Person Re-identification in the Wild

Liang Zheng[1], Hengheng Zhang[2], Shaoyan Sun[3], Manmohan Chandraker[4], Yi Yang[1], Qi Tian[2]
[1]University of Technology Sydney [2]UTSA [3]USTC [4]UCSD
{liangzheng06,manu.chandraker,yee.i.yang,wywqtian}@gmail.com

Abstract

*This paper[1] presents a novel large-scale dataset and comprehensive baselines for end-to-end pedestrian detection and person recognition in raw video frames. Our baselines address three issues: the performance of various combinations of detectors and recognizers, mechanisms for pedestrian detection to help improve overall re-identification (re-ID) accuracy and assessing the effectiveness of different detectors for re-ID. We make three distinct contributions. First, a new dataset, PRW, is introduced to evaluate **P**erson **R**e-identification in the **W**ild, using videos acquired through six near-synchronized cameras. It contains 932 identities and 11,816 frames in which pedestrians are annotated with their bounding box positions and identities. Extensive benchmarking results are presented on this dataset. Second, we show that pedestrian detection aids re-ID through two simple yet effective improvements: a cascaded fine-tuning strategy that trains a detection model first and then the classification model, and a Confidence Weighted Similarity (CWS) metric that incorporates detection scores into similarity measurement. Third, we derive insights in evaluating detector performance for the particular scenario of accurate person re-ID.*

1. Introduction

Automated entry and retail systems at theme parks, passenger flow monitoring at airports, behavior analysis for automated driving and surveillance are a few applications where detection and recognition of persons across a camera network can provide critical insights. Yet, these two problems have generally been studied in isolation within computer vision. Person re-identification (re-ID) aims to find occurrences of a query person ID in a video sequence, where state-of-the-art datasets and methods start from predefined bounding boxes, either hand-drawn [22, 25, 37] or automatically detected [21, 45]. On the other hand, several pedestrian detectors achieve remarkable performance on benchmark datasets [12, 30], but little analysis is available on how they can be used for person re-ID.

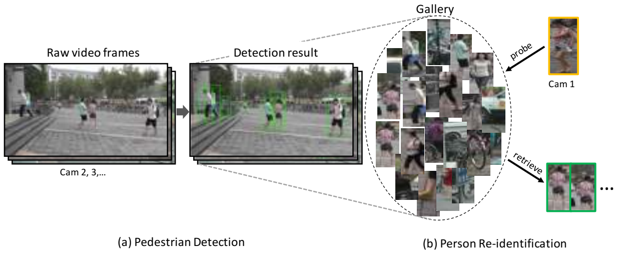


Figure 1: Pipeline of an end-to-end person re-ID system. It consists of two modules: pedestrian detection and person recognition (to differentiate from the overall re-ID). This paper not only benchmarks both components, but also provides novel insights in their interactions.

In this paper, we propose a dataset and baselines for practical person re-ID in the wild, which moves beyond sequential application of detection and recognition. In particular, we study three aspects of the problem that have not been considered in prior works. First, we analyze the effect of the combination of various detection and recognition methods on person re-ID accuracy. Second, we study whether detection can help improve re-ID accuracy and outline methods to do so. Third, we study choices for detectors that allow for maximal gains in re-ID accuracy.

Current datasets lack annotations for such combined evaluation of person detection and re-ID. Pedestrian detection datasets, such as Caltech [10] or Inria [6], typically do not have ID annotations, especially from multiple cameras. On the other hand, person re-ID datasets, such as VIPeR [16] or CUHK03 [21], usually provide just cropped bounding boxes without the complete video frames, especially at a large scale. As a consequence, a large-scale dataset that evaluates both detection and overall re-ID is needed. To address this, Section 3 presents a novel large-scale dataset called

[1] L. Zheng, H. Zhang and S. Sun contribute equally. This work was partially supported by the Google Faculty Award and the Data to Decisions Cooperative Research Centre. This work was supported in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar. This work was supported in part by National Science Foundation of China (NSFC) 61429201. Project page: http://www.liangzheng.com.cn

PRW that consists of 932 identities, with bounding boxes across $11,816$ frames. The dataset comes with annotations and extensive baselines to evaluate the impacts of detection and recognition methods on person re-ID accuracy.

In Section 4, we leverage the volume of the PRW dataset to train state-of-the-art detectors such as R-CNN [15], with various convolutional neural network (CNN) architectures such as AlexNet [19], VGGNet [31] and ResidualNet [17]. Several well-known descriptors and distance metrics are also considered for person re-ID. However, our joint setup allows two further improvements in Section 4.2. First, we propose a cascaded fine-tuning strategy to make full use of the detection data provided by PRW, which results in improved CNN embeddings. Two CNN variants, are derived *w.r.t* the fine tuning strategies. Novel insights can be learned from the new fine-tuning method. Second, we propose a Confidence Weighted Similarity (CWS) metric that incorporates detection scores. Assigning lower weights to false positive detections prevents a drop in re-ID accuracy due to the increase in gallery size with the use of detectors.

Given a dataset like PRW that allows simultaneous evaluation of detection and re-ID, it is natural to consider whether any complementarity exists between the two tasks. For a particular re-ID method, it is intuitive that a better detector should yield better accuracy. But we argue that the criteria for determining a detector as better are application-dependent. Previous works in pedestrian detection [10, 28, 43] usually use Average Precision or Log-Average Miss Rate under IoU > 0.5 for evaluation. However, through extensive benchmarking on the proposed PRW dataset, we find in Section 5 that IoU > 0.7 is a more effective rule in indicating detector influences on re-ID accuracy. In other words, the localization ability of detectors plays a critical role in re-ID.

