A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets

Srikrishna Karanam*, *Student Member, IEEE,* Mengran Gou*, *Student Member, IEEE,*
Ziyan Wu, *Member, IEEE,* Angels Rates-Borras, Octavia Camps, *Member, IEEE,*
and Richard J. Radke, *Senior Member, IEEE*

Abstract—Person re-identification (re-id) is a critical problem in video analytics applications such as security and surveillance. The public release of several datasets and code for vision algorithms has facilitated rapid progress in this area over the last few years. However, directly comparing re-id algorithms reported in the literature has become difficult since a wide variety of features, experimental protocols, and evaluation metrics are employed. In order to address this need, we present an extensive review and performance evaluation of single- and multi-shot re-id algorithms. The experimental protocol incorporates the most recent advances in both feature extraction and metric learning. To ensure a fair comparison, all of the approaches were implemented using a unified code library that includes 11 feature extraction algorithms and 22 metric learning and ranking techniques. All approaches were evaluated using a new large-scale dataset that closely mimics a real-world problem setting, in addition to 16 other publicly available datasets: VIPeR, GRID, CAVIAR, DukeMTMC4ReID, 3DPeS, PRID, V47, WARD, SAIVT-SoftBio, CUHK01, CHUK02, CUHK03, RAiD, iLIDSVID, HDA+, and Market1501. The evaluation codebase and results will be made publicly available for community use.

Index Terms—Person Re-Identification, Camera Network, Video Analytics, Benchmark

1 Introduction

Person re-identification, or re-id, is a critical task in most surveillance and security applications [1], [2], [3] and has increasingly attracted attention from the computer vision community [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. The fundamental re-id problem is to compare a person of interest as seen in a "probe" camera view to a "gallery" of candidates captured from a camera that does not overlap with the probe one. If a true match to the probe exists in the gallery, it should have a high matching score, or rank, compared to incorrect candidates.

Since the body of research in re-id is now quite large, we can begin to draw conclusions about the best combinations of algorithmic subcomponents. In this paper, we present a careful, fair, and systematic evaluation of feature extraction, metric learning, and multi-shot ranking algorithms proposed for re-id on a wide variety of benchmark datasets. Our general evaluation framework is to consider all possible combinations of feature extraction and metric learning algorithms for single-shot datasets and all possible combinations of feature extraction, metric learning, and multi-shot ranking algorithms for multi-shot datasets. In particular, we evaluate 276 such algorithm combinations on 10 single-shot re-id datasets and 646 such algorithm combinations on 7 multi-shot re-id datasets, making the proposed study the **largest and most systematic** re-id benchmark to date. As part of the evaluation, we built a **public code library** with an easy-to-use input/output code structure and uniform algorithm parameters that includes 11 contemporary feature extraction and 22 metric learning and ranking algorithms. Both the code library and the complete benchmark results are publicly available for community use at https://github.com/RSL-NEU/person-reid-benchmark.

Existing re-id algorithms are typically evaluated on academic re-id datasets [4], [24], [25], [26], [27], [28], [29], [30] that are specifically hand-curated to only have sets of bounding boxes for the probes and the corresponding matching candidates. On the other hand, real-world end-to-end surveillance systems include automatic detection and tracking modules, depicted in Figure 1, that generate candidates on-the-fly, resulting in gallery sets that are dynamic in nature. Furthermore, errors in these modules may result in bounding boxes that may not accurately represent a human [3]. While these issues are critical in practical re-id

- *R.J. Radke is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, 12180 (e-mail: rjradke@ecse.rpi.edu).*
- *S. Karanam and Z. Wu are with Siemens Corporate Technology, Princeton, NJ 08540 (e-mail: srikrishna.karanam@siemens.com, ziyan.wu@siemens.com).*
- *M. Gou, A. Rates-Borras, and O. Camps are with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, 02115 (e-mail: mengran@coe.neu.edu, ratesborras.a@husky.neu.edu, camps@coe.neu.edu).*
- **S. Karanam and M. Gou contributed equally to this work. Corresponding author: S. Karanam.*

1. This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. Thanks to Michael Young, Jim Spriggs, and Don Kemer for supplying the airport video data.

applications, they are not well-represented in the currently available datasets. To this end, our evaluation also includes a **new, large-scale dataset** constructed from images captured in a challenging surveillance camera network from an airport. All the images in this dataset were generated by running a prototype end-to-end real-time re-id system using automatic person detection and tracking algorithms instead of hand-curated bounding boxes.

2 Evaluated Techniques

In this section, we summarize the feature extraction, metric learning, and multi-shot ranking techniques that are evaluated as part of the proposed re-id benchmark, which include algorithms published through ECCV 2016. We anticipate that the benchmark will be updated (along the lines of the Middlebury benchmarks [31], [32]) as new techniques are implemented into our evaluation framework.

2.1 Feature extraction

We consider 11 feature extraction schemes that are commonly used in the re-id literature, summarized in Table 1(a). In ELF [4], color histograms in the RGB, YCbCr, and HS color spaces, and texture histograms of responses of rotationally invariant Schmid [33] and Gabor [34] filters are computed. In LDFV [35], local pixel descriptors comprising pixel spatial location, intensity, and gradient information are encoded into the Fisher vector [36] representation. In gBiCov [37], multi-scale biologically-inspired features [38] are encoded using covariance descriptors [39]. In IDE-CaffeNet, IDE-ResNet, and IDE-VGGNet, we use the idea first presented in the DeepFace paper [40] and applied to re-id by Zheng *et al.* [41], in which every person is treated as a separate class and a convolutional neural network is trained for a classification objective. AlexNet [42], ResNet [43], and VGGNet [44] architectures are employed in IDE-CaffeNet, IDE-ResNet and IDE-VGGNet respectively. In each case, we start with a model pre-trained on the ImageNet dataset, and finetune it using the datasets we consider in this evaluation. Specifically, during finetuning, we modify the weights of all the layers of the network. More implementation details are presented in Section 3.2. In DenseColorSIFT [9], each image is densely divided into patches, and color histograms and SIFT features are extracted from each patch. In HistLBP [13], color histograms in the RGB, YCbCr, and HS color spaces and texture histograms from local binary patterns (LBP) [45] features are computed. In LOMO [18], HSV color histograms and scale-invariant LBP [46] features are extracted from the image processed by a multi-scale Retinex algorithm [47], and maximally-pooled along the same horizontal strip. In GOG [48], an image is divided into horizontal strips and local patches in each strip are modeled using a Gaussian distribution. Each strip is then regarded as a set of such Gaussian distributions, which is then summarized using a single Gaussian distribution.

2.2 Metric learning

While using any of the features described in the previous section in combination with the Euclidean distance (l_2) can be used to rank gallery candidates, this would be an unsupervised and suboptimal approach. Incorporating supervision using training data leads to superior performance, which is the goal of metric learning, i.e., learning a new feature space such that feature vectors of the same person are close whereas those of different people are relatively far. We consider 18 metric learning methods that are typically used by the re-id community, summarized in Table 1(b). Fisher discriminant analysis (FDA) [51], local Fisher discriminant analysis (LFDA) [11], marginal Fisher analysis (MFA) [54], cross-view quadratic discriminant analysis (XQDA) [18], and discriminative null space learning (NFST) [57] all formulate a Fisher-type optimization problem that seeks to minimize the within-class data scatter while maximizing between-class data scatter. In practice, scatter matrices are regularized by a small fraction of their trace to deal with matrix singularities. Information-theoretic metric learning (ITML) [52], large-margin nearest neighbor (LMNN) [55], relative distance comparison (PRDC) [6], keep-it-simple-and-straightforward metric (KISSME) [7], and pairwise constrained component analysis (PCCA) [8] all learn Mahalanobis-type distance functions using variants of the basic pairwise constraints principle. kPCCA [8], kLFDA [13], and kMFA [13] kernelize PCCA, LFDA, and MFA, respectively. kCCA [56] adopts canonical correlation analysis to map the kernelized features into a common subspace. For these kernel-based methods, we consider the standard linear, exponential (exp), chi2 (χ^2), and chi2-rbf ($\mathtt{R}_{\chi^2}$) kernels. In RankSVM [5], a weight vector that weights the different features appropriately is learned using a soft-margin SVM formulation. In SVMML [53], locally adaptive decision functions are learned in a large-margin SVM framework.

2.3 Multi-shot ranking

While most re-id algorithms are single-shot, i.e., features are extracted from a single probe image of the person of interest, the multi-shot scenario, in which features are extracted from a series of images of the person of interest, is arguably more relevant to video analysis problems. The simplest way to handle multi-shot data is to compute the average feature vector for each person, effectively resulting in a single-shot problem. However, we also evaluated several algorithms that inherently address multi-shot data, treating it as an image set and constructing affine hulls to compute the distance between a gallery and a probe person. Specifically, we considered the AHISD [58] and RNP [59] algorithms. While these methods were proposed in the context of face recognition, the basic notion of image set matching applies to re-id as well. We also evaluated a multi-shot method based on sparse ranking, in which re-id is posed as a sparse recovery problem. Specifically, we consider SRID [60], where a block sparse recovery problem is solved to retrieve the identity of a probe person, and ISR [50], where the recovered sparse coefficient vector is re-weighted using an iterative scheme to rank gallery candidates and re-identify the person of interest.

