) or not is a weak label in re-ID. Another potentially effective strategy consists of using a classification/identification mode, which makes full use of the re-ID labels. In [76], training identities from multiple datasets jointly form the training set and a softmax loss is employed in the classification network. Together with the proposed impact score for each FC neuron and a domain guided dropout based on the impact score, the learned generic embeddings yield competitive re-id accuracy. On larger datasets, such as PRW and MARS, the classification model achieves good performance without careful training sample selection [21], [77]. Yet the application of the identification loss requires more training instances per ID for model convergence. For comparison, this survey presents some baseline results for both types of models. In Table 2, we implement the identification and verification models on the Market-1501 dataset [44]. All the networks use the default parameter settings, and are fine-tuned from the ImageNet [78] pre-trained models. Images are resized to 224×224 before being fed into the network. The initial learning rate is set to 0.001 and reduced by a factor of 0.1 after each epoch. Training is done after 36 epochs. We can clearly observe that the identification model outperforms the verification model, and that the residual-50 model [68] yields state-of-the-art re-ID accuracy on Market-1501 compared with recent results [71], [72], [75].

The above-mentioned works learn deep features in an end-to-end manner, and there are alternatives that take low-level features as input. In [79], low-level descriptors including SIFT and color histograms are aggregated into a single Fisher Vector [80] for each image. The hybrid network builds fully connected layers on the input Fisher vectors and enforces the linear discriminant analysis (LDA) as an objective function to produce embeddings that have low intra-class variance and high inter-class variance. Wu *et al.* [81] propose concatenating the FC feature and a low-level feature vector, which is followed by another FC layer before the softmax loss layer. This method constrains the FC features using the hand-crafted features.

2.3 Datasets and Evaluation

2.3.1 Datasets

A number of datasets for image-based re-ID have been released, and some commonly used datasets are summarized in Table 3. The most tested benchmark is VIPeR. It contains 632 identities, and two images for each identity. 10 random train/test splits are used for stable performance, and each split has 316 different identities in both the training

Dataset	time	#ID	#image	#camera	label	evaluation
VIPeR	2007	632	1,264	2	hand	CMC
iLIDS	2009	119	476	2	hand	CMC
GRID	2009	250	1,275	8	hand	CMC
CAVIAR	2011	72	610	2	hand	CMC
PRID2011	2011	200	1,134	2	hand	CMC
WARD	2012	70	4,786	3	hand	CMC
CUHK01	2012	971	3,884	2	hand	CMC
CUHK02	2013	1,816	7,264	10 (5 pairs)	hand	CMC
CUHK03	2014	1,467	13,164	2	hand/DPM	CMC
RAiD	2014	43	1,264	4	hand	CMC
PRID 450S	2014	450	900	2	hand	CMC
Market-1501	2015	1,501	32,668	6	hand/DPM	CMC/mAP

TABLE 3: Statistics of some commonly used datasets [16], [36], [44], [82], [83], [84], [85], [86], [87], [88], [89], [90] for image-based re-ID.

and testing sets. These datasets reflect various scenarios. For example, the GRID dataset [84] was collected in an underground station, iLIDS [83] was captured at an airport arrival hall, and CUHK01 [88], CUHK02 [89], CUHK03 [16] and Market-1501 [44] were collected in a university campus. Over recent years, progress can observed in several aspects.

First, the dataset scale is increasing. Many of these datasets are relatively small in size, especially those of the early days, but recent datasets, such as CUHK03 and Market-1501, are larger. Both have over 1,000 IDs and over 10,000 bounding boxes, and both provide good amount of data for training deep learning models. That said, we must admit that the current data volume is still far from satisfactory. The community is in great need of larger datasets.

Second, the bounding boxes tend to be produced by pedestrian detectors (such as DPM [91] and ACF [92]) instead of being hand-drawn. For practical applications, it is infeasible to draw gallery bounding boxes using human labor, so detectors must be used. This may cause the bounding boxes to deviate from ideal ones. It is shown in [16] that using detected bounding boxes usually leads to compromised re-ID accuracy compared to hand-drawn ones due to detector errors such as misalignment. In [44], a number of false detection results (on the background) are included in the gallery, which is inevitable when detectors are used. The experiments in [44] show that re-ID accuracy drops as more distractors are added to the gallery. As a consequence, it is beneficial for the community to study datasets with practical imperfections such as false detection and misalignment.

Third, more cameras are used during collection. For example, each identity in Market-1501 can be captured by up to 6 cameras. This design calls for metric learning methods that have good generalization ability, instead of being carefully tuned between a certain camera pair. In fact, in a city-scale camera network with n, the number of camera pairs is C_n^2, so it is prohibitive to collect annotated data from each camera and train C_n^2 distance metrics.

For more detailed descriptions of these datasets, we refer to survey [5] and website[1].

2.3.2 Evaluation Metrics

When evaluating re-ID algorithms, the cumulative matching characteristics (cmc) curve is usually used. CMC represents

the probability that a query identity appears in different-sized candidate lists. No matter how many ground truth matches there are in the gallery, only the first match is counted in the CMC calculation. So basically, CMC is accurate as an evaluation method only when one ground truth for each query exists. This measurement is acceptable, in practice, when people care more about returning the ground truth match in the top positions of the rank list.

For research integrity, however, when multiple ground truths exist in the gallery, Zheng *et al.* [44] propose using the mean average precision (mAP) for evaluation. The motivation is that a perfect re-ID system should be able to return all true matches to the user. The case might be that two systems are equally competent at spotting the first ground truth, but have different retrieval recall ability. In this scenario, CMC does not have enough discriminative ability but mAP does. Therefore, mAP is used together with CMC for the Market-1501 dataset where multiple ground truths from multiple cameras exist for each query. Later, in [71], [72], [93], mAP results are also reported for datasets with multiple ground truths per query.

2.3.3 Re-ID Accuracy Over the Years

In this section, we summarize re-ID accuracy on several representative datasets over the years in Fig. 3. The presented datasets are VIPeR [82], CUHK01 [88], iLIDS [83], PRID 450S [90], CUHK03 [16], and Market-1501 [44]. We broadly classify the current methods into two types, *i.e.*, hand crafted and deeply learned. For each dataset, representative methods that report the highest re-ID accuracy in the corresponding year are shown. From these results, three major insights can be drawn.

First, a clear trend of performance improvement can be observed from the six datasets over the years. On VIPeR, CUHK01, i-LIDS, PRID 450S, CUHK03, and Market-1501, we observe a performance increase of +51.9%, +56.7%, +35.0%, +42.6%, +57.2%, and +31.62%, respectively. For example, on the most studied dataset VIPeR [82], from the year 2008 to 2016, representative works [13], [24], [30], [32], [85], [94], [95] witness a rank-1 accuracy from 12.0% in 2008 [24] to 63.9% in 2015 [94], an improvement of +51.9%. For the Market-1501 dataset, since its release in 2015, the state-of-the-art results have increased from 44.42% [44] to 76.04% [72], an improvement of 31.62%.

Second, with the exception of VIPeR, deep learning methods yield a new state of the art on the remaining 5 datasets. On these 5 datasets (CUHK01, i-LIDS, PRID 450S, CUHK03, and Market-1501), the performance of deep learning is superior to hand-crafted systems. On CUHK03 and Market-1501, the two largest datasets so far, we observe overwhelming advantage for deep learning [72], [76] compared to the (also extensive) tests of hand-crafted methods. Since VIPeR is relatively small, the advantage of deep learning cannot be tested to the full; instead, a hand-crafted metric learning may be more advantageous in this setting. Considering the cases in image classification and object detection, it is highly possible that deeply learned systems will continue dominating the re-ID community over the next few years.

Third, we speculate that there is still much room for further improvement, especially when larger datasets are to

---

1. http://robustsystems.coe.neu.edu/sites/robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html
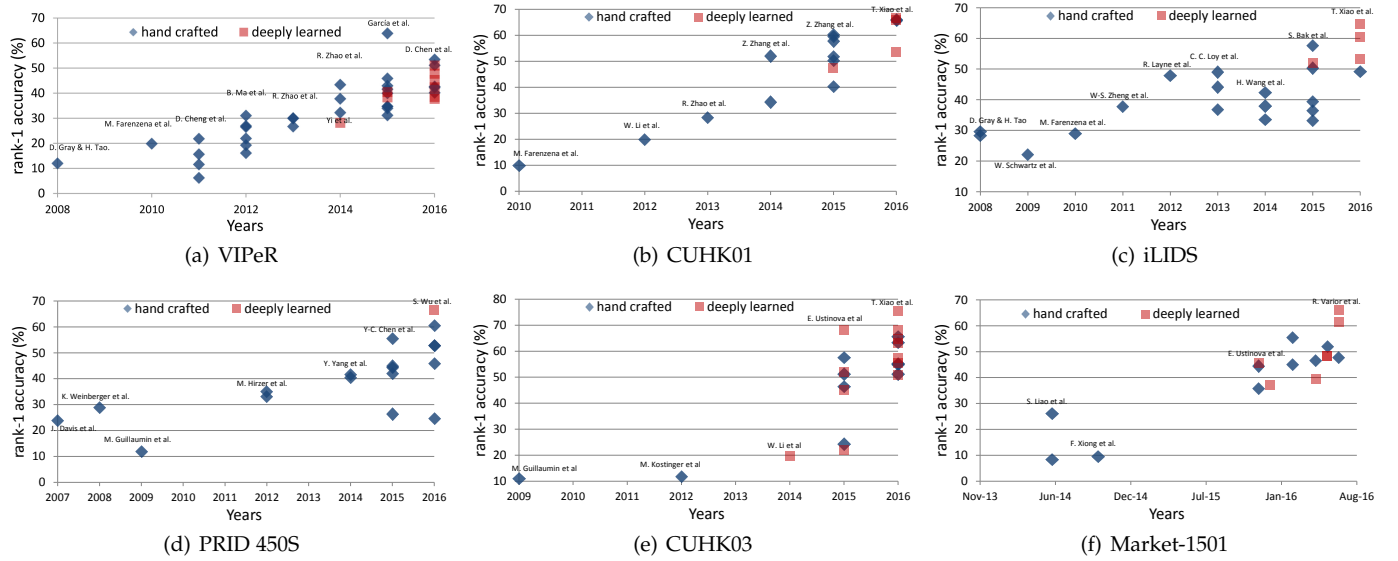
Fig. 3: Person re-ID accuracy on (a) VIPeR [82] (b) iLIDS [83] (c) GRID [84] (d) CUHK01 [88] (e) CUHK03 [16] and (f) Market-1501 [44] over the years. Results from top venues using hand-crafted or deeply learned systems are presented. For CUHK03, we record results on the detected data, and for Market-1501, results using single queries are used. Since Market-1501 was released recently, results on this dataset are plotted according to their publication (or ArXiv) time.