Figure 1 presents the pipeline of the end-to-end re-ID system discussed in this paper. Starting from raw video frames, a gallery is created by pedestrian detectors. Given a query person-of-interest, gallery bounding boxes are ranked according to their similarity with the query. To summarize, our main contributions are:

- A novel large-scale dataset, Person Re-identification in the Wild (PRW), for simultaneous analysis of person detection and re-ID.
- Comprehensive benchmarking of state-of-the-art detection and recognition methods on the PRW dataset.
- Novel insights into how detection aids re-ID, along with an effective fine-tuning strategy and similarity measure to illustrate how they might be utilized.
- Novel insights into the evaluation of pedestrian detectors for the specific application of person re-ID.

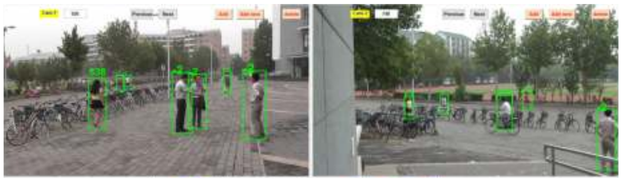


Figure 2: Annotation interface. All appearing pedestrians are annotated with a bounding box and ID. ID ranges from 1 to 932, and -2 stands for ambiguous persons.

2. Related Work

An overview of existing re-ID datasets. In recent years, a number of person re-ID datasets have been exposed [16, 20, 21, 44, 45, 48, 48]. Varying numbers of IDs and boxes exist with them (see Table 1). Despite some differences among them, a common property is that the pedestrians are confined within pre-defined bounding boxes that are either hand-drawn (*e.g*., VIPeR [16], iLIDS [48], CUHK02 [20]) or obtained using detectors (*e.g*., CUHK03 [21], Market-1501 [45] and MARS [44]). PRW is a follow-up to our previous releases [44, 45] and requires considering the entire pipeline for person re-ID from scratch.

Pedestrian detection. Recent pedestrian detection works feature the "proposal+CNN" approach. Pedestrian detection usually employs weak pedestrian detectors as proposals, which allows achieving relatively high recall using very few proposals [24, 27–29]. Despite the impressive recent progress in pedestrian detection, it has been rarely considered with person re-ID as an application. This paper attempts to determine how detection can help re-ID and provide insights in assessing detector performance.

Person re-ID. Recent progress in person re-ID mainly consists in deep learning. Several works [1, 8, 21, 40, 44] focus on learning features and metrics through the CNN framework. Formulating person re-ID as a ranking task, image pairs [1, 21, 40] or triplets [8] are fed into CNN. It is also shown in [47] that deep learning using the identification model [35, 44, 50] yields even higher accuracy than the siamese model. With a sufficient amount of training data per ID, we thus adopt the identification model to learn an CNN embedding in the pedestrian subspace. We refer readers to our recent works [47, 50] for details.

Detection and re-ID. In our knowledge, two previous works focus on such end-to-end systems. In [42], persons in photo albums are detected using poselets [4] and recognition is performed using face and global signatures. However, the setting in [42] is not typical for person re-ID where pedestrians are observed by surveillance cameras and faces are not clear enough. In a work closer to ours, Xu *et al.* [39] jointly model pedestrian commonness and uniqueness, and calculate the similarity between query and each sliding window in a brute-force manner. While [39] works on datasets

Datasets	**PRW**	CAMPUS [38]	EPFL [3]	Market-1501 [45]	RAiD [7]	VIPeR [16]	i-LIDS [48]	CUHK03 [21]
#frame	11,816	214	80	-	-	-	-	-
#ID	932	74	30	1,501	43	632	119	1,360
#annotated box	34,304	1,519	294	25,259	6,920	1,264	476	13,164
#box per ID	36.8	20.5	9.8	19.9	160.9	2	2	9.7
#gallery box	100-500k	1,519	294	19,732	6,920	1,264	476	13,164
#camera	6	3	4	6	4	2	2	2

Table 1: Comparing PRW with existing image-based re-ID datasets [3, 7, 16, 21, 38, 45, 48].

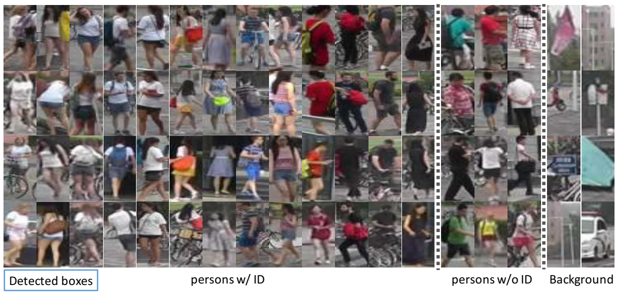


Figure 3: Examples of detected bounding boxes from video frames in the PRW dataset. In "persons w/ID", each column contains 4 detected boxes of the same identity from distinctive views. Column "persons w/o ID" presents persons who do not have an ID in the dataset. Column "background" shows false positive detection results. The detector used in this figure is DPM + RCNN (AlexNet).

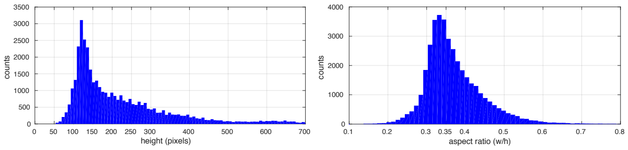
consisting of no more than 214 video frames, it may have efficiency issues with large datasets. Departing from both works, this paper sets up a large-scale benchmark system to jointly analyze detection and re-ID performance.