2.4 Techniques not (yet) considered

As noted in Section 1, the framework we adopt involves evaluating all possible combinations of candidate feature extraction, metric learning, and multi-shot ranking algorithms.

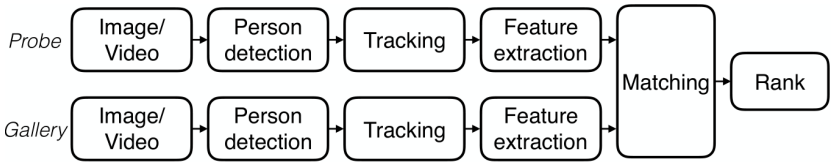


Fig. 1. A typical end-to-end re-id system pipeline.

Feature	Year
ELF [4]	ECCV 08
LDFV [35]	ECCVW 12
gBiCov [37]	BMVC 12
IDE-CaffeNet [42], [49]	NIPS 12, ECCV 16
DenseColorSIFT [9]	CVPR 13
HistLBP [13]	ECCV 14
IDE-VGGNet [44], [49]	ICLR 15, ECCV 16
LOMO [18]	CVPR 15
IDE-ResNet [43], [49]	ICCV 15, ECCV 16
WHOS [50]	T-PAMI 15
GOG [48]	CVPR 16

(a)

Metric	Year	Metric	Year
l_2		kPCCA [8]	CVPR 12
FDA [51]	AE 1936	LFDA [11]	CVPR 13
ITML [52]	ICML 07	SVMML [53]	CVPR 13
MFA [54]	PAMI 07	kMFA [13]	CVPR 13
LMNN [55]	JMLR 08	KCCA [56]	ICDSC 14
RankSVM [5]	BMVC 10	rPCCA [13]	ECCV 14
PRDC [6]	CVPR 11	kLFDA [13]	ECCV 14
KISSME [7]	CVPR 12	XQDA [18]	CVPR 15
PCCA [8]	CVPR 12	NFST [57]	CVPR 16

(b)

TABLE 1
Evaluated feature extraction and metric learning methods.

Methods that do not fall into this evaluation framework include post-rank learning methods [61], [62], unsupervised learning [9], [24], [63], attribute learning [64], [65], [66], ensemble methods [67], [68], [69] and mid-level representation learning [14]. A more comprehensive survey of these and other related methods can be found in the book by Gong *et al.* [2] and papers by Zheng [49], Satta [70], Vezzani [71], and Bedagkar-Gala and Shah [72]. While these methods are currently not part of our evaluation, we plan to expand our study and include them in a future release. We note that while all the evaluated algorithms follow a two-step process of feature and metric learning, we do consider a baseline Siamese CNN [73] algorithm that learns features and metrics together in a single-step approach. While we provide extensive discussion about challenges and opportunities in learning more powerful architectures that follow this single-step approach in Sections 4.5 and Sec 5, our results suggest that the two-step approach also gives competitive performance, and more importantly, can provide re-id specific domain knowledge and insights to aid future research.

3 Datasets

In this section, we briefly summarize the various publicly available datasets that are used in our benchmark evaluation. Table 2 provides a statistical summary of each dataset. Based on difficult examples, we also annotate each dataset with challenging attributes from the following list: viewpoint variations (VV), illumination variations (IV), detection errors (DE), occlusions (OCC), background clutter (BC), and low-resolution images (RES). We also indicate the number of bounding boxes (BBox) and cameras (cam) in each dataset, and the means by which the bounding boxes were obtained: using hand-labeling (hand), aggregated channel features [74] (ACF), or the deformable parts model detector [75] (DPM). Samples of difficult examples are provided as part of the supplementary material [1].

VIPeR [4] consists of 632 people from two disjoint views. Each person has only one image per view. VIPeR suffers from subtantial viewpoint and illumination variations. GRID [76] has 250 image pairs collected from 8 non-overlapping cameras. To mimic a realistic scenario, 775 non-paired people are included in the gallery set, which makes this dataset extremely challenging. GRID suffers from viewpoint variations, background clutter, occlusions and low-resolution images. CAVIAR [24] is constructed from two cameras in a shopping mall. Of the 72 people, we only use 50 people who have images from both cameras. CAVIAR suffers from viewpoint variations and low-resolution images. In the case of 3DPeS [26], the re-id community uses a set of selected snapshots instead of the original video, which includes 192 people and 1,011 images. 3DPeS suffers from viewpoint and illumination variations. PRID [25] is constructed from two outdoor cameras, with 385 tracking sequences from one camera and 749 tracking sequences from the other camera. Among them, only 200 people appear in both views. To be consistent with previous work [28], we use the same subset of the data with 178 people. PRID suffers from viewpoint and illumination variations. V47 [77] contains 47 people walking through two indoor cameras. WARD [27] collects 4,786 images of 70 people in 3 disjoint cameras. SAIVT-Softbio [78] consists of 152 people as seen from a surveillance camera network with 8 cameras. To be consistent with existing work [78], we use only two camera pairs: camera 3 and camera 8 (which we name SAIVT-38) and camera 5 and camera 8 (which we name SAIVT-58). Both these datasets suffer from viewpoint and illumination

1. Supplementary material can be found at https://arxiv.org/abs/1605.09653.

Dataset	# people	# BBox	# distractors	# cam	label	Attributes
VIPeR	632	1,264	0	2	hand	VV,IV
GRID	1025	1,275	775	2	hand	VV,BC,OCC,RES
CAVIAR	72	1,220	0	2	hand	VV,RES
3DPeS	192	1,011	0	8	hand	VV,IV
PRID	178	24,541	0	2	hand	VV,IV
V47	47	752	0	2	hand	-
WARD	70	4,786	0	3	hand	IV
SAIVT-Softbio	152	64,472	0	8	hand	VV,IV,BC
CUHK01	971	3,884	0	2	hand	VV,OCC
CUHK02	1,816	7,264	0	10	hand	VV,OCC
CUHK03	1,360	13,164	0	10	DPM/hand	VV,DE,OCC
RAiD	43	6,920	0	4	hand	VV,IV
iLIDSVID	300	42,495	0	2	hand	VV,IV,BC,OCC
HDA+	53	2,976	0	12	ACF/hand	VV,IV,DE
Market1501	1,501	32,217	2,793+500k	6	DPM	VV,DE,RES
DukeMTMC4ReID	1,852	46,261	21,551	8	Doppia	VV,IV,DE,BC,OCC
Airport	9,651	39,902	31,238	6	ACF	VV,IV,DE,BC,OCC

TABLE 2
The characteristics of the 17 datasets of the re-id benchmark.

variations, and background clutter. CUHK01 [79] has 971 people and 3,884 images captured from 2 disjoint camera views in a college campus setting. CUHK02 [80] has 1816 people and 7,264 images captured from 5 disjoint camera pairs in a college campus. All bounding boxes are manually labeled. CUHK03 [81] has 1360 people and 13,164 images from 5 disjoint camera pairs. Both manually labeled bounding boxes and automatically detected bounding boxes using the DPM detector [75] are provided. We only use the detected bounding boxes in our experiments. CUHK03 suffers from viewpoint variations, detection errors, and occlusions. RAiD [29] includes 43 people as seen from two indoor and two outdoor cameras and suffers from viewpoint and illumination variations. iLIDSVID [28] includes 600 tracking sequences for 300 people from 2 non-overlapping cameras in an airport and suffers from viewpoint and illumination variations, background clutter, and occlusions. HDA+ [82] was proposed to be a testbed for an automatic re-id system. Fully labeled frames for 30-minute long videos from 13 disjoint cameras are provided. Since we only focus on the re-id problem, we use pre-detected bounding boxes generated using the ACF [74] detector. DukeMTMC4ReID [83] has 1852 identities with 46,261 images and 21,551 false alarms from the Doppia person detector [84]. The images are captured from a disjoint 8-camera network located at the Duke University campus. This dataset was constructed specifically for re-id from the DukeMTMC multi-camera multi-target tracking dataset [85]. Market1501 [86] has 1,501 people with 32,643 images and 2,793 false alarms from the DPM person detector [75]. Besides these, an additional 500,000 false alarms and non-paired people are also provided to emphasize practical problems in re-id. Market1501 suffers from viewpoint variations, detection errors and low-resolution. Airport is the new dataset we introduce in the next section.

3.1 A new, real-world, large-scale dataset

In this section, we provide details about a new re-id dataset we designed for this benchmark. The dataset was created using videos from six cameras of an indoor surveillance network in a mid-sized airport; this testbed is described further in [3]. The cameras cover various parts of a central security checkpoint area and three concourses. Each of our cameras has 768×432 pixels and captures video at 30 frames per second. 12-hour long videos from 8 AM to 8 PM were collected from each of these cameras. Under the assumption that each target person takes a limited amount of time to travel through our camera network, each of these long videos was randomly split into 40 five minute long video clips. Each video clip was then run through a prototype end-to-end re-id system comprised of automatic person detection and tracking algorithms. Specifically, we employed the ACF framework of Dollar *et al.* [74] to detect people and a combination of FAST corner features [87] and the KLT tracker [88] to track people and associate any broken "tracklets". The dataset can be requested at http://www.northeastern.edu/alert/transitioning-technology/alert-datasets/alert-airport-re-identification-dataset/.