be released. For example, on the Market-1501 dataset, while the best rank-1 accuracy is 65.88% without using multiple queries [72], mAP is quite low (39.55%). This indicates that although it is relative easy to find the first true match (rank-1 accuracy) among a pool of 6 cameras, it is not trivial to locate the hard positives and thus achieve a high recall (mAP). On the other hand, although we seem to be able to achieve 60% to 70% rank-1 accuracy on these datasets, we must keep in mind that these datasets receive a very small proportion of practical usage. In fact, apart from [44], it is also reported in [96] a 10-fold gallery size increase leads to a 10-fold decrease in rank-1 accuracy, resulting in a single-digit rank-1 score even for the best-performing methods. As a consequence, considering the low mAP (re-ID recall) and the small scale of current datasets, we are more than optimistic that important breakthroughs are to be expected in image-based re-ID.

3 VIDEO-BASED PERSON RE-ID

In literature, person re-ID is mostly explored with single images (single shot). In recent years, video-based re-ID has become popular due to the increased data richness which induces more research possibilities. It shares a similar formulation to image-based re-ID as Eq. 1. Video-based re-ID replaces images q and g with two sets of bounding boxes $\{q_i\}_{i=1}^{n_q}$ and $\{g_j\}_{j=1}^{n_g}$, where n_q and n_g are the number of bounding boxes within each video sequence, respectively. As important as the bounding box features are, video-based methods pay additional attention to multi-shot matching schemes and the integration of temporal information.

3.1 Hand-crafted Systems

The first two trials [12], [13] in 2010 were both hand-crafted systems. They basically use color-based descriptors and optionally employ foreground segmentation to detect the

pedestrian. They use similar image features to image-based re-ID methods, where the major difference is the matching function. As mentioned in Section 1.2, both methods commonly calculate the minimum Euclidean distance between two sets of bounding box features as the set similarity. In essence, such methods should be classified into "multi-shot" person re-ID, where the similarity between two sets of frames plays a critical role. This multi-shot matching strategy is adopted by later works [97], [98]. In [86], multiple shots are used to train a descriminative boosting model based on a set of covariance features. In [99], the SURF local feature is used to detect and describe interest points within short video sequences that are in turn indexed in the KD-tree to speed up matching. In [11], a spatial-temporal graph is generated to identify spatial-temporal stable regions for foreground segmentation. The the local descriptions are then calculated using a clustering method over time to improve matching performance. Cong *et al.* [100] employ the manifold geometric structures from video sequences to construct more compact spatial descriptors with color-based features. Karaman *et al.* [101] propose using the conditional random field (CRF) to incorporate constraints in the spatial and temporal domains. In [102], colors and selected face images are used to build a model over frames that capture the characteristic appearance as well as its variations over time. Karanam *et al.* [103] make use of multi-shots for a person and propose that the probe feature be presented as a linear combination of the same person in the gallery. Multiple shots of an identity can also be employed to enhance body part alignment. In [85], in the effort to look for precise part-to-part correspondence, Cheng *et al.* propose an iterative algorithm in which the fitting of the pictorial structure becomes more accurate after each iteration due to the improvement of part detectors. In [104], pedestrian poses are estimated and frames with the same pose are matched with higher confidence.

The above methods typically build appearance models based on multiple shots, and a recent trend is to incorporate temporal cues in the model. Wang *et al.* [105] propose using spatial-temporal descriptors to re-identify pedestrians. Its features include HOG3D [106] and the gait energy image (GEI) [107]. By designing a flow energy profile (FEP), walking cycles are detected so that frames around the local minimum/maximum are used to extract motion features. Finally, reliable spatial-temporal features are selected and matched through a discriminative video ranking model. In [108], Liu *et al.* propose de-composing a video sequence into a series of units that represent body-actions corresponding to certain action primitives, from which Fisher vectors are extracted for the final representation of the person. Gao *et al.* [109] make use of the periodicity property of pedestrians and divide the walking cycle into several segments which are described by temporally aligned pooling. In [110], a new spatial-temporal descriptor is proposed based on densely computed multi-directional gradients and discarding noisy motion occurring over a short period.

Distance metric learning is also important when matching videos. In [111], a set verification method is proposed in which a transfer ranking is employed to tell whether the query matches one of the images belonging to the same identity. In [89], the multi-shot extension of the proposed local match model minimizes the distance of the best-matched pairs and reduces the number of cross-view transformations. In [112], Zhu *et al.* propose simultaneously learning intra- and inter-video distance metrics to make video representation more compact and to discriminate videos of different identities. You *et al.* [113] propose the top-push distance learning method which optimizes the top-rank matching in video re-ID by selecting discriminative features.

3.2 Deeply-learned Systems

In video-based re-ID, the data volume is typically larger than image-based datasets, because each tracklet contains a number of frames (Table 4).

A basic difference between video-based and image-based re-ID is that with multiple images for each matching unit (video sequence), either a multi-match strategy or a single-match strategy after video pooling should be employed. The multi-match strategy is used in older works [12], [13], which induces higher computational cost and may be problematic on large datasets. On the other hand, pooling-based methods aggregates frame-level features into a global vector, which has better scalability. As a consequence, current video-based re-ID methods typically involve the pooling step. This step can be max/average pooling as [21], [114], or learned by a fully connected layer [115]. In Zheng *et al.*'s system [21], temporal information is not explicitly captured; instead, frames of an identity are viewed as its training samples to train a classification CNN model with softmax loss. Frame features are aggregated by max pooling which yield competitive accuracy on three datasets. These methods are proven to be effective, and yet there is plenty of space for improvement. With respect to this point, the re-ID community can borrow ideas from the community of action/event recognition. For example, Xu *et al.* [116] propose aggregating the column features in the 5th convolutional layer of CaffeNet into

Dataset	time	#ID	#track	#bbox	#cam.	label	evaluation
ETHZ	2007	148	148	8,580	1	hand	CMC
3DPES	2011	200	1,000	200k	8	hand	CMC
PRID-2011	2011	200	400	40k	2	hand	CMC
iLIDS-VID	2014	300	600	44k	2	hand	CMC
MARS	2016	1261	20,715	1M	6	DPM&GMMCP	mAP&CMC

TABLE 4: Statistics of some currently available datasets [21], [86], [105], [122], [123] for video-based re-ID.

Fisher vectors [80] or VLAD [117], in direct CNN feature transfer. Fernando *et al.* [118] propose a learning-to-rank model to capture how frame features evolve over time in a video, which yields video descriptors of video-wide temporal dynamics. Wang *et al.* [119] embed a multi-level encoding layer into the CNN model and produce video descriptors of varying sequence lengths.

Another good practice consists of injecting temporal information in the final representation. In hand-crafted systems, Wang *et al.* [105] and Liu *et al.* [108] use pure spatial-temporal features on the iLIDS-VID and PRID-2011 datasets and report competitive accuracy. In [21], however, it is shown that the spatial-temporal features are not sufficiently discriminative on the MARS dataset, because many pedestrians share similar waling motion under the same camera, and because motion feature of the same person can be distinct in different cameras. The point made in [21] is that appearance features are critical in a large-scale video re-ID system. That said, this survey calls for attention to the recent works of [114], [115], [120], in which appearance features (*e.g.,* CNN, color and LBP) are used as the starting point to be fed into RNN networks to capture the time flow between frames. In [114], features are extracted from consecutive video frames through a CNN model, and then fed through a recurrent final layer, so that information flow between time-steps is allowed. The features are then combined using max or average pooling to yield an appearance feature for the video. All these structures are incorporated into a siamese network. A similar architecture is used in [120]. Their difference is two-fold. First, a particular type of RNN, the Gated Recurrent Unit (GRU) is used in [120]. Second, an identification loss is adopted in [114], which is beneficial for loss convergence and performance improvement. While the two works [114], [120] employ the siamese network for loss computation, Yan *et al.* [115] and Zheng *et al.* [21] use the identification model which classifies each input video into their respective identities. In [115], hand-crafted low-level features such as color and LBP are fed into several LSTMs and the LSTM outputs are connected to a softmax layer. In action recognition, Wu *et al.* [121] propose extracting both appearance and spatial-temporal features from a video and build a hybrid network to fuse the two types of features. In this survey, we note that perhaps the discriminative combination of appearance and spatial-temporal models is an effective solution in future video re-ID research.