Finally, we would like to refer readers to [36], concurrent to ours and published in the same conference, which also releases a large dataset for end-to-end person re-ID.

3. The PRW Dataset

3.1. Annotation Description

The videos are collected in Tsinghua university and are of total length 10 hours. This aims to mimic the application in which a person-of-interest goes out of the field-of-view of the current camera for a short duration and needs to be located from nearly cameras. A total of 6 cameras were used, among which 5 are 1080×1920 HD and 1 is 576×720 SD. The video captured by each camera is annotated every 25 frames (1 second in duration). We first manually draw a bounding box for all pedestrians who appear in the frames and then assign an ID if it exists in the Market-1501 dataset. Since all pedestrians are boxed, when we are not sure about a person's ID (ambiguity), we assign -2 to it. These ambiguous boxes are used in detector training and testing, but are excluded in re-ID training and testing. Figure 2 and Figure 3 show the annotation interface and sample detected boxes, respectively.



(a) height distribution

(b) aspect ratio distribution

Figure 4: Distribution of pedestrian height and aspect ratio (width/height) in the PRW dataset.

A total of 11,816 frames are manually annotated to obtain 43,110 pedestrian bounding boxes, among which 34,304 pedestrians are annotated with an ID ranging from 1 to 932 and the rest are assigned an ID of -2. In Table 1, we compare PRW with previous person re-ID datasets regarding numbers of frames, IDs, annotated boxes, annotated boxes per ID, gallery boxes and number of cameras. Specifically, since we densely label all the subjects, the number of boxes for each identity is almost twice that of Market-1501. Moreover, when forming the gallery, the detectors produce 100k-500k boxes depending on the threshold. The distinctive feature enabled by the PRW dataset is the end-to-end evaluation of person re-ID systems. This dataset provides the original video frames along with hand-drawn ground truth bounding boxes, which makes it feasible to evaluate both pedestrian detection and person re-ID. But more importantly, PRW enables assessing the influence of pedestrian detection on person re-ID, which is a topic of great interest for practical applications but rarely considered in previous literature.

3.2. Evaluation Protocols

The PRW dataset is divided into a training set with $5,704$ frames and 482 IDs and a test set with $6,112$ frames and 450 IDs. We choose this split since it enables the minimum ID overlap between training and testing sets. Detailed statistics of the splits are presented in Table 2.

Pedestrian Detection. A number of popular pedestrian datasets exist, to name a few, INRIA [6], Caltech [10] and KITTI [13]. The INRIA dataset contains 1,805 128×64 pedestrian images cropped from personal photos; the Caltech dataset provides $\sim$350k bounding boxes from $\sim$132k frames; the KITTI dataset has 80k labels for the pedestrian class. With respect to the number of annotations, PRW ($\sim$43k

Datasets	#frame	#ID	#ped.	#ped. w/ ID	#ped. w/o ID
Train	5,134	482	16,243	13,416	2,827
Val.	570	482	1,805	1,491	314
Test	6,112	450	25,062	19,127	5,935

Table 2: Training/validation/testing split of the PRW dataset.

boxes) is a medium-sized dataset for pedestrian detection. The training and testing splits are as described above and in Table 2. Following the protocols in KITTI as well as generic object detection, we mainly use the precision-recall curve and average precision to evaluate detection performance. We also report the log-average miss rate (MR) proposed in [10]. The former calculate the average precision corresponding to ten recalls uniformly sampled from $[0, 1]$ [15], while MR is the average miss rate at 9 False Positive Per Image (FPPI) uniformly sampled from $[10^{-2}, 10^{0}]$ [10]. More statistics about the annotated pedestrians can be viewed in Fig. 4.

Person Re-identification. A good re-ID system possesses two characteristics. First, all pedestrians are accurately localized within each frame, that is, 100% recall and precision. Second, given a probe pedestrian, all instances of the same person captured by disjoint cameras are retrieved among the top-ranked results.

Re-ID is a 1:N search process. On the one hand, queries are produced by hand-drawn bounding boxes, as in practice, it takes acceptable time and effort for a user to draw a bounding box on the person-of-interest. For each ID, we randomly select one query under each camera. In total, we have 2,057 query images for the 450 IDs in the test set, averaging 4.57 (maximum 6) queries/ID. On the other hand, "N" denotes the database or gallery. A major difference between PRW and traditional re-ID datasets [7,16,21,23,45,48] is that the gallery in PRW varies with the settings of pedestrian detectors. Different detectors will produce galleries with different properties; even for the same type of detector, varying the detection threshold will yield galleries of different sizes. A good detector will be more likely to recall the person-of-interest while keeping the database small.

The IDs of the gallery boxes are determined by their intersection-over-union (IoU) scores with the ground truth boxes. In accordance to the practice in object detection, the detected boxes with IoU scores larger than 0.5 are assigned with an ID, while those with IoU less than 0.5 are determined as distractors [45]. Now, assume that we are given a query image I and a gallery $\mathcal{G}$ generated by a specific detector. We calculate the similarity score between the query and all gallery boxes to obtain a ranking result. Following [45], two metrics are used to evaluate person re-ID accuracy – mean Average Precision (mAP), which is the mean across all queries' Average Precision (AP) and the rank-1, 10, 20 accuracy denoting the possibility to locate at least one true positive in the top-1, 10, 20 ranks.