Unlike other datasets that capture image data from public environments such as universities [83], [85] [41] [14], shopping locations [24], or publicly accessible spots in transportation gateways [30], the Airport dataset provides data captured from video streams inside the secure area, post the security checkpoint, of a major airport. It is generally very difficult to obtain data from such a camera network, in which configuration settings (e.g., network topology and placement of cameras) are driven by security requirements. For instance, most academic datasets summarized in Table 2 have images taken from cameras with optical axes parallel to the ground plane, as opposed to the real world where the angle is usually much larger due to constraints on where and how the cameras can be installed. This aspect is explicitly captured by the Airport dataset. Unlike other

datasets that primarily capture images of people in a university setup (e.g., Market1501, CUHK, DukeMTMC4ReID), the Airport dataset captures images of people from an eclectic mix of professions, leading to a richer, more diversified set of images. Another key difference with existing datasets is the temporal aspect; we capture richer time-varying crowd dynamics, i.e., the density of people appearing in the source videos naturally varies according to the flight schedule at each hour. Such time-varying behavior can help evaluate the temporal performance of re-id algorithms, an understudied area [89].

Since all the bounding boxes were generated automatically without any manual annotation, this dataset accurately mimics a real-world re-id problem setting. A typical fully automatic re-id system should be able to automatically detect, track, and match people seen in the gallery camera, and the proposed dataset exactly reflects this setup. In total, from all the short video clips, tracks corresponding to 9,651 unique people were extracted. The number of bounding box images in the dataset is 39,902, giving an average of 3.13 images per person. The sizes of detected bounding boxes range from 130×54 to 403×166. 1,382 of the 9,651 people are paired in at least two cameras. A number of unpaired people are also included in the dataset to simulate how a real-world re-id system would work: given a person of interest in the probe camera, a real system would automatically detect and track all the people seen in the gallery camera. Therefore, having a dataset with a large number of unpaired people greatly facilitates algorithmic re-id research by closely simulating a real-world environment. While this aspect is discussed in more detail in our system paper [3], we briefly describe how this dataset can be used to validate detection and tracking algorithms typically used in an end-to-end re-id system. Specifically, since we have both valid and invalid detections in our dataset, we can use them interchangeably to evaluate the impact of the detection module. For instance, adding invalid detections to the gallery would help evaluate the need for more detection accuracy at the cost of compute time. Since we have access to multiple broken tracklets for each person, we can interchangeably use them to evaluate the impact of the tracking module. For instance, manually associating all broken tracklets can help evaluate the need for more tracking accuracy at the cost of compute time. We can also fuse these two concepts together to evaluate the need for more detection and tracking accuracy together, helping understand the upper-bound performance of real-world systems. A sample of the images available in the dataset is shown in Figure 2. As can be seen from the figure, these are the kind of images one would expect from a fully automated system with detection and tracking modules working in a real-world surveillance environment. As noted in Table 2, the Airport dataset suffers from all challenging attributes except low resolution. That is because relatively small detections are rejected by the person detector.

3.2 Evaluation protocol

3.2.1 Datasets, and training and testing splits.

Based on the number of images for each probe person, we categorize the datasets into either the single-shot or multi-shot setting. We employ the single-shot evaluation protocol for VIPeR, GRID, 3DPeS, DukeMTMC4ReID, CUHK01, CUHK02, CUHK03, HDA+, Market1501, and Airport. For the other 7 datasets, we employ the multi-shot evaluation protocol. In the Airport dataset, we fix one of the 6 cameras as the probe view and randomly pick paired people from 20 of the 40 short clips as the training set, with the rest forming the testing set. In the case of CUHK03, DukeMTMC4ReID, GRID, HDA+, and Market1501, we use the partition files provided by the respective authors. In particular, for the CUHK03 dataset, as noted in the previous section, we only use the "detected" bounding boxes in all reported experiments. In RAiD, we fix camera 1 as the probe view, resulting in three sub-datasets, RAiD-12, RAiD-13, and RAiD-14, corresponding to the 3 possible gallery views. RAiD-12 has 43 people, of which we use 21 people to construct the training set and the rest to construct the testing set. The other two sub-datasets have 42 people each, which we split into equal-sized training and testing sets. In WARD, we fix camera 1 as the probe view, resulting in two sub-datasets, WARD-12 and WARD-13, corresponding to the 2 possible gallery views. Both these sub-datasets have 70 people each. We split VIPeR, CUHK01, CUHK02, GRID, CAVIAR, 3DPeS, PRID, WARD-12, WARD-13 and iLIDSVID into equal-sized training and testing sets. SAIVT-38 has 99 people, of which we use 31 people for training and the rest for testing. SAIVT-58 has 103 people, of which we use 33 people for training and the rest for testing. We note that in the cases of iLIDSVID, PRID, and SAIVT, the split protocol used here is the same as in previous works that propose multi-shot re-id algorithms [28], [90], [91], [92], [93] to ensure evaluation consistency. Finally, for each dataset, we use 10 different randomly generated training and testing sets and report the overall average results.

3.2.2 Evaluation framework.

In the single-shot evaluation scheme, for each dataset, we consider two aspects: type of feature and type of metric learning algorithm. We evaluate all possible combinations of the 11 different features and 18 different metric learning algorithms listed in Table 1. Since we also evaluate four different kernels for the kernelized algorithms, the total number of algorithm combinations is 276.[2] In the multi-shot evaluation scheme, we consider three aspects: type of feature, type of metric learning algorithm, and type of ranking algorithm. Additionally, we consider two evaluation subschemes: using the average feature vector as the data representative (called AVER), and clustering the multiple feature vectors for each person and considering the resulting cluster centers as the representative feature vectors for each person (called CLUST). AVER effectively converts each dataset into an equivalent single-shot version. However, in the case of CLUST, we do not consider kernelized metric learning algorithms and other non-typical algorithms such as RankSVM and SVMML because only AVER can be employed to rank gallery candidates. Consequently, we use the remaining 9 metric learning algorithms and the baseline l_2 method, in which we use the features in the original space without any projection. These 10 algorithms are used in combina-

2. We evaluate only linear and exp kernels for LDFV, GOG, IDE-CaffeNet, IDE-ResNet, and IDE-VGGNet.

Fig. 2. Samples of images from the proposed Airport dataset. See supplementary material for more snapshots.

tion with the 11 different features and 4 different ranking algorithms. In total, we evaluate 646 different algorithm combinations for each multi-shot dataset.

3.2.3 Implementation and parameter details.

We normalize all images of a particular dataset to the same size, which is set to 128×48 for VIPeR, GRID, CAVIAR and 3DPeS and 128×64 for all other datasets. To compute features, we divide each image into 6 horizontal rectangular strips. In the case of LOMO, since the patches are fixed to be square-shaped, we obtain 12 patches for a 128×48 image and 18 patches for a 128×64 image.

In the case of IDE-ResNet and IDE-VGGNet, we resize each image to 224×224 pixels following [44]. For IDE-CaffeNet, we resize each image to 227×227. We start training with a model pre-trained on the ImageNet dataset and train the fully connected layers fc7 and fc8 from scratch. The number of output units in the fc7 layer is set to 4096 for IDE-VGGNet and IDE-CaffeNet, and 2048 for IDE-ResNet. Since we consider each person to be a different class, we set the number of output units in the fc8 layer to the number of unique people in our training set. Depending on the training split, this number varies from 2560 to 2580. Once the model is trained, we use the output of the fc7 layer as the image descriptor, giving a 4096-dimensional feature vector in the case of IDE-VGGNet and IDE-CaffeNet, and a 2048-dimensional feature vector in the case of IDE-ResNet.

In metric learning, we set the projected feature space dimension to 40. We set the ratio of the number of negative to positive pairwise constraints to 10.[3] In the case of CLUST, we set the number of clusters to 10, which we determine using the k-means algorithm.

Datasets	Best Combination	1	5	10
VIPeR	GOG-XQDA	41.1	71.1	82.1
GRID	IDE-ResNet-KISSME	26.6	43.1	50.9
3DPeS	IDE-ResNet-NFST$_{exp}$	53.4	77.8	85.6
CUHK01	GOG-NFST$_{exp}$	55.6	77.7	84.8
CUHK02	GOG-NFST$_{exp}$	57.9	79.3	85.7
CUHK03	GOG-kLFDA$_{exp}$	62.1	88.7	94.2
HDA+	IDE-ResNet-NFST$_{\ell}$	84.1	84.5	85.8
Market1501	IDE-ResNet-NFST$_{exp}$	64.3	80.9	86.2
Airport	IDE-ResNet-NFST$_{exp}$	42.7	67.5	76.0
DukeMTMC4ReID	IDE-ResNet-NFST$_{exp}$	54.6	68.6	73.7
PRID	GOG-KISSME-SRID	91.5	97.8	98.8
V47	IDE-ResNet-KISSME-RNP	100.0	100.0	100.0
CAVIAR	GOG-KISSME-RNP	55.6	79.6	95.6
WARD-12	GOG-KISSME-SRID	99.7	100.0	100.0
WARD-13	GOG-KISSME-ISR	97.7	98.6	99.1
SAIVT-38	GOG-KISSME-SRID	96.5	100.0	100.0
SAIVT-58	GOG-KISSME-RNP	72.6	89.9	93.0
RAiD-12	IDE-ResNet-KISSME-AHISD	100.0	100.0	100.0
RAiD-13	GOG-KISSME-SRID	81.9	94.8	96.2
RAiD-14	GOG-KISSME-SRID	95.7	96.2	99.5
iLIDSVID	GOG-KISSME-SRID	75.7	90.1	93.6

TABLE 3
Top performing algorithmic combinations on each dataset, where we show the re-id performance (%) at ranks 1, 5, and 10. Read as feature-metric for single-shot and feature-metric-ranking for multi-shot.