3.3 Datasets and Evaluation

Several multi-shot re-ID datasets exist, *e.g.,* ETH [122], 3DPES [123], PRID-2011 [86], iLIDS-VID [105], and MARS [21]. Some statistics of these datasets are summarized in Table 4. The ETH dataset uses a single moving camera. It contains three sequences and multiple images from each sequence

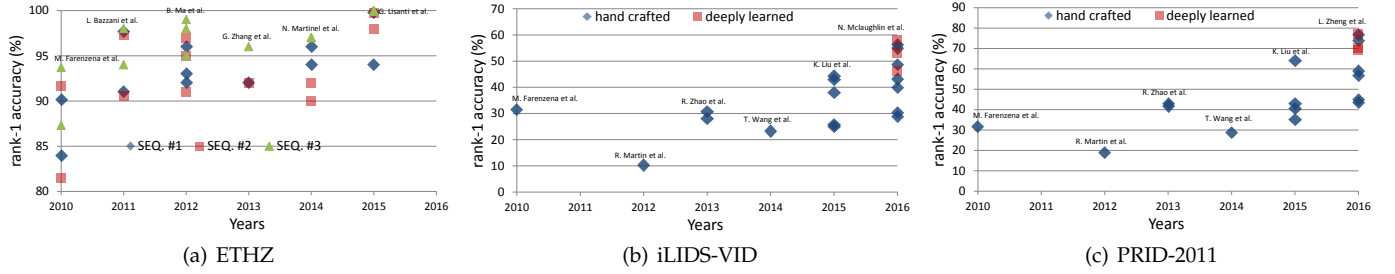(a) ETHZ   (b) iLIDS-VID   (c) PRID-2011

Fig. 4: Video-based person re-ID accuracy on (a) ETH sequence 1 [82] (b) PRID-2011 [86] and (c) iLIDS-VID [105] over the years. Results from top venues using hand-crafted or deeply learned systems are presented. For ETHZ, we report results obtained by 5 images per video sequence, and state-of-the-art results on SEQ. #1, SEQ. #2, and SEQ. #3 are drawn.

are provided. This dataset is relatively easy and the re-ID accuracy of the multi-shot scenario is nearly 100% [124]. The 3DPeS dataset is collected with 8 non-overlapping outdoor cameras. Although the videos are released, this dataset is typically used for single-shot re-ID. PRID-2011 and iLIDS-VID are similar in that both datasets were captured by 2 cameras and each identity has 2 video sequences. iLIDS-VID has 300 identities captured under indoor scenes. PRID-2011 has 385 and 749 identities for each outdoor camera, respectively, and in this dataset 200 identities are observed in both cameras. During testing, 178 identities are used for PRID-2011 following the proposal by [105]. It is generally believed that iLIDS-VID is more challenging than PRID-2011 due to extremely heavy occlusion. The MARS dataset [21] was recently released which is a large-scale video re-ID dataset containing 1,261 identities in over 20,000 video sequences. It is produced using the DPM detector [91] and the GMMCP tracker [125]. Due to its recent release, we have not provided an extensive summary of results for the MARS dataset. Figure 4 presents the evaluation of the state-of-the-art results on three representative video (multi-shot) re-ID datasets, *i.e.,* ETHZ, iLIDS-VID, PRID-2011. Two major conclusions are drawn:

First, the ETHZ dataset has reached its performance saturation. In 2015, Lisanti *et al.* [124] and Martinel *et al.* [126] report rank-1 accuracies approximating 100%. In [124], using 5 images per sequence, the rank-1 accuracy of ETHZ sequence 1, 2, and 3 is 99.8%, 99.7%, and 99.9%, respectively. Results with 10 frames per sequence is higher, achieving 100% [124], [126]. The primary reason is that the ETHZ dataset has relatively fewer identities, and the image variance is low due to the use of only one moving camera. This may be the first re-ID dataset to almost accomplish its initial objectives.

Second, active video re-ID research is still being conducted on the iLIDS-VID and PRID-2011 datasets. Since their introduction, we observe clear improvement of their rank-1 accuracy (including the ETHZ dataset). For iLIDS-VID, Wang *et al.* [105] report a rank-1 accuracy of 23.3%, and an absolute improvement of 34.7% can be seen when compared to McLaughlin *et al.* [114]. On PRID-2011, Wang *et al.* [105] report a rank-1 accuracy = 19.0%, and two years later, Zheng *et al.* [21] improve this score by 58.3% using the max pooling of CNN features fine-tuned on the MARS dataset.

Third, deep learning methods are producing overwhelmingly superior accuracy in video-based re-ID. On both the iLIDS-VID and PRID-2011 datasets, the best performing

methods are based on the convolutional neural network with optional insertion of a recurrent neural network [21], [114]. Compared to image-based re-ID, the amount of training data is clearly larger in video re-ID. MARS provides over 500k training frames, compared to 13k in the Market-1501 dataset [44], from which MARS was extended. With these training data, it is feasible to train discriminative networks not only for video-based re-ID, but also for image-based datasets. We also note that, while the rank-1 accuracy on the MARS dataset reaches 68.3%, its mAP is still relatively low (49.3%), and when evaluating the performance of each camera pair, performance is further lowered. As a consequence, we believe that the research of video re-ID still has good potential for improvement.

4 FUTURE: DETECTION, TRACKING AND PERSON RE-ID

4.1 Previous Works

Although person re-ID originates from multi-camera tracking, it is now studied as an independent research topic. In this survey, we view re-ID as an important future direction that will join pedestrian detection and tracking as a scenario, but in a more independent role. Specifically, we consider an end-to-end re-ID system[2] that takes raw videos as input and integrates pedestrian detection and tracking, along with re-identification.

Until recently, most re-ID works are based on two assumptions: first, that the gallery of pedestrian bounding boxes is given; second, that the bounding boxes are hand-drawn, *i.e.,* with perfect detection quality. However, in practice, these two assumptions do not hold. On the one hand, the gallery size varies with the detector threshold. A lower threshold produces more bounding boxes (a larger gallery, higher recall, and lower precision), and vice versa. When the detection recall/precision undergoes changes due to different thresholds, re-ID accuracy does not remain stable. On the other hand, when pedestrian detectors are used, detection errors typically exist with the bounding boxes, such as misalignment, miss-detection, and false alarms. Moreover, when pedestrian trackers are used, tracking errors may lead to outlier frames within a tracklet, *i.e.,* background or pedestrians with different identities. So the quality of pedestrian detection and tracking may have direct influence

---

2. Here, "end-to-end" means spotting a query person from raw videos.

on re-ID accuracy, which has been rarely discussed in the re-ID community. In the following, we will review the several works devoted to this direction.

In initial attempts to address the second problem, several datasets, *i.e.*, CUHK03 [16], Market-1501 [44], and MARS [21], were introduced. These datasets do not assume perfect detection/tracking outputs and are a step closer to practical applications. For example, Li *et al.* [16] show that on CUHK03, re-ID accuracy using the detected bounding boxes is lower than that obtained with hand-drawn bounding boxes. Later works also report this observation [42], [127]. These findings are closely related to practical applications. On MARS, tracking errors (Fig. 8) as well as detection errors are presented, but it remains unknown how tracking errors will affect re-ID accuracy.

Despite the fact that the datasets make progress by introducing detection/tracking errors, they do not evaluate explicitly how detection/tracking affects re-ID, which provides critical insights into how to select detectors/trackers among the vast number of existing works in an end-to-end re-ID system. To our knowledge, the first work on end-to-end person re-ID was proposed by Xu *et al.* [18] in 2014. They use the term "commonness" to describe how an image bounding box resembles a pedestrian, and the term "uniqueness" to indicate the similarity between the gallery bounding box and the query. Commonness and uniqueness are fused by their product in an exponential function. This method works by eliminating the impact of false background detections. Although Xu *et al.* [18] considers the impact of detection on re-ID, its limitation is a lack of comprehensive benchmarking and consideration of the dynamic issue of the gallery.

In 2016, Xiao *et al.* [128] and Zheng *et al.* [77] simultaneously introduce an end-to-end re-ID system based on large-scale datasets. The two works take raw video frames and a query bounding box as input (Fig. 5). One is required to first perform pedestrian detection on the raw frames, and the resulting bounding boxes form the re-ID gallery. Then, classic re-ID is leveraged. This process, called "person search" in [18], [128], is no longer restricted to re-ID (Fig. 5(b)): it pays equal attention to the detection module (Fig. 5(a)). A very important aspect of this pipeline is that a better pedestrian detector tends to produce higher re-ID accuracy, given the same set of re-ID feature. In [77], [128], extensive baselines are implemented on the person re-identification in the wild (PRW), and the large-scale person search (LSPS) datasets, respectively and this argument usually holds. Another interesting topic is whether pedestrian detection helps person re-ID. In [18], [77], detection confidence is integrated in the final re-ID scores. In [128], pedestrian detection and re-ID are jointly considered in a CNN model which resembles faster R-CNN [129], while in [77], the ID-discriminative embedding (IDE) is shown to be superior when fined-tuned on a CNN model pre-trained on the R-CNN model [130] for pedestrian detection. These methods provide initial insights on how weakly labeled detection data helps improve re-ID accuracy.