Combining pedestrian detection, we plot mAP (or rank-1, rank-20 accuracy) against the average number of detected boxes per image to present the end-to-end re-ID performance. Conceptually, with few detection boxes per image, the detections are accurate but recall is low, so a small mAP is expected. When more boxes are detected, the gallery is filled with an increasing number of false positive detections, so mAP will first increase due to higher recall and then drop due to the influence of distractors.

4. Base Components and Our Improvements

4.1. Base Components in the End-to-End System

Pedestrian detection. Recent pedestrian detectors usually adopts the "proposal+CNN" approach [5,32]. Instead of using objectness proposals such as selective search [33], hand-crafted pedestrian detectors are first applied to generate proposals. Since these weak detectors are discriminatively trained on pedestrians, it is possible to achieve good recall with very few proposals (in the order of 10). While RCNN is slow with 2000 proposals, extracting CNN features from a small number of proposals is fast, so we use RCNN instead of the fast variant [14]. Specifically, the feature for detection can be learnt through the RCNN framework by classifying each box into 2 classes, namely pedestrian and background. In this paper, three CNN architectures are tested: AlexNet [19], VGGNet [31] and ResNet [17].

Person re-identification. We first describe some traditional methods. For image descriptors, we test 6 state-of-the-art methods, namely BoW [45], LOMO [22], gBiCov [25], HistLBP [37], SDALF [11] and the IDE recognizer we propose in Section 4.2. For metric learning, the 4 tested methods are KISSME [18], XQDA [22], DVR [34] and DNS [41].

For CNN-based methods, it is pointed out in [47] that the identification model outperforms the siamese model given sufficient training data per class. In this work, the training samples per ID consist of both hand-drawn and detected boxes, and the average number of training samples per ID is over 50. So we can readily adopt the identification CNN model. Note that the training data do not include background detections due to their imbalance large number compared with the boxes for each ID. We do not apply any data augmentation. During training, a CNN embedding is learned to discriminate different identities. During testing, features of the detected bounding box are extracted from FC7 (AlexNet) after RELU, following which Euclidean distance or learned metrics are used for similarity calculation. We name the descriptor as ID-discriminative Embedding (IDE). The implementation details of IDE can be viewed in [47,50][2].

4.2. Proposed Improvements

Cascaded fine-tuning strategy. In [47], the IDE descrip-

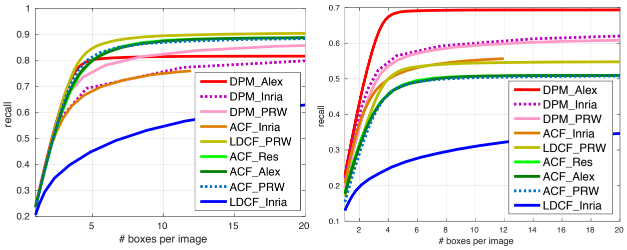
[2] github.com/zhunzhong07/IDE-baseline-Market-1501

tor is fine-tuned using the Market-1501 dataset [45] on the ImageNet pre-trained model. In this paper, we name this descriptor as IDE$_{imgnet}$ and treat it as a competing method. For the proposed cascaded fine-tuning strategy, we insert another fine-tuning step in the generation process of IDE$_{imgnet}$. That is, build on the ImageNet pre-trained model, we first train a 2-class recognition model using the detection data, *i.e.*, to tell whether an image contains a pedestrian or not. Then, we train a 482-class recognition model using the training data of PRW. The two fine-tuning process which is called "cascaded fine-tuning", results in a new CNN embedding, denoted as IDE$_{det}$. The two types of CNN embeddings are summarized below:

- **IDE**$_{imgnet}$. The IDE model is directly fine-tuned on the ImageNet pre-trained CNN model. In what follows, when not specified, we use the term **IDE** to stand for **IDE**$_{imgnet}$ for simplicity.
- **IDE**$_{det}$. With the ImageNet pre-trained CNN model, we first train an R-CNN model on PRW which is a two-class recognition task comprising of pedestrians and the background. Then, we fine-tune the R-CNN model with the IDE method, resulting in IDE$_{det}$.

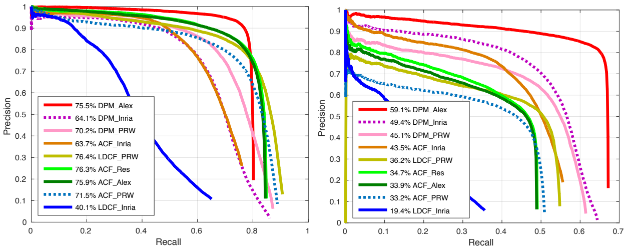
Through the cascaded fine-tuning strategy, the learned descriptor has "seen" more background training samples as well as more pedestrians (labeled as "-2") that are provided by the detection label of PRW. Therefore, the learned descriptor IDE$_{det}$ has improved discriminative ability to reduce the impact of false detections on the background. In the experiment, the performance of the two variants will be compared and insights will be drawn.

Confidence Weighted Similarity. Previous works treat all gallery boxes as equal in estimating their similarity with the query. This results in a problem: when populated with false detections on the background (inevitable when gallery gets larger with the use of detectors), re-ID accuracy will drop with the gallery size [45]. This work proposes to address this problem by incorporating detection confidence into the similarity measurement. Intuitively, false positive detections will receive lower weights and will have reduced impact on re-ID accuracy. Specifically, detector confidences of all gallery boxes are linearly normalized to $[0, 1]$ in a global manner. Then, the cosine distance between two descriptors are calculated, before multiplying the normalized confidence. Note that the **IDE feature is extracted from FC7 after RELU in AlexNet, so there are no negative entries in the IDE vector.** The cosine distance remains non-negative with IDE vectors, and is compatible with the detection scores. Currently, this baseline method supports cosine (Euclidean) distance between descriptors, but in future works, more sophisticated weightings corresponding to metric learning methods may also be considered, which should be a novel research direction in person re-ID.