4 Results and Discussion

We first summarize the results of the overall evaluation, and then discuss several aspects of these results in detail.

The overall cumulative match characteristic (CMC) curves for two representative single- and multi-shot datasets are shown in in Figure 3. The CMC curve is a plot of the re-identification rate at rank-k. The individual performance of each algorithm combination on all datasets as well as complete CMC curves can be found in the supplementary material. As can be seen from the CMC curves, the "spread" in the performance of the algorithms for each dataset is huge, indicating the progress made by the re-id community over the past decade. However, on most datasets, the performance is still far from the point where we would consider re-id to be a solved problem. In Table 3, we summarize the overall CMC curves by reporting the algorithm combination that achieved the best performance on each dataset as measured by the rank-1 performance. We note that IDE-ResNet [43] and GOG [48] perform the best among the 11 evaluated feature extraction algorithms, with them being a part of the best performing algorithm combination in 6 of the 10 single-shot and all the 11 multi-shot datasets respectively. In the case of multi-shot evaluation, the combination of KISSME [7] as the metric learning algorithm and SRID [60] as the multi-shot ranking algorithm is the best performing algorithm combination, with it resulting in the best performance on 6 of the 11 datasets.

In general, we observe that the algorithms give better performance on multi-shot datasets than on single-shot datasets. While this may be attributed to multi-shot datasets having more information in terms of the number of images per person, it is important to note that the single-shot datasets considered here generally have a significantly higher number of people in the gallery. It is quite natural to expect re-id performance to go down as the number of gallery people increases because we are now searching for the person of interest in a much larger candidate set.

3. This is set to 1 for kPCCA and rPCCA on Market1501 due to system memory issues.

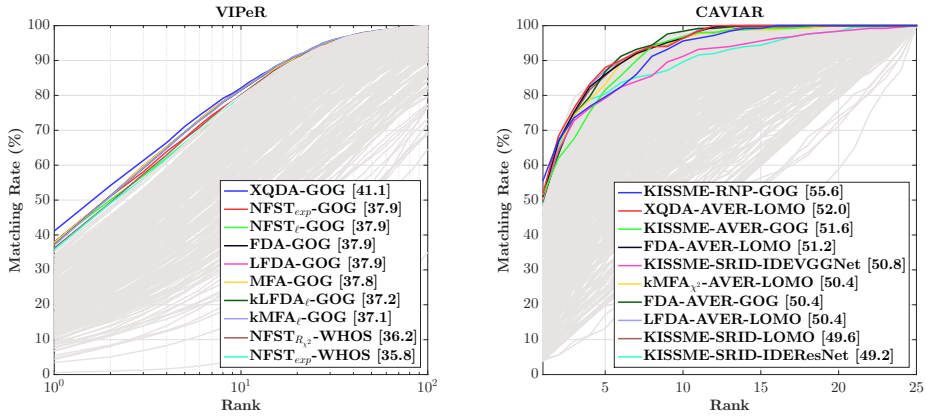


Fig. 3. CMC curves for the single-shot dataset VIPeR and the multi-shot dataset CAVIAR. The algorithmic combinations with the ten best rank-1 performances (indicated in the legend) are shown in color and all the others are shown in gray. CMC curves for all other datasets can be found in the supplementary material.

4.1 Single shot analysis: features and metric learning

Single-shot re-id involves two key aspects: features and metric learning. In this section, we isolate the impact of the best performing algorithm in these two areas. First, we note that IDE-ResNet is the best performing feature extraction algorithm in our evaluation. To corroborate this observation, we study the impact of IDE-ResNet in comparison with other feature extraction algorithms both in the presence as well as the absence of any metric learning. In the first experiment, we use the baseline Euclidean distance to rank gallery candidates in the originally computed feature space, which can be regarded as an unsupervised method. As can be noted from the results shown in Figure 4(a)[4], IDE-ResNet gives the best performance on all the 10 datasets.

Next, we study how IDE-ResNet performs in comparison with other features in the presence of metric learning. In this experiment, we fix NFST$_{\text{exp}}$ as our metric learning algorithm and rank gallery candidates using all the 11 evaluated feature extraction algorithms. The rank-1[5] results for this experiment are shown in Figure 4(b). As can be noted from the graph, IDE-ResNet gives the best performance on 4 of the 7 datasets shown in the figure, with GOG giving the best performance on the remaining 3 datasets.

These experiments show that IDE-ResNet is indeed the best performing feature extraction algorithm. This should not be surprising given the powerful modeling and generalization ability of the ResNet architecture, which is also evidenced by its strong performance in other computer vision domains and applications [43].

Here, we also note that GOG, despite being a hand-crafted feature extraction algorithm, results in competitive respectable performance. This is because color and texture are the most descriptive aspects of a person image and GOG describes the global color and texture distributions using a local Gaussian distributions of pixel-level features. Another critical reason for the success of GOG is the hierarchical modeling of local color and texture structures. This is a critical step because typically a person's clothes consists of local parts, each of which has certain local properties.

Next, we analyze the performance of different metric learning algorithms[6], in the context of IDE-ResNet, the best performing feature extraction algorithm. In this experiment, we fix IDE-ResNet as the feature extraction algorithm and study how different metric learning algorithms perform. The results of this experiment are shown in Figure 4(c), from which we can note that NFST$_{\text{exp}}$ gives the best performance on Market1501, DukeMTMC, 3DPeS, and Airport, with XQDA and kLFDA not being too far behind. These results further corroborate what we observe in Table 3, with NFST, kLFDA, and XQDA being among the best performing metric learning algorithms.

From the above discussion, we can infer the following: while NFST$_{\text{exp}}$ gives the best overall performance, kLFDA and XQDA also emerge as strong and competitive metric learning algorithms. It is interesting to note that all these three algorithms learn the distance metric by solving some form of generalized eigenvalue decomposition problems, similar to traditional Fisher discriminant analysis. While kLFDA and XQDA directly employ Fisher-type objective functions, NFST uses the Foley-Shannon transform [94], which is very closely related to the Fisher discriminant analysis. This suggests that the approach of formulating discriminant objective functions in terms of data scatter matrices is most suitable to the re-id problem.

4.2 Multi-shot analysis: features, metric learning, and ranking

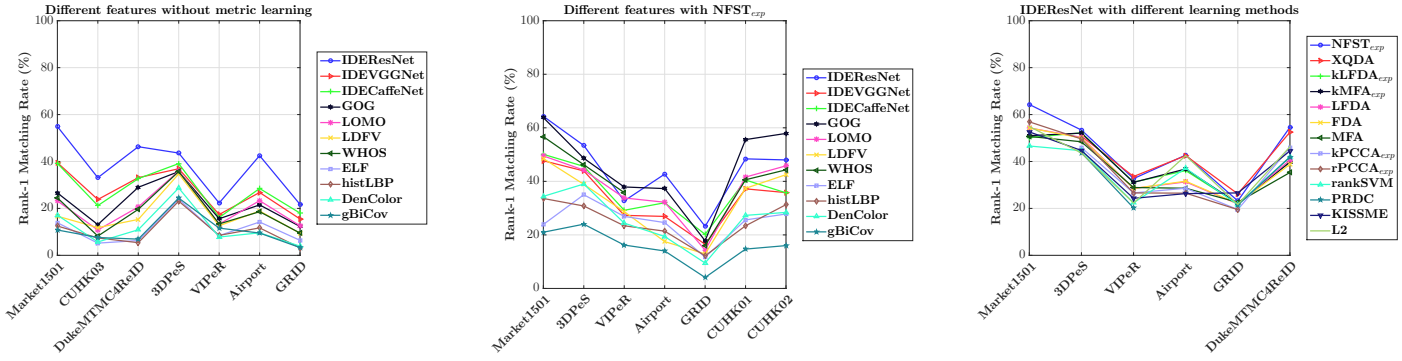
Multi-shot re-id involves three aspects: features, metric learning, and ranking. As noted previously, GOG, KISSME, and SRID emerged as the best performing algorithmic combination. On all the datasets, as expected, a custom ranking algorithm resulted in the best performance, with SRID performing the best on 6 of these 11 datasets. In this section, we provide further empirical results analyzing the impact of using a multi-shot ranking algorithm. To this end, we fix GOG as the feature extraction scheme.

First, we evaluate the impact of using a multi-shot ranking algorithm instead of AVER. Here, we compare the

4. In the graph, we only show results on 7 datasets for brevity. Please consult supplementary material for complete results.