Nevertheless, in the so-called "end-to-end" systems [18], [77], [128], pedestrian tracking is not mentioned nor have we known any existing works/datasets addressing the influence of tracking on re-ID. This work views it as an "ultimate" goal to integrate detection, tracking, and retrieval into one framework, and evaluate the impact of each module on the

Dataset	LSPS	PRW	CAMPUS	EPFL
#frames	18,184	11,816	214	80
#ID	8,432	932	74	30
#annotated bbox	99,809	34,304	1,519	294
#box per ID	11.8	36.8	20.5	9.8
#gallery box	50-200k	50-200k	1,519	294
#camera	-	6	3	4
Evaluation	CMC&mAP	CMC&mAP	CMC	CMC

TABLE 5: Datasets [18], [77], [128], [133] for end-to-end person re-identification (search).

overall re-ID performance. This survey therefore calls for large-scale datasets that provide bounding box annotations to be used for the three tasks.

4.2 Future Issues

4.2.1 System Performance Evaluation

A proper evaluation methodology is a critical and sometimes tricky topic. Generally there is no single "correct" protocol, especially for the under-explored end-to-end re-ID task. An end-to-end re-ID system departs from most current re-ID studies in dynamic galleries based on the specific detector/tracker used and their parameters. Moreover, it also remains mostly unknown how to evaluate detection/tracking performance in the scenario of person re-ID. As a consequence, this survey raises questions of system evaluation on two aspects.

First, it is critical to use effective evaluation metrics for pedestrian detection and tracking in re-ID. The evaluation protocol should be able to quantify and rank detector/tracker performance in a realistic and unbiased manner and informative of re-ID accuracy. Pedestrian detection, for example, mostly employs the log-average miss rate (MR) which is averaged over the precision range of $[10^{-2}, 10^0]$ FPPI (false positives per image). Some also use average precision (AP) following the routine in PASCAL VOC [134]. Dollár *et al.* [135] argue that using the miss rate against FPPI is preferred to precision recall curves in certain tasks such as automotive applications, since there may be an upper limit on the acceptable FPPI. As opposed to the automotive applications of pedestrian detection, person re-ID aims to find a person which does not necessarily care about the false positive rates. So essentially we can employ both the miss rate and average precision to evaluate pedestrian detection for person re-ID.

An important parameter in the AP/MR computation is the intersect over union (IoU) score. A detected bounding box is considered correct if its IoU score with the ground truth bounding box is larger than a threshold. Typically the threshold is set to 0.5, and yet Zhang *et al.* [136] study the difference between a "perfect single frame detector" and an automatic detector under various IoU scores. The KITTI benchmark [137] requires an IoU of 0.7 for car detection, but 0.5 for pedestrians. For person re-identification, this problem is open to proposals. Some clues about it still exist and if we dive closer to the conclusions drawn in [77], we should be aware of the observation that using a larger IoU score (*e.g.*, 0.7) is a better evaluation criteria than a low IoU (*e.g.*, 0.5). Figure 6 presents the relationship between detection accuracy (AP) and re-ID accuracy (rank-1 or mAP) on the PRW dataset. A linear relation is clearly presented between the two tasks
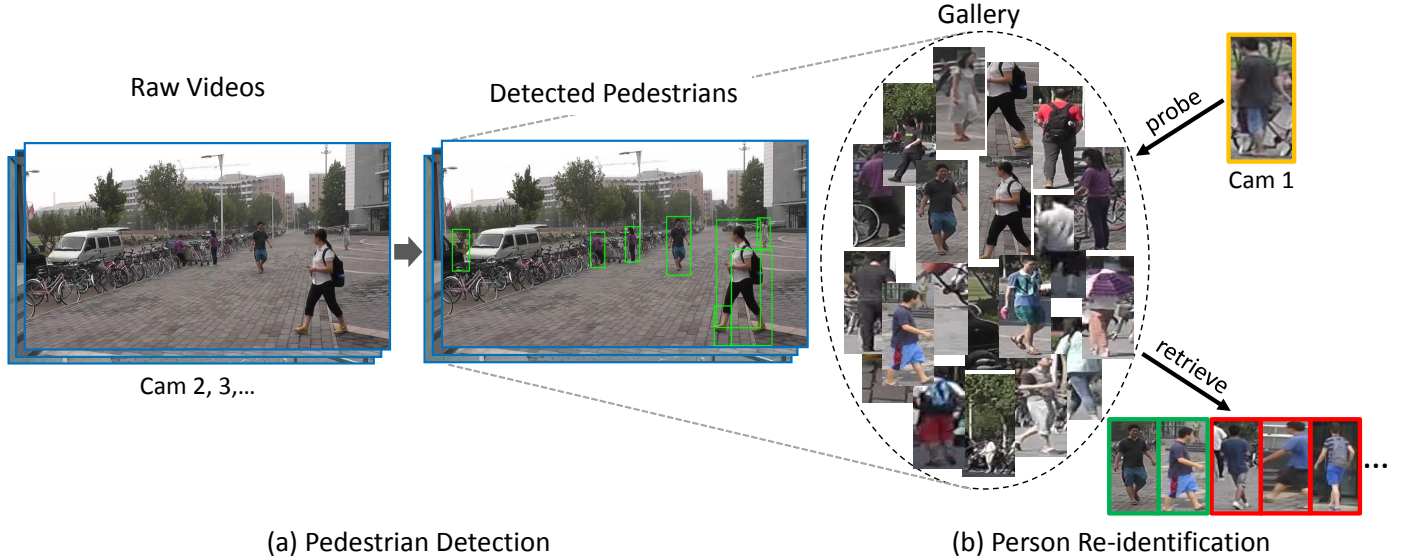
(a) Pedestrian Detection   (b) Person Re-identification

Fig. 5: An end-to-end person re-ID system that includes pedestrian detection and re-identification.

under IoU = 0.7, while a scattered plot exists under IoU = 0.5. The correlation between detectors and recognizors is therefore more consistent with a larger IoU. Nevertheless, it is still far from satisfactory.

Given the consideration that bounding box localization quality is important for re-ID accuracy, it is a good idea to study IoU thresholds when assessing detector quality and see if it accords with re-ID accuracy. Our intuition is that a larger IoU criteria enforces better localization results, but there has to be some limit, because the difference in detector performance tends to diminish when IoU gets larger [136]. It would also feasible to explore the usage of the average recall (AR) proposed in [138] for IoU from 0.5 to 1 and plot the AR for a varying number of detector thresholds. Such an evaluation metric considers both recall and localization, and we speculate that it may be especially informative in re-ID where pedestrian detection recall and bounding box quality are of vital importance.

While there are at least some clues to guide the evaluation of pedestrian detection, how to evaluate tracking under person re-ID is largely unknown. In the multiple object tracking (MOT) benchmark [139], multiple evaluation metrics are used, including multiple object tracking precision (MOTP) [140], mostly track (MT) targets (percentage of ground truth persons whose trajectories are covered by the tracking results for at least 80%), the total number of false positives (FP), the total number of ID switches (IDS), the total number of times a trajectory is fragmented (Frag), the number of frames processed per second (Hz), *etc*. It might be possible that some of the metrics are of limited indication ability such as the processing speed, because tracking is an off-line step. For re-ID, we envision that tracking precision is critical as it is undesirable to have outlier images in the tracklets which compromise the effectiveness of pooling. We also speculate that 80% might not be an optimal threshold for evaluating MT under re-ID. As suggested by [105], extracting features within a walking cycle is a good practise, so generating long tracking sequences may not bring much re-ID improvement. In the future, once datasets are released to evaluate tracking

and re-ID, an urgent problem is thus to design proper metrics to evaluate different trackers.

The second question *w.r.t* the evaluation procedure concerns the re-ID accuracy of the entire system. In contrast to traditional re-ID in which the gallery is fixed, in an end-to-end re-ID system, the gallery varies with the detection/tracking threshold. A stricter threshold indicates higher detection/tracking confidence, so the gallery is smaller and background detections are fewer and vice versa. Furthermore, the gallery size has a direct impact on re-ID accuracy. Let us take an extreme case as an example. When the detection/tracking threshold is very strict, the gallery can be very small, and it is even possible that the ground truth matches are excluded. At the other extreme, when the detection/tracking threshold is set to a very loose value, the gallery would be very large and contain a number of background detections which may exert a negative effect on re-ID, as shown in [44]. Therefore, it is predictable that too strict or too loose a threshold leads to inferior galleries, and it is preferred that the re-ID evaluation protocol reflect how the re-ID accuracy changes with the gallery dynamics. In [77], Zheng *et al.* plot rank-1 accuracy and mAP against a different number of detections per image. It is observed that the curves first rise and then drop after they peak. In the PRW dataset, the peak is positioned at 4-5 detections per images, which can serve as an estimation of the average number of pedestrians per image. In [128], a similar protocol is employed, *i.e.*, the rank-1 matching rate is plotted against detection recall, and reaches its maximum value when recall = 70%. When recall further increases, the prevalence of false detections will compromise the re-ID accuracy. Some other ideas could be explored, *e.g.,* plotting re-ID accuracy against FPPI. Keeping in mind that the gallery size depends on the detector threshold, other new evaluation metrics that are informative and unbiased could be designed in the future.