(a) Recall, IoU>0.5 (b) Recall, IoU>0.7

Figure 5: Detection recall at two IoU criteria. "Inria" and "PRW" denote models trained on INRIA [6] and the proposed PRW datasets, respectively. "Alex" and "Res" denote RCNN models fine-tuned with AlexNet [19] and ResidualNet [17], respectively. For IoU>0.7, we use warm colors for detectors with higher AP, and cold colors for bad detectors. Best viewed in color.



(a) Precision-Recall, IoU>0.5 (b) Precision-Recall, IoU>0.7

Figure 6: Precision-recall curves at two IoU criteria. Detector legends are the same as Fig. 5 (Best viewed in color). The Average Precision number is shown before the name of each method.

5. Experiments

5.1. Evaluation of Pedestrian Detection

First, we evaluate the detection recall of several important detection models on PRW. This serves as an important reference to the effectiveness of proposals for RCNN based methods. These models include Deformable Part Model (DPM) [12], Aggregated Channel Features (ACF) [9] and Locally Decorrelated Channel Features (LDCF) [26]. We also test their respective RCNN versions. We retrain these models on the PRW training set and plot detection recall against average number of detection boxes per image on the testing set. The results are shown in Fig. 5. It is observed that recall is relatively low for the off-the-shelf detectors. After being retrained on PED1K dataset, LDCF yields recall of 89.3% on 11.2 proposals per image; ACF produces recall of 88.81% with 34.5 proposals per image; DPM will have a recall of 86.81% with 32.3 proposals on average. These results are collected under IoU > 0.5. When IoU increases to 0.7, detector recalls deteriorate significantly. In all, recall for the best detectors reaches around 90% for IoU > 0.5, and around 60% for IoU > 0.7.

The detection methods without RCNN mentioned above are used as proposals, and are subsequently coupled with RCNN based models. Specifically, we fine tune RCNN with three CNN models – AlexNet (Alex) [19], VGGNet (VGG) [31], and ResidualNet (Res) [17]. Additionally, we report results of the Histogram of Oriented Gradient (HOG) [6]. True positives are defined by IoU > 0.5 or IoU > 0.7. We report both the Average Precision (AP) and Log Average Miss Rate (MR). Experimental results are presented in Fig. 6. As IoU increases, detector performance deteriorates significantly which is observed in [10]. Under IoU > 0.7, the best detector is DPM+AlexNet, having an AP of 59.1%, which is +9.7% higher than the second best detector. The reason that DPM has robust performance under larger IoU is that it consists of multiple components (parts) which adapts well to pedestrian deformation, while channel feature based methods typically set aspect ratio types that are less robust to target variations. In both detection recall and detection accuracy experiments, we find that detector rankings are different from IoU > 0.5 to IoU > 0.7. With respect to detection time, it takes 2.7s, 1.4s, and 6.5s on average on a 1080×1920 frame for ACF, LDCF and DPM, respectively, using MATLAB 2015B on a machine with 16GB memory, K40 GPU and Intel i7-4770 Processor. RCNN requires 0.2s for 20 proposals.

From these benchmarking results, it is shown that the usage of RCNN effectively increases detection performance given a proposal type. For example, when using ACF as proposal, the inclusion of AlexNet increases AP from 74.16% to 76.23% (+2.07%). Further, when different CNN models are used for a given proposal, we find that ResidualNet outperforms the others in general: AP of ResNet is +0.41% higher than AlexNet.

Similar to the performance of proposals, under IoU $>$ 0.7, detector performance deteriorates significantly which is observed in [10]. For example, LDCF yields the highest recall under IoU $>$ 0.5, while it only ranks 4th under IoU $>$ 0.7. When measured under IoU $>$ 0.7, the DPM detectors are superior, probably because DPM deals with object deformation by detecting parts and adapts well to PRW where pedestrians have diverse aspect ratios (see Fig. 4(b)).

5.2. Evaluation of Person Re-identification

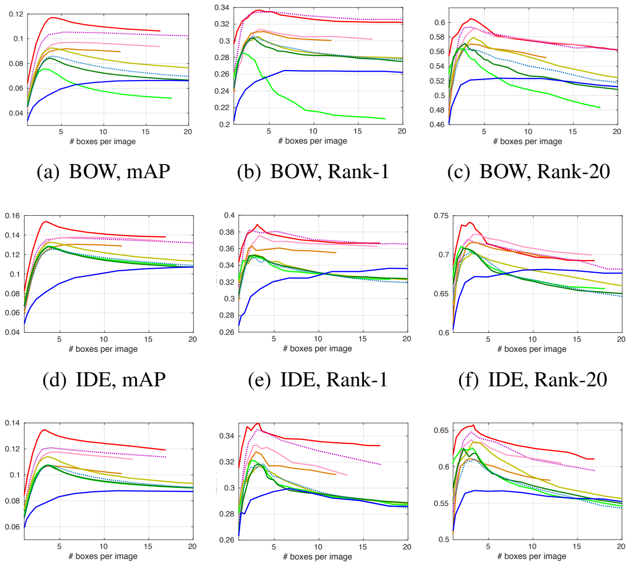
We benchmark the performance of some recent descriptors and distance metrics on the PRW dataset. Various types of detectors are used – DPM, ACF, LDCF and their related RCNN methods. The descriptors we have tested include the Bag-of-Words vector [45], the IDE descriptor described in Section 4.2, SDALF [11], LOMO [22], HistLBP [37], and gBiCov [25]. The used metric learning methods include Kissme [18], XQDA [22], and the newly proposed DNS [41]. The results are shown in Fig. 7 and Table 3.