5. Complete CMC curves can be found in the supplementary material.

6. A discussion on the training time of these algorithms is provided in the supplementary material.



(a) (b) (c)

Fig. 4. Rank-1 results for single shot datasets illustrating the impact of IDE-ResNet and NFST$_{exp}$.

performance of GOG-KISSME-AVER and GOG-KISSME-SRID. The results are shown in Figure 5(a). As can be noted from the graph, with the exception of CAVIAR, SRID generally gives superior performance when compared to AVER. This, and our observations from Table 3, suggest that using a multi-shot ranking algorithm that exploits the inherent structure of the data instead of naive feature averaging will give better performance. Furthermore, we also note that a multi-shot ranking algorithm in itself is not sufficient to give good performance because that would be purely an unsupervised approach. Combining a metric learning algorithm with the ranking technique adds a layer of supervision to the overall framework and will provide a significant performance boost.

Next, we analyze the performance of the feature extraction and metric learning algorithms and compare the observed trends with those in the single-shot case. In the feature extraction case, we fix SRID as the ranking algorithm and KISSME as the metric learning algorithm. The rank-1 results for this experiment are shown in Figure 5(b)-(c). We see very clear trends in this case, with GOG giving the best results on all the 11 datasets. These results are not surprising given the strong performance of GOG in the single-shot case. In the metric learning case, we fix SRID as the ranking algorithm and GOG as the feature extraction algorithm, with Figure 5(d) showing the rank-1 results. We see very clear trends in this case as well, with KISSME giving the best results across all datasets.

4.3 Additional observations

In this section, we report additional empirical observations. Most contemporary feature extraction schemes produce high-dimensional data, introducing significant computational overhead. To this end, we analyze the impact of an unsupervised dimensionality reduction scheme, principal component analysis (PCA). We fix GOG, features with the highest dimensionality in our evaluation framework, as the feature extraction scheme and perform experiments with and without PCA. We set the dimension of the PCA-reduced space to 100. The results are shown in the first two bars (pink and yellow) in Figure 6(a). The results without PCA are better than those with PCA on all the datasets shown in the graphs. This observation is not surprising given that PCA can result in the undesirable removal of the most discriminative features.

All hand-crafted feature extraction algorithms use some form of localized feature computation by dividing the image into pre-defined strips. Here, we analyze the impact of the number-of-strips parameter. To this end, we perform experiments with 6, 9, and 15 horizontal strips in the best hand-crafted feature extraction algorithm, GOG, with Euclidean distance as the metric in the single-shot case and Euclidean distance as the metric and AVER as the ranking strategy in the multi-shot case. The rank-1 results are shown in bars 2–5 in Figure 6(a) [7]. While it is reasonable to expect superior performance as we increase the number of strips, thereby increasing the feature space dimensionality, it is important to note that in this process, we may have fewer samples to estimate the Gaussians in each strip and also increase the amount of background/noise/non-informative content in the feature descriptor. We also note that there does not seem to be any significant performance variations as we increase the number of strips. Given the computational complexity involved in working with higher dimensional feature spaces due to increased number of strips, these results suggest that 6 strips, which is in fact the widely used number in the re-id community, seems to be a reasonable choice, giving better or close performance to the other choices in most cases.

Finally, we also empirically study how re-id accuracy varies vis-a-vis computational requirements for various values of the PCA dimension. In Figure 6(b), we show results of this experiment for values of PCA dimension ranging from 50 to 500 for a small-scale dataset, VIPeR, and a large-scale dataset, Market1501. The numbers on top of the bars are the training times in seconds. As can be noted from the results, as we increase the PCA dimension, the training time increases (quite substantially for the large-scale dataset), while the accuracy saturates beyond a certain value of the PCA dimension. This empirically substantiates the sufficiency of a relatively small value for performing dimensionality reduction using algorithms such as the PCA.

4.4 Attribute-based analysis

In this section, we analyze the performance of the different feature extraction schemes with respect to the different attributes used to characterize datasets in Table 2. The goal of this experiment is to study which features are good in

7. The rank-1 results for the multi-shot case are provided in the supplementary material.

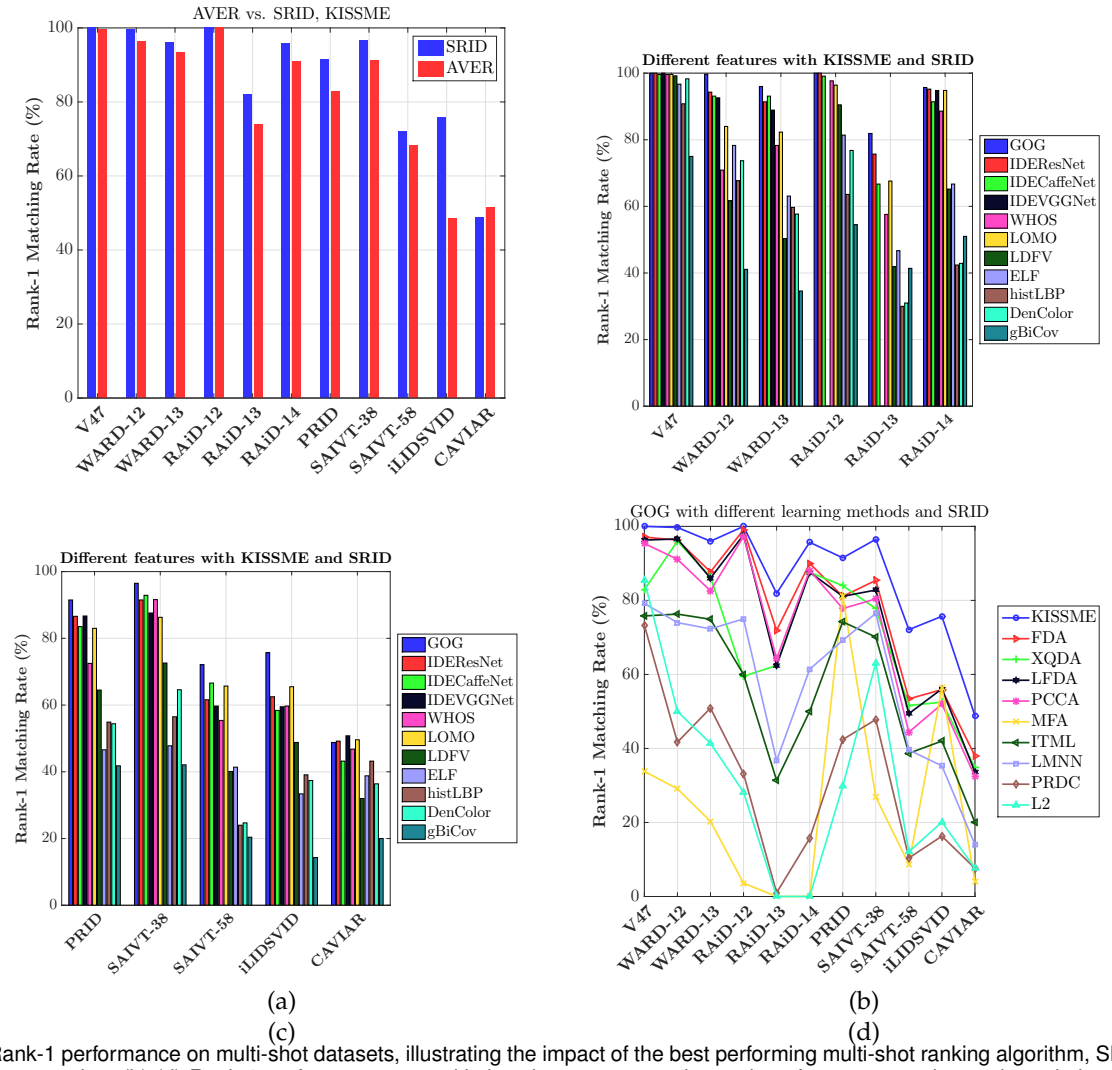


Fig. 5. (a): Rank-1 performance on multi-shot datasets, illustrating the impact of the best performing multi-shot ranking algorithm, SRID over AVER, naive feature averaging. (b)-(d) Rank-1 performance on multi-shot datasets comparing various feature extraction and metric learning algorithms with SRID as the ranking algorithm.