We also point out another re-ID evaluation protocol in end-to-end systems. In practice, when being presented with a query bounding box/video sequence, while it is good to locate the identity in a certain frame and tell its coordinates
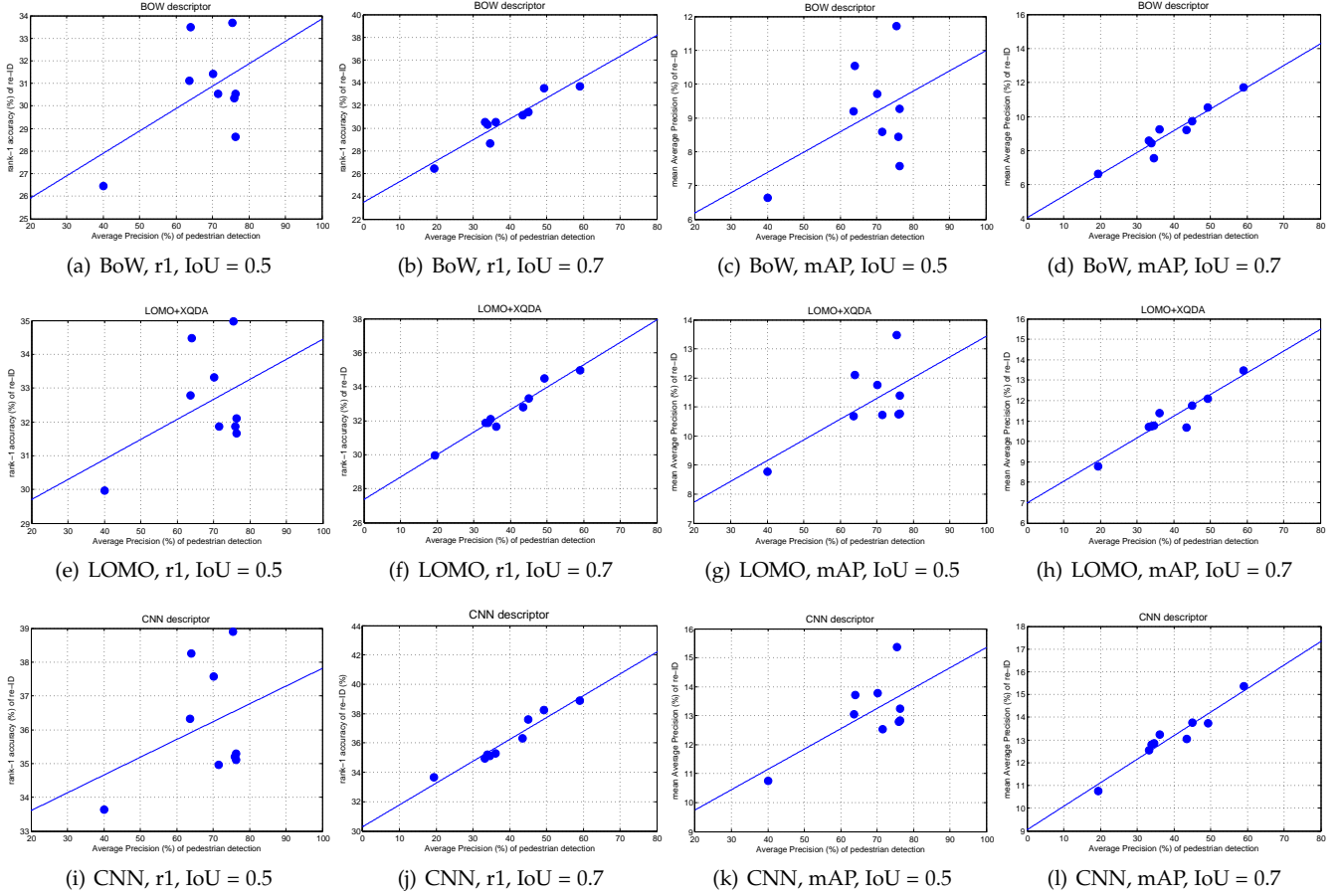
(a) BoW, r1, IoU = 0.5   (b) BoW, r1, IoU = 0.7   (c) BoW, mAP, IoU = 0.5   (d) BoW, mAP, IoU = 0.7

(e) LOMO, r1, IoU = 0.5   (f) LOMO, r1, IoU = 0.7   (g) LOMO, mAP, IoU = 0.5   (h) LOMO, mAP, IoU = 0.7

(i) CNN, r1, IoU = 0.5   (j) CNN, r1, IoU = 0.7   (k) CNN, mAP, IoU = 0.5   (l) CNN, mAP, IoU = 0.7

Fig. 6: Person re-ID accuracy (mAP and rank-1) versus pedestrian detection accuracy (AP) on the PRW dataset [77]. Three re-ID methods are evaluated, *i.e.,* BoW [44], LOMO + XQDA [20], and CNN [77]. 9 detectors are evaluated, *i.e.,* 1) DPM [91] + RCNN (AlexNet), 2) DPM pre-trained on INRIA [131], 3) DPM re-trained on PRW, 4) ACF [92] pre-trained on INRIA, 5) ACF + RCNN (AlexNet), 6) ACF + RCNN (ResidualNet), 7) ACF re-trained on PRW, 8) LDCF [132] re-trained on PRW, and 9) LDCF pre-trained on INRIA. We can observe clearly the linear relation between re-ID and detection accuracy under IoU = 0.7 instead of IoU = 0.5.

by pedestrian detection/tracking, it is also acceptable that the system only knows which frame(s) the identity re-appears in. The specific location of the query person can then be found by human labor which is efficient. In essence, determining the exact frame(s) where the queried person appears is a relatively easier task than a "detection/tracking+re-identification" task, because detection/tracking errors may not exert a large influence. In this scenario, re-ID accuracy should be higher than the standard re-ID task. Also, mean average precision can be used *w.r.t* the retrieved video frames. Since this task does not require locating persons very precisely, we can thus use relaxed detectors/proposals or trackers aiming at improving frame-level recall. Detectors/proposals can be learned to locate a rough region of pedestrians with a loose IoU restriction, and put more emphasis on matching, *i.e.,* finding a particular person from a larger bounding box/spatial-temporal tube.

4.2.2 The Influence of Detector/Tracker on Re-ID

Person re-ID originates from pedestrian tracking [9], in which tracklets from multiple cameras are associated if they are determined to be of the same identity. This line of research

treats re-ID as a part of the tracking system, and does not evaluate the impact of localization/tracking accuracy on re-ID accuracy. However, even since the independence of re-ID, most studies have been conducted on hand-drawn image bounding boxes which is an idealized situation that hardly meets reality. Therefore, in an end-to-end re-ID system, it is critical that the impact of detection/tracking on re-ID be understood and that methods be proposed that detection/tracking methods/data can help re-ID.

First, pedestrian/tracking errors do affect re-ID accuracy, but the intrinsic mechanism and feasible solutions are still open to challenge. Detection errors (Fig. 7) may lead to pedestrian misalignment, scale changes, part missing and most importantly, false positives and miss detections, which compromise the re-ID performance and pose new challenges for the community [16], [44], [96].

A few re-ID works explicitly take the detection/tracking errors into account. In [29], Zheng *et al.* propose fusing local-local and global-local matches to address partial re-ID problems with severe occlusions or missing parts. In [18], Xu *et al.* compute a "commonness" score by matching the GMM encoded descriptor with a prior distribution. The score
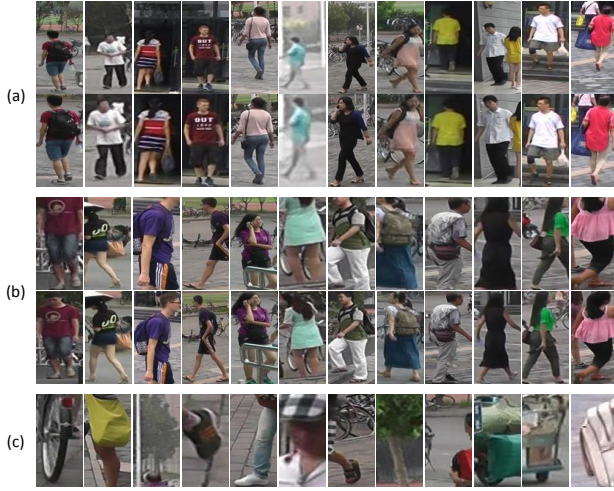
Fig. 7: Detection errors in the Market-1501 dataset [44]. (a) misalignment and scale variances; (b) part missing; (c) false positives. In (a) and (b), the first and second rows represent DPM-detected and hand-drawn bounding boxes which have an IoU > 0.5.

Fig. 8: Tracking errors/artifacts in the MARS dataset [21]. Each row represents a tracklet generated by the DPM detection + GMMCP tracker [125]. First row: detection error and tracking error; second row: detection error; third row: occlusions in tracking; last row: tracking error.

can be used to eliminate false positives which do not contain or provide good localization of a human body. In a similar way, Zheng *et al.* [77] propose integrating detector confidence (after square root) into the re-ID similarity score, according to which the gallery bounding boxes are ranked. These works address detection errors after they happen. Nevertheless, there is a possibility that detection/tracking errors could be avoided at an earlier stage. For example, in the network designed by Xiao *et al.* [128], a localization loss is added in the fast R-CNN [141] sub-module. It regulates localization quality which is critical for an effective re-ID system.

Future investigations are in need to reveal the dependence of person re-ID on detection/tracking quality. Since the idea to develop detector/trackers that are error-free is too idealistic, we advocate research into how detection confidence can be integrated into re-ID matching scores, *i.e.,* how to correct errors by effectively identifying outliers, and how to train context models that do not rely solely on detected bounding boxes. For example, using clustering algorithms to filter out inconsistent frames within a tracklet can be effective in purifying tracking sequences. In another example, detected bounding boxes could be enlarged to include possibly missing body parts and learn discriminative visual features that in turn use or discard the enriched contextual information.