Detector	Recognizer	#detection=3			#detection=5			#detection=10		
		mAP	r1	r20	mAP	r1	r20	mAP	r1	r20
DPM	BOW	8.9	30.4	58.3	9.7	31.1	58.6	9.6	30.5	57.7
DPM	IDE	12.7	37.2	72.2	13.7	36.9	72.1	13.7	36.6	70.8
DPM	IDE$_{det}$	17.2	45.9	77.9	18.8	45.9	77.4	19.2	45.7	76.0
DPM-Alex	SDALF+Kiss.	12.0	32.6	63.8	13.0	32.5	63.4	12.4	31.8	
DPM-Alex	LOMO+XQ.	13.4	34.9	66.5	13.0	34.1	64.0	12.4	33.6	62.5
DPM-Alex	HistLBP+DNS	14.1	36.8	70.0	13.6	35.9	67.8	12.7	35.0	65.7
DPM-Alex	IDE	15.1	38.8	74.1	14.8	37.6	71.4	14.1	36.9	69.8
DPM-Alex	IDE$_{det}$	**20.2**	**48.2**	78.1	20.3	47.4	77.1	19.9	47.2	76.4
DPM-Alex	IDE$_{det}$+CWS	20.0	**48.2**	**78.8**	**20.5**	**48.3**	**78.8**	**20.5**	**48.3**	**78.8**
ACF	LOMO+XQ.	10.5	31.5	61.6	10.5	30.9	59.5	9.7	29.7	57.4
ACF	gBiCov+Kiss.	9.8	31.1	60.1	9.9	30.3	58.3	9.0	29.0	55.9
ACF	IDE$_{det}$	16.6	44.8	75.9	17.5	43.8	76.0	17.0	42.9	74.5
ACF-Res	IDE	12.4	35.0	70.4	12.5	33.8	68.6	11.5	33.0	66.7
ACF-Alex	LOMO+XQ.	10.5	31.8	60.7	10.3	30.6	59.4	9.5	29.6	57.1
ACF-Alex	IDE$_{det}$	17.0	45.2	76.6	17.5	43.6	75.1	16.6	42.7	73.7
ACF-Alex	IDE$_{det}$+CWS	17.0	45.2	76.8	17.8	45.2	76.8	17.8	45.2	76.8
LDCF	BoW	8.2	30.1	56.9	9.1	29.8	57.0	8.3	28.3	55.3
LDCF	LOMO+XQ.	11.2	31.6	62.9	11.0	31.1	62.2	10.1	29.6	58.6
LDCF	gBiCov+Kiss.	9.5	30.7	58.8	9.6	30.1	58.4	8.8	28.7	56.7
LDCF	IDE	12.7	35.3	70.1	34.4	13.1	69.4	12.2	33.1	68.0
LDCF	IDE$_{det}$	17.5	45.3	76.2	18.3	44.6	75.6	17.7	43.8	74.3
LDCF	IDE$_{det}$+CWS	17.5	45.5	76.3	18.3	45.5	76.4	18.3	45.5	76.4

Table 3: Benchmarking results of various combinations of detectors and recognizers on the PRW dataset.

The unsupervised descriptor BoW yields decent performance on PRW dataset: around 10% in mAP and 30% in rank-1 accuracy. Improvements can be found when metric learning methods are employed. for example, when coupling SDALF and Kissme, mAP increases to 12.0% and 32.6% in mAP and rank-1 accuracy, respectively. We observe that for hand-crafted features, "HistLBP+DNS" outperforms others when built on the DPM-AlexNet detector. These results generally agree with observations in prior works [41]. We conjecture that given a fixed detector, re-ID accuracy will display similar trends as prior studies [21, 37, 41]. The IDE descriptor yields significantly higher accuracy compared with the others. For example, IDE$_{det}$ exceeds "HistLBP+DNS" by +6.2% in mAP when on average 3 bounding boxes are detected per image. This validates the effectiveness of the CNN-based descriptor. When different detectors are employed, detectors with higher AP under IoU $>$0.7 are generally beneficial towards higher overall re-ID accuracy.

The number of detected bounding boxes per image also has an impact on re-ID performance. When too few (*e.g.,* 2) bounding boxes are detected, it is highly possible that our person-of-interest is not detected, so the overall re-ID accuracy can be compromised. But when too many bounding boxes are detected, distractors may exert negative influence on the re-ID accuracy, so accuracy will slowly drop as the number of bounding boxes per image increases (Fig. 7). Nevertheless, one thing we should keep in mind is that with more bounding boxes, the timings for person retrieval also increase. Currently most works do not consider retrieval efficiency due to the small volume of the gallery. PRW, on



(a) BOW, mAP (b) BOW, Rank-1 (c) BOW, Rank-20

(d) IDE, mAP (e) IDE, Rank-1 (f) IDE, Rank-20

(g) LOMO, mAP (h) LOMO, Rank-1 (i) LOMO, Rank-20

Figure 7: Re-id accuracy (mAP, rank-1 accuracy, and rank-20 accuracy) with 9 detectors and 3 recognizers. Detector legends are the same as Fig. 5 and Fig. 6. Given a recognizer, we find that the performance of overall re-ID accuracy is more consistent with detection accuracy under IoU>0.7 than IoU>0.5, which suggests that IoU>0.7 is a better criterion for detector evaluation under the application of re-ID.