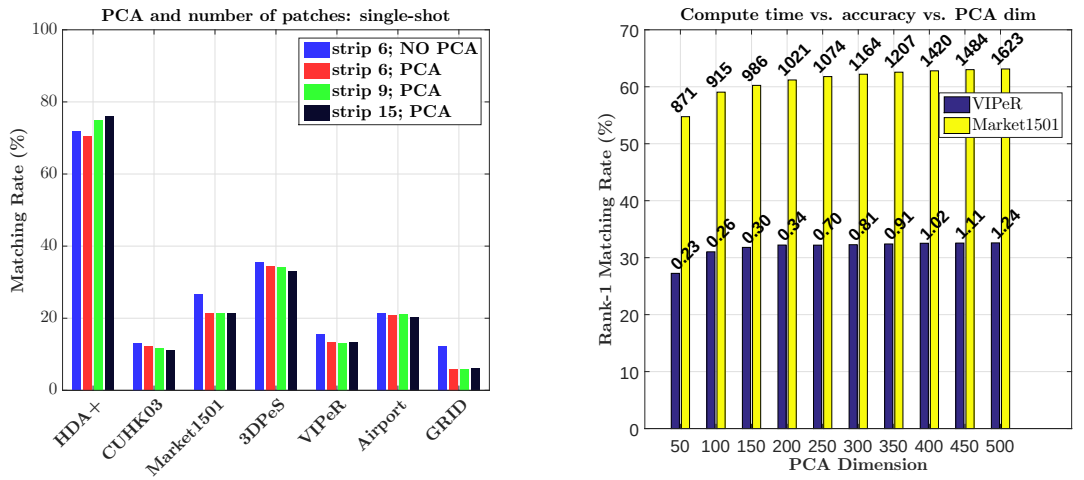


Fig. 6. Rank-1 performance on (a) single-shot datasets illustrating the impact of PCA and number of strips. (b) Rank-1 and training time (in seconds) for VIPeR and Market1501 for various values of the PCA dimension.

	Attr	IDERes	IDEVGG	IDECaffe	GOG	LOMO	LDFV	WHOS	ELF	HistLBP	DenColor	gBiCov
Single-shot	VV	**40.7**	30.4	29.2	**26.6**	18.3	**19.6**	17.2	11.7	9.0	11.7	15.5
	IV	**45.8**	36.3	33.5	**34.7**	22.6	**26.7**	20.7	16.6	12.0	15.8	23.1
	BC	**36.8**	25.1	26.4	**20.9**	**18.8**	14.5	15.9	9.0	6.8	8.1	6.6
	OCC	**35.3**	23.8	24.4	**19.3**	**15.2**	13.3	13.7	6.7	5.8	6.9	7.5
	RES	**38.3**	27.5	28.6	**19.5**	**17.9**	13.1	17.0	10.2	7.9	10.5	7.1
Multi-shot	VV	**49.1**	46.7	37.9	**36.7**	**20.2**	11.9	15.7	10.5	8.2	9.2	8.4
	IV	**65.1**	61.5	50.7	**51.5**	**24.1**	13.7	21.7	11.2	9.1	7.3	8.6
	BC	**61.0**	52.3	53.6	**43.7**	17.2	14.5	**34.2**	5.2	6.6	9.2	12.4
	OCC	**42.0**	37.4	27.5	**23.2**	6.8	5.5	**12.9**	5.7	3.1	8.3	5.1
	RES	**34.0**	26.8	24.4	**28.4**	**25.2**	12.4	14.0	18.0	15.6	13.6	11.2

TABLE 4
Mean rank-1 performance across all single- and multi-shot datasets with respect to various attributes and features. The best deep learning feature is shown in red and the best and second best hand-crafted features are shown in green and blue respectively.

certain scenarios. To this end, we use Euclidean distance as the metric, and in the multi-shot case, AVER as the ranking algorithm, and report the mean rank-1 performance on all datasets for each attribute group. The results obtained are shown in Table 4. We observe the following trends from the results. In all the scenarios, IDE-ResNet resulted in the best performance, with IDE-VGGNet and IDE-CaffeNet following closely behind. For a few attributes, GOG gave competitive, albeit lower, performance when compared to the IDE features (e.g., VV, IV, and RES).

While the IDE results are not surprising given the relatively strong supervision from annotated training data, hand-crafted features can provide us with insights on learning more re-id specific domain knowledge. While we only discuss the best performing hand-crafted algorithms- GOG, LDFV, and LOMO- these insights can be quite useful across the spectrum as researchers think about designing feature learning architectures. First, as opposed to other hand-crafted algorithms, these three methods explicitly model local pixel distributions. While this is intuitively obvious, we note that is an extremely important step in describing person images. Additionally, since viewpoint invariance is an extremely important attribute for any re-id descriptor, these results suggest that incorporating local region information and horizontal strip pooling, as done explicitly in both GOG and LOMO, is critical to achieve viewpoint invariant descriptors. Furthermore, we note that WHOS results in strong performance on BC and RES in the single-shot case, primarily due to the use of a mask that filters out background clutters, resulting in a more localized features representation, similar in spirit to the approach noted above.

4.5 Impact of datasets on re-id research

Datasets play an extremely important role in validating a newly proposed algorithm. From Table 3, we note that the V47 dataset can be considered to be solved, with the best performing algorithm achieving a rank-1 performance of 100%. However, the performance on other datasets is still far from ideal. These datasets therefore present opportunities and challenges to develop better algorithms. For instance, the performance on VIPeR is still very low despite it being the most popular dataset in the re-id community. The performance on GRID is the lowest (26.6% at rank-1) and this is in part due to the presence of a large number of distractor people in the gallery. The newly proposed Airport dataset has the next lowest performance (42.7% at rank-1). This is due to the presence of a large distractor set as well as the significant real-world challenges described in Section 3.1. These observations suggest that as the number of distractor people in the gallery increases, the performance of a re-id algorithm goes down. This is not surprising since now we have a larger set of people to compare the probe person against, leading to more avenues for the re-id algorithm to fail.

Generally speaking, there are two key aspects that need to be considered while constructing new datasets: they have to be both **large** and **realistic**. As noted above, the presence of a distractor set helps mimic the real-world nature to a certain extent. An important point we would like to emphasize relates to the notion of distractors as commonly used in constructing datasets; in most cases, these correspond to false detections provided by a person detector. While this is somewhat reasonable, in the real world, we have both false alarms and actual person images, so having a number of other unpaired person images, in addition to false positives from a person detector, is crucial. Given that end-users, mostly security personnel, will be searching for a person of interest among hundreds of thousands of people, having a re-id dataset this large would help to quickly scale-up re-id algorithms to become practically relevant. Furthermore, and crucially, datasets must be relevant to real-world application scenarios. Most current dataset construction mechanisms focus on capturing images of people under different camera views, illumination conditions, or other factors such as occlusion. While these factors are important, what is missing is more fundamental: since re-id primarily finds applications in crime detection and prevention, the perpetrator or the person of interest may re-appear in a different (e.g., changed clothes, hair style, etc.) appearance. Again, for re-id algorithms to be used in the real world, we need datasets that capture this subtle yet extremely important aspect. A more extensive discussion is provided in Section 5.

Our categorization of datasets according to their attributes also provides a useful reference point for constructing new datasets. We note that none of the 14 datasets have all 6 attributes. While MARS [41], a recently proposed dataset, has a large number of images, constructing datasets that are of the size of ImageNet, in terms of the number of people, positive examples, and under an eclectic mix of conditions as noted above, will assist in the application of some of the recent algorithmic advances in feature learning using CNNs [43], [95]. We believe focusing on the aspects

discussed in this section, and several others discussed in Section 5, when constructing new datasets, would help rapidly accelerate progress in person re-id.

5 Insights and Recommendations for Researchers

Our systematic study of re-id algorithms across many datasets helps us characterize what re-id algorithms are currently capable of doing, as well as what is missing and can be done in the future. To this end, we here discuss insights gained from this exercise, as well as posit research directions and recommendations for re-id researchers that would help develop better algorithms.

We noted that IDE-ResNet resulted in the best performance among all the evaluated feature extraction methdods, with IDE-VGGNet and IDE-CaffeNet close behind. Again, this is expected due to powerful, generalizable features that CNNs are capable of learning giving enough supervision. While adapting more recent improvements in architecture design and feature learning [96], [97] will naturally give better performance, we would, in particular, like to note the strong performance given by GOG. A primary reason for its success is the hierarchical modeling of local pixel distributions, first at the patch level and then at the strip level which is inspired by LOMO. There are two key takeaways from this observation: local features and hierarchical modeling. Intuitively, this should not be surprising since using local features helps mitigate potential issues with background clutter and noise, two challenges that are critical for re-id. Following recent advances in feature learning, a particularly promising research direction would be to learn **local patch representations**, which can then be aggregated into an image-level descriptor using existing aggregation schemes [98]. The region proposal network of Faster R-CNN [99] is a potential candidate to generate local patch proposals and learn representations, which can be trained in an end-to-end fashion. There have been some recent efforts to this end [100], [101], where specific architectures are designed to learn local body-region features. Strip-level pooling is another important aspect that is unique to the re-id problem. Because people are roughly vertically aligned in person images, translation-invariant pooling for each local strip (typically constructed horizontally) can help mitigate issues caused by viewpoint variations across cameras. Consequently, recent advances in learning translation invariant local features such as bilinear CNNs [102] or gated CNNs with specially designed convolution operations [103] would be particularly relevant.

As noted above, we can potentially use a localized feature representation approach to mitigate issues caused by background clutter and noise. In WHOS, another hand-crafted feature representation algorithm, a simple background filter was used to address this issue. Specifically, features were weighted according to their distance to the center. GOG also has the similar trick by weighting patches based on the distance to the center line. While issues like occlusion can certainly create problems, this assumption is not entirely unreasonable. An immediate idea to improve this strategy would be to learn camera-specific background distributions [23]. At its core, the fundamental idea of the strategy described here and in the previous paragraph is to focus on the "person" part of the image as much as possible; we can do this in a much more sophisticated fashion using recent advances in image segmentation [104], [105], learning the image representation while simultaneously segmenting out the background.