Secondly, we should be aware that detection and tracking, if appropriately designed, may be of help to re-ID. In [77], the IDE network fine-tuned on the R-CNN model [64] is proved to be more effective than the one fine-tuned directly on an ImageNet pre-trained model. This illustrates the importance of using the excessive amount of labeled data in pedestrian detection, *i.e.,* pedestrians with ID annotations and false positive detections. In [128], the end-to-end network integrates the loss of background detections, which is assumed to improve the discriminative ability of the learned embedding. The integration of detection scores into re-ID similarities [18], [77] can also be viewed as an alternative that detection helps re-ID.

It may seem not quite straightforward that pedestrian detection/tracking could help re-ID or the reverse, but if we consider the analogy of generic image classification and fine-grained classification, we may think of some clues. From example, fine-tuning the ImageNet pre-trained CNN model on the fine-grained datasets is an effective way for faster convergence and higher fine-grained recognition accuracy. It is also a good idea to jointly train a pedestrian detection and re-ID model by back-propagating the re-ID loss to the (fast) RCNN part. Being able to discriminate different identities may be beneficial to the task of discriminating pedestrians from the background. The latter could also be helpful to the former.

One of the ideas that can be explored is the use of unsupervised tracking data. In videos, tracking a pedestrian is not too difficult a task, though tracking errors are inevitable. Facial recognition, color, and non-background information are useful tools to improve tracking performance like in Harry Potter's Marauder's Map [142]. Within a tracking sequence, the appearance of a person may undergo variances to some extent, but it can be expected that most of the bounding boxes are of the same person. In this scenario, each tracklet represents a person which contains a number of noisy but roughly usable training samples. We can therefore make use of racking results to train pedestrian verification/identification models, so as to alleviate the reliance on large-scale supervised data. As another promising idea, it is worth trying to pre-train CNN models using the detection/tracking data using the auto-encoder [143] or the generative adversarial nets (GAN) [144]. It would also be interesting to directly learn person descriptors using such unsupervised networks to help address the data issue in person re-ID.

5 FUTURE: PERSON RE-ID IN VERY LARGE GALLERIES

The scale of data has increased significantly in the re-ID community in recent years, *e.g.,* from several hundred gallery images in VIPeR [82] and iLIDS [83] to over 100k as in PRW [77] and LSPS [128], which gives rise to the predominance of

visual words

local
descriptor

$f \longrightarrow q(f) \longrightarrow$
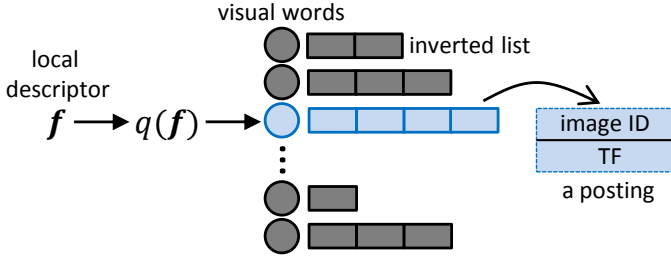
inverted list

image ID

TF

a posting

Fig. 9: An illustration of the inverted index.

deep learning methods. However, it is apparent that current datasets are still far from a practical scale. Supposing that in a region-scale surveillance network with 100 cameras, if one video frame is used per second for pedestrian detection, and an average of 10 bounding boxes are produced from each frame, then, running the system for 12 hours will produce $3,600 \times 12 \times 1 \times 10 \times 100 = 43.2 \times 10^6$ bounding boxes. But to our knowledge, previously no work has reported re-ID performance in such a large gallery. It seems that the largest gallery used in the literature is 500k [44], and evidence suggests that mAP drops over 7% compared to Market-1501 with a 19k gallery. Moreover, in [44], approximate nearest neighbor search [145] is employed for fast retrieval but at the cost of compromised accuracy.

From both a research and an application perspective, person re-ID in very large galleries should be a critical direction in the future. Attempts to improve both the accuracy and efficiency issues should be made.

On the one hand, robust and large-scale learning of descriptors and distance metrics is much more important. This coincides with current research [71], [73], [75], [81]. Following large-scale image recognition [78], person re-ID will progress to large-scale evaluations. Although current methods address the re-ID problem between one or several pairs of cameras in a very limited time window, robustness in a camera networks over a long time period has not been well considered. In [36], [146], the re-ID consistency within a camera network is jointly optimized with pair-wise matching accuracy, but the testing datasets (WARD [87] and RAiD [36]) only have 3 and 4 cameras and less than 100 identities. In a network with n cameras, the number of camera pairs is $\mathcal{O}(n^2)$. Considering the long recording time and lack of annotated data, it is typically prohibitive to train distance metrics or CNN descriptors in a pair-wise manner. As a consequence, training a global re-ID model with adaptation to various illumination condition and camera location is a priority. Toward this goal, an option is to design unsupervised descriptors [44], [97] which aim to find visually similar persons and treat visually dissimilar ones as false matches. But unsupervised methods may be prone to lighting changes.

On the other hand, efficiency is another important issue in such a large-scale setting. Although computation time could almost be omitted in small datasets [82], [83], in our experiment using MATLAB 2014 on a server with 3.1GHz Intel Xeon E5-2687w v3 (10 cores), 64GB memory, it takes 8.50s to compute the distance between a 100-dim floating vector with a number of 10 million 100-dim vectors. If we use a 4,096-dim floating-point vector extracted from the

CaffeNet [14] and C++ programming, the time used increases dramatically to 60.7s including 33.2s for the distance calculation and 26.8s for the data to load from the disk. It is clear that the query time increases dramatically according to the feature dimensions and gallery size, which is not desirable for practical use. To our knowledge, previous works in person re-ID rarely focus on efficiency issues, and therefore effective solutions are lacking, but fortunately, we can resort to the image retrieval community for answers, and this survey provides two possible directions.

Inverted index-based. The inverted index is a *de facto* data structure in the Bag-of-Words (BoW) based retrieval methods [22], [147], [148]. Based on the quantization result of local descriptors, the inverted index has k entries where k denotes the codebook size. The indexing structure thus has k entries and each is attached to an inverted list, in which the local descriptors are indexed. The structure of the baseline inverted index is shown in Fig. 9. A posting stores the image ID and the term frequency (TF) of the indexed descriptor and in a series of works, a number of other meta data can be stored, such as binary signature [148], feature coordinates [149], *etc*. For basic knowledge and state-of-the-art advances of the inverted index in instance retrieval, we refer readers to a recent survey [19].

In person re-ID, the use of local descriptors is popular [30], [34], [44]. The color and texture features are typically extracted from local patches. While some previous works use sophisticated matching algorithms [30], it is preferred that the matching procedure can be accelerated using the inverted index under a large gallery. A codebook is usually needed to quantize a local descriptor to visual words, and since the local descriptors are high-dimensional, a large codebook is needed to reduce quantization error. Under these circumstances, the inverted index is ready for use which saves memory costs to a large extent and, if properly employed, can have approximately the same accuracy compared to quantization-free cases.

Hashing-based. Hashing has been an extensively studied solution to approximate nearest neighbor search, which aims to reduce the cost of finding exact nearest neighbors when the gallery is large or distance computation is costly [23]. Learning to hash is popular in the community following the milestone work Spectral Hashing [150]. It refers to learning hash functions, $y = h(x)$, mapping a vector x to a compact y, and aims at finding the true nearest neighbor at high-ranks in the rank list while keeping the efficiency of the search process. Some classic learning to hash methods include product quantization (PQ) [117], iterative quantization (ITQ) [151], *etc*. Both methods are efficient in training and have fair retrieval accuracy. They do not require labeled data, so are applicable for re-ID tasks when large amount of training data may not be available.

Another application of supervised hashing consists of image retrieval [152], [153], [154], [155], which is the interest of this section. The hash function is learned end-to-end through a deep learning network which outputs a binary vector given an input image. This line of works focus on several image classification datasets such as CIFAR-10 [156] and NUS-WIDE [157], in order to leverage the training data that is lacking in generic instance retrieval datasets [22], [148]. In person re-ID, the application scenario fits well

with deep hashing for image retrieval. In large galleries, efficient yet accurate hash methods are greatly needed, which is a less-explored direction in re-ID. As shown in Table 1, training classes are available in re-ID datasets, and the testing procedure is a standard retrieval task, so the current arts in supervised hashing are readily to be adopted in re-ID in the light of the increasing size of the datasets [16], [44]. The only relevant work we find is [158] which learns hash functions in a triplet-loss CNN network with regularizations to enforce adjacency consistency. This method is tested on the CUHK03 dataset which contains 100 identities in each test split, so in this sense, performance evaluation on very large galleries is still lacking. As a consequence, this survey calls for very large re-ID datasets that will evaluate the scalability of re-ID methods and scalable algorithms especially those using hash codes to further push this task to real-world applications.

6 OTHER IMPORTANT YET UNDER-DEVELOPED OPEN ISSUES

6.1 Battle Against Data Volumn

Annotating large-scale datasets has always been a focus in the vision community. This problem is even more challenging in person re-ID, because apart from drawing a bounding box of a pedestrian, one has to assign him an ID. ID assignment is not trivial since a pedestrian may re-enter the fields of view (FOV) or enter another observation camera a long time after the pedestrian's first appearance. This makes collaborative annotation difficult, as it is costly for two collaborators to communicate on the annotated IDs. These difficulties partially explain why current datasets typically have a very limited number of images for each ID. The last two years have witnessed the release of several large-scale datasets, *e.g.*, Market-1501 [44], PRW [77], LSPS [128], and MARS [21], but they are still far from satisfaction in views of practical applications. In this survey, we believe two alternative strategies can help bypass the data issue.