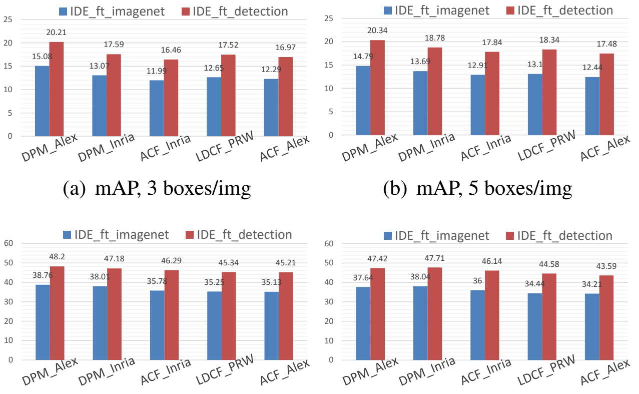
the other hand, may produce over 100k bounding boxes, so efficiency may become an important issue in future research.

5.3. Impact of detectors on re-identification

Criteria for detector assessment. How does the detector performance affect re-ID? This is a critical question in an end-to-end re-id system. Broadly speaking, a better detector would result in a higher re-id accuracy. So how to assess detector quality in the scenario of person re-ID? When only considering pedestrian detection, the community uses AP or MR defined under IOU > 0.5. In this paper, we argue that, apart from providing high recall and precision, it is of crucial importance that a detector give good localization results. Specifically, we find that IoU > 0.7 is a more effective criteria than IoU > 0.5 for detection evaluation in the scenario of person re-ID, which is the third contribution of this work.

To find how re-ID accuracy varies with detector performance, we systematically test 9 detectors (as described in Fig. 5 and Fig. 6) and 3 recognizers. The 3 recognizers are: 1) $5,600$-dimensional Bag-of-Words (BoW) descriptor [45], the state-of-the-art unsupervised descriptor, 2) $4,096$-dimensional CNN embedding feature described in Section 4.2 using AlexNet, and 3) LOMO+XQDA [22], a state-of-the-art supervised recognizer. From the results in Fig. 7 and Table 3, a key finding is that *given a recognizer, the re-ID performance is consistent with detector performance evaluated using the IoU* > 0.7 *criterion*. In fact, if we use the IoU > 0.5 criterion as most commonly employed in pedestrian detection, our study shows that the detector rankings do not have accurate predictions on re-ID accuracy. Specifically, while the "DPM_Alex" detector ranks 4th in average precision (75.5%) with the IoU > 0.5 rule, it enables superior re-ID performance which is suggested in its top ranking under IoU > 0.7. The same observations hold for the other 8 detectors. This conclusion can be attributed to the explanation that under normal circumstances, a better localization result will enable more accurate matching between the query and gallery boxes. As an insight from this observation, when a pool of detectors is available in a practical person re-ID system, a good way for choosing the optimal one is to rank the detectors according to their performance under IoU > 0.7. In addition, recent research on partial person re-ID [49] may be a possible solution to the problem of misalignment.

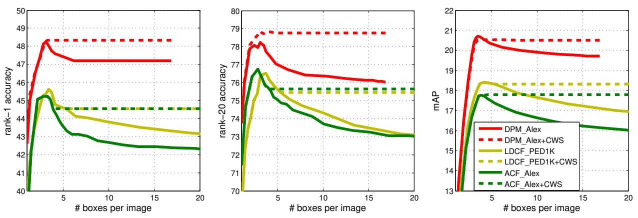
Effectiveness of cascade fine-tuning. This paper introduces two IDE variants. For the first variant, we fine-tune IDE directly from AlexNet pre-trained on ImageNet, denoted as IDE$_{imgnet}$. For the second variant, we first fine-tune a pedestrian detection model (2 classes, pedestrian and background) from AlexNet pre-trained on ImageNet, and then we further fine tune it using the identification model on PRW. We denote the second variant as IDE$_{det}$, which is the learned embedding by the cascaded fine-tuning method. Experimental results related to the IDE variants are presented in Table 3 and Fig. 8.



(a) mAP, 3 boxes/img (b) mAP, 5 boxes/img

(c) rank-1 acc., 3 boxes/img (d) rank-1 acc., 5 boxes/img

Figure 8: Plots of mAP and rank-1 accuracy using two variants of the IDE with 5 detectors. Fine-tuning on the pedestrian-background detection model improves over fine-tuning on the Imagenet model, proving the effectiveness of the proposed cascaded fine-tuning method.



(a) rank-1, CWS (b) rank-20, CWS (c) mAP, CWS

Figure 9: Effectiveness of the proposed Confidence Weighted Similarity (CWS) on the PRW dataset. We test three detectors and the IDE$_{det}$ descriptor fine-tuned on the pedestrian-background detection model. We find that CWS reduces the impact of distractors when the number of detected bounding boxes increases.