In the context of multi-shot or video-based re-id, we have much more information than a single image for each person, and this can be in the form of a set of images or a video sequence. In addition to the spatial aspect, we can exploit the temporal dimension as well to obtain better feature representations. For instance, we can borrow ideas from the C3D network of Tran *et al.* [106] to learn **spatio-temporal feature representations** [107] for each available video sequence, following which existing frameworks can be employed to learn discriminative distance metrics. An immediate follow-up to this idea would be the concept of spatio-temporal Siamese networks. While existing Siamese network frameworks learn to tell pairs of images apart, we can extend them, in conjunction with C3D-like networks, to tell pairs of video sequences apart. In multi-shot ranking, we demonstrated that using a custom ranking method gives much better performance compared to using the feature averaging scheme. In practical re-id applications, an image sequence of a person will typically undergo several variations such as background clutter, occlusion, and illumination variations. Developing custom multi-shot ranking algorithms that take this data variance into account will give better performance. Another promising future research direction in this context would be to integrate multi-shot ranking with metric learning. While most existing methods treat these two topics separately, developing a unified metric learning and multi-shot ranking framework that exploits the several aspects of multi-shot data can potentially lead to further performance gains.

While existing re-id datasets may not be large enough to use recent advances in feature learning, we can use smart augmentation strategies to generate a large number of synthetic images. While common strategies such as random 2-D translations can be readily applied [73], we can use sophisticated methods such as CycleGAN [108] and LSRO [109] to generate realisitic and meaningful images. For instance, starting from a base person image, we can generate images with different attributes, such as with and without sunglasses, with and without a backpack, and so on, each with a different branch of CycleGAN [110]. Furthermore, we can combine these attributes to generate fused images– for instance, an image of a person with glasses wearing a backpack. Such strategies can be used to generate meaningfully augmented datasets with large number of **diverse** images, which can in turn help fuel development of new deep learning approaches for re-id, some of which have been discussed in the previous paragraphs.

Among the various attributes we as humans use to re-identify people, **walking patterns** or gait are critical [111]. When we know who we are looking for, we can pick her/him up far away in a large crowd just by looking at the way s/he walks. Clearly, and intuitively, this information must be exploited by re-id algorithms. While there has been some work along these lines in the past [23], [28], [112], much more work needs to be done, specifically in using such

dynamic gait-based models in conjunction with appearance modeling via feature learning strategies.

An interesting but hardly addressed line of research in re-id is the use of **camera calibration** information. Typically, when one is constructing a new dataset, calibration information for the source cameras is easily obtainable. Such information can prove very handy in reducing the search space of candidates in the gallery. For instance, in a real-world application of re-id [1], [3], [113], we can use calibration information to estimate motion patterns of people. This can be used to filter candidates that only move in a certain direction, if the operator of the system is sufficiently confident about the trajectory of the person of interest. Furthermore, we can also use calibration information to estimate the ground plane of the scene, thereby helping estimate the height of the candidates seen in the gallery. This information can also be used as a filter to reduce the search space in the gallery.

Another very interesting aspect that is missing in existing re-id algorithms is the notion of **context**: when, where, and with whom is/was the person of interest moving? It is not unreasonable to assume that people often walk in groups, and this information can provide a strong prior to reduce the search space of gallery candidates. For instance, tracking groups of people can help model behavioral trends for the person of interest, providing a rich visual sense of context that can be used to perform re-id. Zheng *et al.* [114] used this notion of context to perform person matching in a multi-camera setup, and we believe this is a promising direction to pursue given the availability of several multi-camera re-id datasets.

Other promising directions for future research include using **multi-modal data** to alleviate potential problems with appearance feature learning. For instance, existing person re-id algorithms would fail in dark rooms or in cases where people wear similar clothes, for instance, in a laboratory or factory setting that mandates a dress code. In such scenarios, we could use depth information to estimate gait to perform re-id [115]. Furthermore, much recent work has focused on learning rich visual representations using RGB-D data [116], and this can be readily applied to the re-id problem.

Finally, we conclude with some thoughts on the current state of performance evaluation of re-id algorithms. We believe that researchers should take a broader view of how algorithms perform and not just look at raw rank-1 or mAP numbers. Specifically, as noted in Camps *et al.* [3], a re-id algorithm is only a small, but critical, part of a larger, fully automated system that tracks people in multi-camera networks. To this end, we should focus on creating datasets that accurately mimic scenarios where such systems would be employed. While the airport dataset we propose in this paper is a step towards this direction, much more work needs to be done to achieve practical, realistic evaluation mechanisms for re-id algorithms. For instance, a re-id application may stem from a crime detection/prevention perspective where the perpetrator, or the person of interest, may re-appear after several days or months, and in an entirely new appearance. Clearly, this shows we need datasets that have richer "temporal" aspects than those used in this paper. Specifically, datasets that have multiple re-appearances of people, ideally spanning across days or months, and appearances, will help in developing evaluation metrics that shed light on practical, real-world usability of re-id algorithms. A very early idea to this end is proposed in Karanam *et al.* [89] where the notion of "rank persistence" is used to evaluate re-id algorithms. The motivation for this approach is from the crime detection application discussed above; since we do not know when the perpetrator re-appears, the system continuously searches and ranks observed candidates. Once the person of interest in observed, we would like her/him to **stay** in the top-k rank list for **as long as possible**. We contend that researchers should focus on developing, and evaluating, re-id algorithms with this notion of persistence of the person of interest over time, resulting in algorithms that are meaningful from a real-world perspective.

REFERENCES

[1] Y. Li *et al.*, "Real-world re-identification in an airport camera network," in *ICDSC*, 2014.
[2] S. Gong *et al.*, *Person re-identification*. Springer, 2014, vol. 1.
[3] O. Camps *et al.*, "From the lab to the real world: Re-identification in an airport camera network," *CSVT*, vol. PP, no. 99, 2016.
[4] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008.
[5] B. Prosser *et al.*, "Person re-identification by support vector ranking." in *BMVC*, 2010.
[6] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011.
[7] M. Koestinger *et al.*, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.
[8] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.
[9] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.
[10] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *CVIU*, vol. 117, no. 2, pp. 130–144, 2013.
[11] S. Pedagadi *et al.*, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013.
[12] L. An *et al.*, "Reference-based person re-identification," in *AVSS*. IEEE, 2013, pp. 244–249.
[13] F. Xiong *et al.*, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014.
[14] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014.
[15] Z. Wu, Y. Li, and R. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *T-PAMI*, vol. 37, no. 5, pp. 1095–1108, 2015.
[16] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015.
[17] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, 2015.
[18] S. Liao *et al.*, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.
[19] W.-S. Zheng *et al.*, "Partial person re-identification," in *ICCV*, 2015.
[20] X. Li *et al.*, "Multi-scale learning for low-resolution person re-identification," in *ICCV*, 2015.
[21] S. Messelodi and C. M. Modena, "Boosting fisher vector based scoring functions for person re-identification," *IVC*, vol. 44, pp. 44–58, 2015.
[22] D. Chen *et al.*, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *CVPR*, 2015, pp. 1565–1573.
[23] M. Gou *et al.*, "Person re-identification in appearance impaired scenarios," in *BMVC*, 2016.
[24] D. S. Cheng *et al.*, "Custom pictorial structures for re-identification." in *BMVC*, 2011.
[25] M. Hirzer *et al.*, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, 2011.

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

[27] N. Martinel, C. Micheloni, and C. Piciarelli, "Distributed signature fusion for person re-identification," in *ICDSC*, 2012.

[28] T. Wang *et al.*, "Person re-identification by video ranking," in *ECCV*, 2014.

[29] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *ECCV*, 2014.

[30] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *ICIP*, 2013.

[31] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-3, pp. 7–42, 2002.

[32] S. Baker *et al.*, "A database and evaluation methodology for optical flow," *IJCV*, vol. 92, no. 1, pp. 1–31, 2011.

[33] C. Schmid, "Constructing models for content-based image retrieval," in *CVPR*, 2001.

[34] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological cybernetics*, vol. 61, no. 2, pp. 103–113, 1989.

[35] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV Workshops*, 2012.

[36] J. Sánchez *et al.*, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, 2013.

[37] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *IVC*, vol. 32, no. 6, pp. 379–390, 2014.

[38] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[39] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *T-PAMI*, vol. 30, no. 10, pp. 1713–1727, 2008.

[40] Y. Taigman *et al.*, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.

[41] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[43] K. He *et al.*, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

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

[45] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[46] S. Liao *et al.*, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *CVPR*, 2010.

[47] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *T-IP*, vol. 6, no. 7, pp. 965–976, 1997.

[48] T. Matsukawa *et al.*, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016.

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

[50] G. Lisanti *et al.*, "Person re-identification by iterative re-weighted sparse ranking," *T-PAMI*, vol. 37, no. 8, pp. 1629–1642, 2015.

[51] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics (AE)*, vol. 7, no. 2, pp. 179–188, 1936.

[52] J. V. Davis *et al.*, "Information-theoretic metric learning," in *ICML*, 2007.