First, how to use annotations from tracking and detection datasets remains under-explored. Compared to re-ID, tracking and detection annotations do not require ID assignment when a person re-enters FOV: the majority of effort has been spent on bounding box drawing. In [77], it is shown that adding more pedestrian and background training data in the R-CNN stage benefits the following training of the IDE descriptor. In [50], [75], attribute annotations from independent datasets are employed to represent the re-ID images. Since the attributes can be annotated through collaboration among workers and have good generalization ability, they are also good alternatives to the deficiency of re-ID data. As a consequence, external resources are valuable for training re-ID systems when training data is lacking.

Apart from the pre-training/unsupervised strategies as mentioned in Section 4.2.2, a novel solution is to retrieve hard negatives from the unlabeled data which can be viewed as "true positives" in metric learning/CNN training. This strategy has been evaluated in object classification where a small portion of labels are disturbed before training [159]. It can efficiently enlarge the training set, and at the same time reduce the risk of model over-fitting. Our preliminary experiments show that this direction yields decent improvement over the baselines.
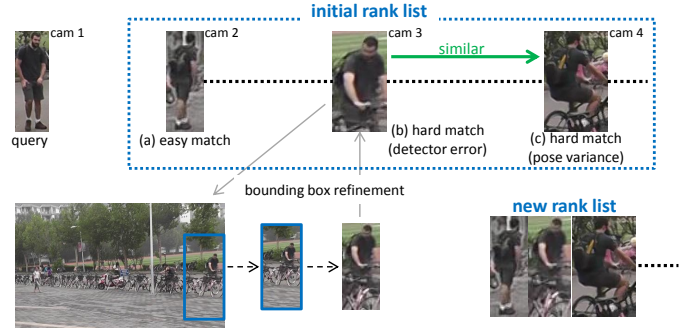

Fig. 10: An example of re-ranking in re-ID. Given a query image, an initial rank list is obtained, in which an easy match (a) is ranked top, while two hard matches (b) and (c) have low ranks. The detection error in (b) can be corrected by retrieving the corresponding video frame and performing a finer search for the best bounding box within a local neighborhood. In this example, (c) is visually similar to (b) but not the query, so after (b) is retrieved, (c) can be found by similarity propagation.

The second strategy is transfer learning that transfers a trained model from the source to the target domain. Previously, supervised learning require large numbers of labeled data which limits the re-ID system to scale to other cameras. In [160], an unsupervised topic model is proposed to discover saliant image patches for re-ID matching and simultaneously remove background clusters. In [161], a weakly supervised method is proposed which requires full annotations from other re-ID dataset and a few samples captured in the target scenario. In [162], [163], unsupervised transfer learning is proposed in which the target dataset is unlabeled. Ma *et al.* [162] employ a cross-domain ranking SVM, while Peng *et al.* [163] formulate the transfer problem as a dictionary learning task, which learns the shared invariant latent variables and is biased towards the target dataset. These methods indicate that it is feasible to learn a fair re-ID model from the source, and that it is beneficial to mine discriminative cues from the unsupervised data. Transfering CNN models to other re-ID datasets can be more difficult because the deep model provides a good fit to the source. Xiao *et al.* [76] gather a number of source re-ID datasets and jointly train a recognition model for the target dataset. According to our experience, the usage of off-the-shelf metric learning methods [20], [55] can also correct the transfer effect to some extent, but unsupervised transfer learning is still an open issue for the deeply learned models.

6.2 Re-ranking Re-ID Results

The re-identification process (Fig. 5(b)) can be viewed as a retrieval task, in which re-ranking is an important step to improve the retrieval accuracy. It refers to the re-ordering of the initial ranking result from which re-ranking knowledge can be discovered. For a detailed survey of search re-ranking methods, we refer the readers to [164].

A few previous works exist on this topic. Re-ranking can be performed either with human in the loop or fully automatically. When online human labeling is involved, Liu *et al.* [165] propose the post-rank optimisation (POP)

method which allows a user to provide an easy negative and, optionally, a few hard negatives from the initial rank list. The sparse human feedback enables on-the-fly automatic discriminative feature selection of the query person. In an improvement, Wang *et al.* [96] design the human verification incremental learning (HVIL) model which does not require any pre-labelled training data and learns cumulatively from human feedback to provide instance model update. A number of incrementally learned HVIL models are combined into a single ensemble model for use when human feedback is no longer available. In a similar nature, Martinel *et al.* [166] propose finding the most relevant gallery images for a query, sending them to the human labeler, and finally using the labels to update the re-ID model. Automatic re-ranking methods have also been studied in several works. Zheng *et al.* [167] propose a query-adaptive fusion method to combine rank results of several re-ID systems. Specifically, the shape of the initial score curves is used and it is argued that the curve exhibits an "L" shape for a good feature. In [95], various metrics are ensembled based on the direct optimization of the CMC curve. García *et al.* [94] analyze the unsupervised discriminant context information in the rank list. This is further combined with a re-ranking metric learned in the offline. Leng *et al.* [168] use the idea of reciprocal k-nearest neighbors [169] to refine the initial rank list based constructing images relations in the offline steps.

Re-ranking is still an open direction in person re-ID, while it has been extensively studied in instance retrieval. The application scenario can be depicted as follows. When searching for a person-of-interest, it is likely that its images under certain cameras are very difficult to find due to intensive image variations. But we may be able to find the true matches under some cameras which are more similar to the hard positives. So in this manner, hard positives can be found once the easy ones are returned. Re-ranking methods in instance retrieval can be readily adopted in person re-ID [44], [169], [170], [171]. Since training data is available in re-ID (Table 1), it is possible to design re-ranking methods based on training distribution. For example, when doing k-NN re-ranking [170], the validity of the returned results can be determined from the training set according to the scores. Since re-ID is focused on pedestrians, re-ranking methods can be specifically designed. For example, after obtaining the initial rank list, a subset of the top-ranked images can be selected, and the video frames containing them can be retrieved. We can subsequently find the best localization through expensive sliding window method without incurring much computation burdens, so as to allieviate the impact of detector misalignment.

6.3 Open-World Person Re-ID

Most existing re-ID works can be viewed as an identification task (Eq. 1). Query identities are assumed to exist in the gallery and the tasks aim to determine the ID of the query. By contrast, open-world re-ID systems study the person verification problem. That is, based on identification tasks, the open-world problem adds another condition to Eq. 1,

$$\text{sim}(q, g_{i^*}) > h, \qquad (4)$$

where h is the threshold above which we can assert that query q belongs to identity i^*; otherwise, q is determined

an outlier identity which is not contained in the gallery, although i^* is the first ranked identity in the identification process.

In literature, open-world person re-ID is still at its early stage, and several works are proposed to help define this task. In [172], Zheng *et al.* design a system consisting of a watch list (gallery) of several known identities and a number of probes including target and non-target ones. Their work aims to achieve high true target recognition (TTR) and low false target recognition (FTR) rate which calculate rate of the number of queries that are verified as the target identities to the total number of queries. In [173], Liao *et al.* divide open-world re-ID into two sub-tasks, *i.e.,* detection and identification; the former decides whether a probe identity is present in the gallery and the latter assigns an ID to the accepted probe. Consequently two different evaluation metrics, the detection and identification rate (DIR) and the false accept rate (FAR) are proposed, based on which a receiver operating characteristic (ROC) curve can be drawn.

Open-world re-ID still remains a challenging task as evidenced by the low recognition rate under low false accept rate, as shown in [172], [173]. The challenge mainly lies in two aspects *i.e.,* detection and recognition, both of which are limited to the unsatisfying matching accuracy - a research focus in standard re-ID tasks. As indicated in [173], a 100% FAR corresponds to the standard close-set re-ID and its accuracy is limited by the current state of the art; a lower FAR is accompanied by lower re-ID accuracy due to the low recall of the true matches. As a consequence, from a technical perspective, the critical goal is to improve matching accuracy, based on which probabilistic models can be designed for novelty detection (verification) methods. Moreover, when focusing on re-ID accuracy, open-world re-ID should also consider the dynamics of the gallery [174]. In a dynamic system with constantly incoming bounding boxes, a new identity will be added to the "watch list" if it is determined to not belong to any existing gallery identities, and vice versa. Enrolling new identities dynamically enables automatic database construction and facilitates the re-ID process with a pre-organized gallery.

7 CONCLUDING REMARKS

Person re-identification, foretold in the oldest stories, is gaining extensive interest in the modern scientific community. In this paper, a survey of person re-identification is presented. First, a brief history of person re-ID is introduced and its similarities and differences to image classification and instance retrieval are described. Then, existing image and video-based methods are reviewed, which are categorized into hand-crafted and deeply-learned systems. Positioned inbetween image classification and instance retrieval, person re-ID has a long way from becoming an accurate and efficient application. Therefore, departing from previous surveys, this paper places more emphasis on the under-developed but critical future possibilities, such as the end-to-end re-ID systems that integrate pedestrian detection and tracking, and person re-ID in very large galleries, which we believe are necessary steps toward practical systems. We also highlight some important open issues that may attract further attention from the community. They include solving the data volume

issue, re-ID re-ranking methods, and open re-ID systems. All in all, the integration of discriminative feature learning, detector/tracking optimization, and efficient data structures will lead to a successful person re-identification system.