Two major conclusions can be drawn from the above experiments. First, we observe that the accuracy of IDE is superior to that of hand-crafted descriptors (in accordance with [47]), and is further improved in combination with state-of-the-art metric learning schemes. Second, it is noticeable from Fig. 8 that IDE$_{det}$ yields considerably higher re-ID accuracy than IDE$_{imgnet}$. Specifically, when using the DPM detector trained on INRIA dataset and considering 3 detection boxes per image, IDE$_{det}$ results in $+4.52\%$ and $+9.17\%$ improvement in mAP and rank-1 accuracy, respectively. Very similar improvements can be observed for the other 4 detectors and using 5 detection boxes per image. This indicates that when more background and pedestrian samples are "seen", the re-ID feature is more robust against outliers. This illustrates that the proposed cascaded fine-tuning method is effective in improving the discriminative ability of the learned embeddings. In fact, a promising direction is to utilize more background and pedestrian samples without ID that are cheaper to collect in order to pre-train the IDE model. Experiment of the two IDE variants provides one feasible solution of how detection aids re-ID.

Effectiveness of Confidence Weighted Similarity (CWS) We test the CWS proposed in Section 4.2 on the PRW dataset with three detectors and the IDE$_{det}$ descriptor. The results are shown in Fig. 9. The key observation is that CWS is effective in preventing re-ID accuracy from dropping as the number of detections per image increase. As discussed before, more distractors are present when the database get larger and CWS addresses the problem by suppressing the scores of false positive results. In Table 3, the best results on the PRW dataset are achieved when CWS is used, which illustrates the effectiveness of the proposed similarity. We will extend CWS to include metric learning representations in the future work.

Figure 10 presents some sample re-ID results. For the failure case in row 3, the reason is that too many pedestrians are wearing similar clothes. For row 4, the query is cropped by the camera, leading to compromised pedestrian matching.

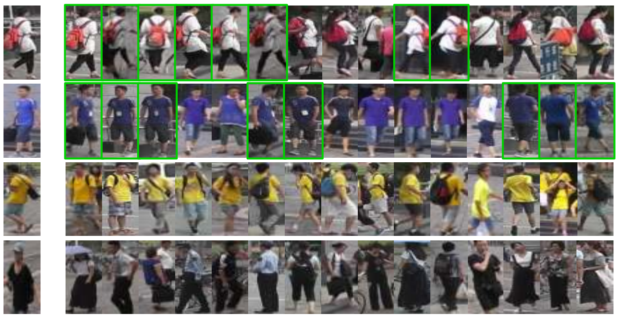


Figure 10: Sample re-ID results on the proposed PRW dataset with the DPM_Alex detector and the proposed IDE descriptor. Rows 1 and 2 are success cases, while Rows 3 and 4 are failure cases due to similar clothing and truncation, respectively. With truncated queries, partial re-ID methods [49] might be especially important.

6. Conclusions and Future Work

We have presented a novel large-scale dataset, baselines and metrics for end-to-end person re-ID in the wild. The proposed PRW dataset has a number of features that are not present in previous re-ID datasets, allowing the first systematic study of how the interplay of pedestrian detection and person re-ID affects the overall performance. Besides benchmarking several state-of-the-art methods in the fields of pedestrian detection and person re-ID, this paper also proposes two effective methods to improve the re-ID accuracy, namely, ID-discriminative Embedding and Confidence Weighted Similarity. For IDE, we find that fine-tuning an R-CNN model can be a better initialization point for IDE training. Further, our extensive experiments serve as a guide to selecting the best detectors and detection criteria for the specific application of person re-ID.

Our work also enables multiple directions for future research. First, it is critical to design effective bounding box regression schemes to improve person matching accuracy. Second, given the baseline method proposed in this paper to incorporate detection confidence into similarity scores, more sophisticated re-weighting schemes can be devised. This direction could not have been enabled without a dataset that jointly considers detection and re-ID. In fact, re-ranking methods [2, 46, 51] will be critical for scalable re-ID. Third, while it is expensive to label IDs, annotation of pedestrian boxes without IDs is easier and large amounts of pedestrian data already exist. According to the IDE results reported in this paper, it will be of great value to utilize such weakly-labeled data to improve re-ID performance. Finally, effective partial re-ID algorithms [49] can be important for end-to-end systems on the PRW dataset (Fig. 10).

References

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

[2] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017.

[3] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *TPAMI*, 33(9):1806–1819, 2011.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[5] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

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

[8] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *TPAMI*, 36(8):1532–1545, 2014.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34(4):743–761, 2012.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[14] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[16] 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*, volume 3, 2007.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[18] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

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

[20] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[21] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[22] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[23] S. Liao, Z. Mo, Y. Hu, and S. Z. Li. Open-set person re-identification. *arXiv:1408.0872*, 2014.

[24] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, 2014.

[25] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *IVC*, 32(6):379–390, 2014.

[26] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved pedestrian detection. In *NIPS*, 2014.

[27] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012.

[28] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013.

[29] W. Ouyang, X. Zeng, and X. Wang. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, 2013.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

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

[32] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, 2015.

[33] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[34] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[35] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[36] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. CVPR, 2017.

[37] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV 2014*. 2014.

[38] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013.

[39] Y. Xu, B. Ma, R. Huang, and L. Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM MM*, 2014.

[40] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[41] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[42] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, 2015.

[43] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *CVPR*, 2016.

[44] 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 *ECCV*, 2016.

[45] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *CVPR*, 2015.

[46] 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 *CVPR*, 2015.

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

[48] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, volume 2, page 6, 2009.

[49] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong. Partial person re-identification. In *ICCV*, 2015.

[50] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv:1701.07717*, 2017.

[51] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *arXiv preprint arXiv:1701.08398*, 2017.