[53] Z. Li *et al.*, "Learning locally-adaptive decision functions for person verification," in *CVPR*, 2013.

[54] S. Yan *et al.*, "Graph embedding and extensions: a general framework for dimensionality reduction," *T-PAMI*, vol. 29, no. 1, pp. 40–51, 2007.

[55] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, pp. 207–244, 2009.

[56] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," 2014.

[57] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016, pp. 1239–1248.

[58] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *CVPR*, 2010.

[59] M. Yang *et al.*, "Face recognition based on regularized nearest points between image sets," in *FG*, 2013.

[60] S. Karanam, Y. Li, and R. Radke, "Sparse re-id: Block sparsity for person re-identification," in *CVPR Workshops*, 2015.

[61] C. Liu *et al.*, "POP: Person re-identification post-rank optimisation," in *ICCV*, 2013.

[62] J. Garcia *et al.*, "Person re-identification ranking optimisation by discriminant context information analysis," in *ICCV*, 2015.

[63] S. Bak *et al.*, "Multiple-shot human re-identification by mean riemannian covariance grid," in *AVSS*, 2011.

[64] C. Su *et al.*, "Multi-task learning with low rank attribute embedding for person re-identification," in *ICCV*, 2015.

[65] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes." in *BMVC*, 2012.

[66] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *CVPR*, 2015.

[67] M. Eisenbach *et al.*, "Evaluation of multi feature fusion at score-level for appearance-based person re-identification," in *IJCNN*, 2015.

[68] L. Zheng *et al.*, "Query-adaptive late fusion for image search and person re-identification," in *CVPR*, 2015.

[69] N. Martinel, C. Micheloni, and G. L. Foresti, "A pool of multiple person re-identification experts," *PRL*, vol. 71, pp. 23–30, 2016.

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

[71] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 29, 2013.

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

[73] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[74] P. Dollár *et al.*, "Fast feature pyramids for object detection," *T-PAMI*, vol. 36, no. 8, pp. 1532–1545, 2014.

[75] P. F. Felzenszwalb *et al.*, "Object detection with discriminatively trained part-based models," *T-PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[76] C. C. Loy, T. Xiang, and S. Gong, "Time-delayed correlation analysis for multi-camera activity understanding," *IJCV*, vol. 90, no. 1, pp. 106–129, 2010.

[77] S. Wang *et al.*, "Re-identification of pedestrians with variable occlusion and scale," in *ICCV Workshops*, 2011.

[78] A. Bialkowski *et al.*, "A database for person re-identification in multi-camera surveillance networks," in *DICTA*, 2012.

[79] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning." Springer, 2012, pp. 31–44.

[80] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013, pp. 3594–3601.

[81] W. Li *et al.*, "DeepReId: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[82] D. Figueira *et al.*, "The HDA+ data set for research on fully automated re-identification systems," in *ECCV Workshops*, 2014.

[83] M. Gou *et al.*, "DukeMTMC4ReID: A large-scale multi-camera person re-identification dataset," in *CVPR Workshops*, 2017.

[84] R. Benenson *et al.*, "Ten years of pedestrian detection, what have we learned?" *arXiv preprint arXiv:1411.4304*, 2014.

[85] E. Ristani *et al.*, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV Workshops*, 2016.

[86] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

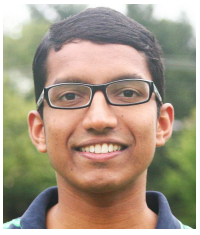
[87] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *T-PAMI*, vol. 32, no. 1, pp. 105–119, 2010.

[88] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Imaging Understanding Workshop*, 1981.

[89] S. Karanam, E. Lam, and R. J. Radke, "Rank persistence: Assessing the temporal performance of real-world person re-identification," in *ACM/IEEE International Conference on Distributed Smart Cameras*, 2017.

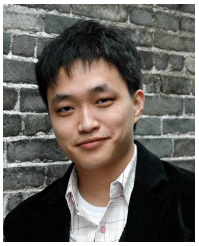
[90] T. Wang *et al.*, "Person re-identification by discriminative selection in video ranking," *T-PAMI*, vol. 38, no. 12, pp. 2501–2514, 2016.

[91] K. Liu *et al.*, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *ICCV*, 2015.
[92] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016.
[93] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with block sparse recovery," *Image and Vision Computing*, vol. 60, pp. 75–90, 2017.
[94] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors," *IEEE Transactions on Computers*, vol. 100, no. 3, pp. 281–289, 1975.
[95] C. Szegedy *et al.*, "Going deeper with convolutions," in *CVPR*, 2015.
[96] G. Huang *et al.*, "Densely connected convolutional networks," in *CVPR*, 2017.
[97] W. Chen *et al.*, "Beyond triplet loss: a deep quadruplet network for person re-identification," *arXiv preprint arXiv:1704.01719*, 2017.
[98] H. Jégou *et al.*, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.
[99] S. Ren *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[100] D. Li *et al.*, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, 2017, pp. 384–393.
[101] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017, pp. 1077–1085.
[102] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *ICCV*, 2015, pp. 1449–1457.
[103] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*. Springer, 2016, pp. 791–808.
[104] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.
[105] G. Lin *et al.*, "Exploring context with deep structured models for semantic segmentation," 2017.
[106] D. Tran *et al.*, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.
[107] L. Wang *et al.*, "Temporal segment networks: towards good practices for deep action recognition," in *ECCV*, 2016.
[108] J.-Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.
[109] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, 2017.
[110] Y. Gong *et al.*, "Learning compositional visual concepts with mutual consistency," *arXiv preprint arXiv:1711.06148*, 2017.
[111] J. E. Cutting, L. T. Kozlowski *et al.*, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin of the psychonomic society*, vol. 9, no. 5, pp. 353–356, 1977.
[112] Z. Liu *et al.*, "Enhancing person re-identification by integrating gait biometric," *Neurocomputing*, 2015.
[113] S. M. Assari, H. Idrees, and M. Shah, "Human re-identification in crowd videos using personal, social and environmental constraints," in *ECCV*. Springer, 2016, pp. 119–136.
[114] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *BMVC*, 2009.
[115] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *CVPR*, June 2016.
[116] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015, pp. 567–576.



Srikrishna Karanam Srikrishna Karanam is a Research Scientist in the Vision Technologies and Solutions group at Siemens Corporate Technology, Princeton, NJ. He has a Ph.D. degree in Computer & Systems Engineering from Rensselaer Polytechnic Institute. His research interests include computer vision and machine learning with focus on all aspects of image indexing, search, and retrieval for object recognition applications.

Mengran Guo Mengran Gou is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at Northeastern University. He received an M.S. degree from the Pennsylvania State University and B.Eng. degree from Harbin Institute of Technology in China. His research interests are person re-identification and activity recognition.



Ziyan Wu Ziyan Wu received a Ph.D. degree in Computer and Systems Engineering from Rensselaer Polytechnic Institute in 2014. He has B.S. and M.S. degrees in Engineering from Beihang University. He joined Siemens Corporate Research as a Research Scientist in 2014. His current research interests include 3D object recognition and autonomous perception.

Angels Rates-Borras Angels Rates-Borras received her B.S. degree in Electrical Engineering from Universitat Politecnica de Catalunya, Barcelona, Spain. She is currently a Ph.D. student in the Department of Electrical and Computer Engineering at Northeastern University, Boston. Her research interests are mainly focused on person re-identification, scene understanding, video analysis and activity recognition.

Octavia Camps Octavia Camps received B.S. degrees in computer science in 1981 and in electrical engineering in 1984, from the Universidad de la Republica (Montevideo, Uruguay), and M.S. and Ph.D. degrees in electrical engineering in 1987 and 1992, from the University of Washington, respectively. Since 2006 she is a Professor in the Electrical and Computer Engineering Department at Northeastern University. From 1991 to 2006 she was a faculty member at the departments of Electrical Engineering and Computer Science and Engineering at The Pennsylvania State University. In 2000, she was a visiting faculty at the California Institute of Technology and at the University of Southern California and in 2013 she was a visiting faculty at the Computer Science Department at Boston University. Her main research interests include dynamics-based computer vision, image processing, and machine learning. She is a member of IEEE.

Richard J. Radke Richard J. Radke joined the Electrical, Computer, and Systems Engineering department at Rensselaer Polytechnic Institute in 2001, where he is now a Full Professor. He has B.A. and M.A. degrees in computational and applied mathematics from Rice University, and M.A. and Ph.D. degrees in electrical engineering from Princeton University. His current research interests involve computer vision problems related to human-scale, occupant-aware environments, such as person tracking and re-identification with cameras and range sensors. Dr. Radke is affiliated with the NSF Engineering Research Center for Lighting Enabled Service and Applications (LESA), the DHS Center of Excellence on Explosives Detection, Mitigation and Response (ALERT), and Rensselaer's Experimental Media and Performing Arts Center (EMPAC). He received an NSF CAREER award in March 2003 and was a member of the 2007 DARPA Computer Science Study Group. Dr. Radke is a Senior Member of the IEEE and a Senior Area Editor of *IEEE Transactions on Image Processing*. His textbook *Computer Vision for Visual Effects* was published by Cambridge University Press in 2012.