ACKNOWLEDGMENTS

REFERENCES

[1] A. Plantinga, "Things and persons," *The Review of Metaphysics*, pp. 493–519, 1961.

[2] A. O. Rorty, "The transformations of persons," *Philosophy*, vol. 48, no. 185, pp. 261–275, 1973.

[3] N. B. Cocchiarella, "Sortals, natural kinds and re-identification," *Logique et analyse*, vol. 80, pp. 439–474, 1977.

[4] T. D'Orazio and G. Cicirelli, "People re-identification and tracking from multiple cameras: a review," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 1601–1604.

[5] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[6] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. Springer, 2014, vol. 1.

[7] R. Satta, "Appearance descriptors for person re-identification: a comprehensive review," *arXiv preprint arXiv:1307.5748*, 2013.

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

[9] T. Huang and S. Russell, "Object identification in a bayesian context," in *IJCAI*, vol. 97, 1997, pp. 1276–1282.

[10] W. Zajdel, Z. Zivkovic, and B. Krose, "Keeping track of humans: Have i seen this person before?" in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2005, pp. 2081–2086.

[11] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1528–1535.

[12] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 1413–1416.

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

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[15] D. Yi, Z. Lei, S. Liao, S. Z. Li *et al.*, "Deep metric learning for person re-identification." in *ICPR*, vol. 2014, 2014, pp. 34–39.

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

[17] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[18] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 937–940.

[19] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: A decade survey of instance retrieval," *arXiv preprint arXiv:1608.01807*, 2016.

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

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

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[23] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *arXiv:1606.00185*, 2016.

[24] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[25] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *BMVC*, vol. 2, no. 5, 2010, p. 6.

[26] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[27] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3567–3574.

[28] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2666–2672.

[29] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4678–4686.

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

[31] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. 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.

[32] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.

[33] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2528–2535.

[34] ——, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 144–151.

[35] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3200–3208.

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

[37] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, "Hierarchical gaussianization for image classification," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 1971–1977.

[38] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1565–1573.

[39] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325.

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

[41] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European Conference on Computer Vision*. Springer, 2014, pp. 536–551.

[42] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[43] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification."

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

[45] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1512–1523, 2009.

[46] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1363–1372.

[47] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes." in *BMVC*, vol. 2, no. 3, 2012, p. 8.

[48] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu, "Attribute-restricted latent topic model for person re-identification," *Pattern recognition*, vol. 45, no. 12, pp. 4204–4213, 2012.

[49] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *European Conference on Computer Vision Workshops*. Springer, 2012, pp. 391–401.

[50] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3739–3747.

[51] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4184–4193.

[52] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016.

[53] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, p. 78, 2006.

[54] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems*, vol. 15, pp. 505–512, 2003.

[55] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[56] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, 2005, pp. 1473–1480.

[57] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 209–216.

[58] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *European Conference on Computer Vision*. Springer, 2012, pp. 780–793.

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

[60] Y. Yang, S. Liao, Z. Lei, and S. Z. Li, "Large scale similarity learning using similar pairs for person verification," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[61] B. Scholkopft and K.-R. Mullert, "Fisher discriminant analysis with kernels," *Neural networks for signal processing IX*, vol. 1, no. 1, p. 1, 1999.

[62] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision*. Springer, 2014, pp. 1–16.

[63] X. Liu, H. Wang, Y. Wu, J. Yang, and M.-H. Yang, "An ensemble color model for human re-identification," in *IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 868–875.

[64] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[65] F. Radenović, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," *arXiv:1604.02426*, 2016.

[66] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

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

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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

[70] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[71] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *European Conference on Computer Vision*, 2016.

[72] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*, 2016.

[73] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *arXiv preprint arXiv:1606.04404*, 2016.

[74] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[75] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European Conference on Computer Vision*, 2016.

[76] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[77] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," *arXiv preprint arXiv:1604.02531*, 2016.

[78] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[79] L. Wu, C. Shen, and A. v. d. Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *arXiv preprint arXiv:1606.01595*, 2016.

[80] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010, pp. 143–156.

[81] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.

[82] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3, no. 5. Citeseer, 2007.

[83] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proceedings of the British Machine Vision Conference*, 2009, pp. 23.1–23.11.

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

[85] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference*, 2011.

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

[87] N. Martinel and C. Micheloni, "Re-identify people in wide area camera network," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 31–36.

[88] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision*, 2012, pp. 31–44.

[89] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3594–3601.

[90] P. M. Roth, M. Hirzer, M. Koestinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, ser. Advances in Computer Vision and Pattern Recognition, S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds. London, United Kingdom: Springer, 2014, pp. 247–267.

[91] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[92] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

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

[94] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in *ICCV*, 2015.

[95] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[96] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *European Conference on Computer Vision*, 2016.

[97] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *British Machive Vision Conference*, 2012, p. 11.

[98] ——, "Local descriptors encoded by fisher vectors for person re-identification," in *European Conference on Computer Vision*. Springer, 2012, pp. 413–422.

[99] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *ACM/IEEE International Conference on Distributed Smart Cameras*, 2008, pp. 1–6.

[100] D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi, "Video sequences association for people re-identification across multiple non-overlapping cameras," in *International Conference on Image Analysis and Processing*. Springer, 2009, pp. 179–189.

[101] S. Karaman and A. D. Bagdanov, "Identity inference: generalizing person re-identification scenarios," in *European Conference on Computer Vision*. Springer, 2012, pp. 443–452.

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

[103] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 33–40.

[104] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1354–1362.

[105] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision*, 2014, pp. 688–703.

[106] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[107] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[108] K. Liu, B. Ma, W. Zhang, and R. 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.

[109] C. Gao, J. Wang, L. Liu, J.-G. Yu, and N. Sang, "Temporally aligned pooling representation for video-based person re-identification," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 4284–4288.

[110] Z. Liu, J. Chen, and Y. Wang, "A fast adaptive spatio-temporal 3d feature for video-based person re-identification," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 4294–4298.

[111] W.-S. Zheng, S. Gong, and T. Xiang, "Transfer re-identification: From person to set-based verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2650–2657.

[112] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *IJCAI*, 2016.

[113] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[114] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[115] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *European Conference on Computer Vision*, 2016.

[116] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative cnn video representation for event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1798–1807.

[117] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[118] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[119] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal pyramid pooling based convolutional neural networks for action recognition," *arXiv preprint arXiv:1503.01224*, 2015.

[120] L. Wu, C. Shen, and A. van den Hengel, "Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach," *arXiv:1606.01595*, 2016.

[121] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 461–470.

[122] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *IEEE International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[123] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. ACM, 2011, pp. 59–64.

[124] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1629–1642, 2015.

[125] A. Dehghan, S. Modiri Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4091–4099.

[126] N. Martinel, A. Das, C. Micheloni, and A. 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.

[127] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[128] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "End-to-end deep learning for person search," *arXiv preprint arXiv:1604.01850*, 2016.

[129] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[130] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[131] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[132] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[133] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.

[134] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[135] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[136] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[137] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[138] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 814–830, 2016.

[139] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv preprint arXiv:1504.01942*, 2015.

[140] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[141] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[142] S.-I. Yu, Y. Yang, and A. Hauptmann, "Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3714–3720.

[143] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder." in *Interspeech*. Citeseer, 2010, pp. 1692–1695.

[144] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[145] J. Wang and S. Li, "Query-driven iterated neighborhood graph search for large scale indexing," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 179–188.

[146] A. Chakraborty, A. Das, and A. Roy-Chowdhury, "Network consistent data association," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[147] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.

[148] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European conference on computer vision*, 2008, pp. 304–317.

[149] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 809–816.

[150] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in neural information processing systems*, 2009, pp. 1753–1760.

[151] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 817–824.

[152] X. Liu, X. Fan, C. Deng, Z. Li, H. Su, and D. Tao, "Multilinear hyperplane hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5119–5127.

[153] Z. Zhang, Y. Chen, and V. Saligrama, "Efficient training of very deep neural networks for supervised hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1487–1495.

[154] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1556–1564.

[155] V. Erin Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2475–2483.

[156] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[157] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, p. 48.

[158] 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.

[159] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian, "Disturblabel: Regularizing cnn on the loss layer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[160] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," 2014.

[161] X. Wang, W.-S. Zheng, X. Li, and J. Zhang, "Cross-scenario transfer person re-identification," 2015.

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

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

[164] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Computing Surveys*, vol. 46, no. 3, p. 38, 2014.

[165] C. Liu, C. Change Loy, S. Gong, and G. Wang, "Pop: Person re-identification post-rank optimisation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 441–448.

[166] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *European Conference on Computer Vision*, 2016.

[167] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1741–1750.

[168] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen, "Person re-identification with content and context re-ranking," *Multimedia Tools and Applications*, vol. 74, no. 17, pp. 6989–7014, 2015.

[169] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 777–784.

[170] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3013–3020.

[171] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2911–2918.

[172] 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. 591–606, 2016.

[173] S. Liao, Z. Mo, J. Zhu, Y. Hu, and S. Z. Li, "Open-set person re-identification," *arXiv preprint arXiv:1408.0872*, 2014.

[174] B. DeCann and A. Ross, "Modelling errors in a biometric re-identification system," *IET Biometrics*, vol. 4, no. 4, pp. 209–219, 2015.