A Comprehensive Study on Large-Scale Person Retrieval in Real Surveillance Scenarios

Da Li[1,2], Zhang Zhang[1,2], Caifeng Shan[3], Liang Wang[1,2], Tieniu Tan[1,2]
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[2] Center for Research on Intelligent Perception and Computing (CRIPAC), CASIA
[3] Artificial Intelligence Research, CAS (CAS-AIR)

Abstract

Person retrieval is a hot research topic due to its important application potential for public security. Though existing algorithms have achieved impressive progresses on current public datasets, it is still a challenging task in the real surveillance scenarios due to the various viewpoints, pose variations and occlusions. Moreover, few of the existing works study the problem of person retrieval on large-scale gallery set, where lots of distractions may deteriorate the retrieval results heavily. To have a deep understanding on the above challenges, we perform a comprehensive study on current state-of-the-art person retrieval algorithms with a large-scale benchmark in real surveillance scenarios. In the study, two kinds of techniques, i.e., attribute recognition and person re-identification, including eight algorithms, are evaluated at both algorithm level and system level. Here, the system-level evaluations investigate the effects of the combinations of the above algorithms with the module of person detection, where lots of distractions in person detection results pose a big challenge for person retrieval in real scenes. Extensive evaluations with large gallery sizes (up to 243k) and comprehensive analyses are presented in the study, which will guide researchers to develop more advanced algorithms in future.

1. Introduction

Person retrieval aims at searching the person(s) of interest from the captured large-scale surveillance image/video data with specific visual attributes or person images. It is an important application for public security, such as to find the criminal suspects and lost elderly or children. Depending on the different query conditions, person retrieval can be classified into the attribute-based retrieval and the image-based retrieval, in which attribute recognition and person re-identification (ReID) are the two core techniques.

Benefiting from the release of large-scale datasets, such as PA-100k [23] and PETA [4] for the attribute-based retrieval, Market1501 [47] and DukeMTMC-reID [28] for the image-based retrieval, and RAP [17] for the both categories, as well as the success of deep learning models, the performance of person retrieval has been improved significantly. However, it is still a challenging task to obtain satisfying results in real surveillance scenes due to the large gallery size, low image quality, various viewpoints, large pose variations and occlusions. Though, these issues encourage the researchers to develop more robust algorithms, the main limitations of existing works can be summarized as follows.

First, existing works usually pay attention to improve the retrieval performances with a limited gallery size, few of them explore the influences of the large-scale gallery set with lots of distractions which is a common situation in real surveillance scenarios. These distractions may deteriorate the retrieval results heavily. The dataset of Market1501 [47] provides a large-scale gallery set with 500k distractions. However, most of these distractions are easy samples, *e.g.* background, which do not accord with the real situations.

Second, the algorithms for attribute recognition and person ReID are often developed independently with self input-output definitions, train/test dataset and evaluation metrics. However, a real person retrieval system usually consists of several coherent vision algorithms, such as person detection and attribute recognition or person ReID. Therefore, its performance depends on all vision algorithms in the system execution pipeline. Some recent studies on person search [41, 49, 39, 1] start to consider the person detection and ReID as a unified system. They are usually developed and evaluated on the datasets with both the person location and identity are annotated. However, the gallery size in real applications is always very large so that is hard to annotate all the samples in such large-scale gallery set. How to evaluate the performance with partial ground truth is still an open issue [48]. Venetianer and Deng [36] provide a fuzzy ground-truth based evaluation approach for both system-level and component-level evaluation. However, this work only showed a conceptual framework without practical test in a real system.

To have a deep understanding on the above problems, we perform a comprehensive study on current state-of-the-art person retrieval algorithms with a large-scale benchmark in real surveillance scenarios. These algorithms are collected from the Large-Scale Person Retrieval Challenge (LSPRC) which is held based on the proposed benchmark. In the study, two kinds of techniques, i.e., attribute recognition and person ReID, totally eight algorithms, are evaluated at both algorithm level and system level. Different from the existing works, the system-level evaluations investigate the effects of the combinations of the above algorithms with the module of person detection, where lots of distractions in person detection results pose a big challenge in person retrieval in real scenes. Thus, the contributions of this paper include following aspects:

- We present a large-scale benchmark for person retrieval in real surveillance scenarios with standard dataset and metrics, in which the dataset is a subset of RAP [17] and is named RAP-LSPRC in this paper.
- We propose a system-level evaluation paradigm to measure the retrieval performance through combining the algorithms for attributes recognition or ReID with the real person detector as a unified system so that the effects of lots of detected distractions are considered.
- We give a comprehensive study for eight state-of-the-art person retrieval algorithms, which are the submissions in LSPRC, on the proposed benchmark. The results of the extensive evaluations will guide researchers to develop more advanced algorithms in future.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 introduces the proposed benchmark. Section 4 overviews the eight algorithms submitted to LSPRC. Section 5 shows the experimental results and corresponding analysis. Finally, we conclude this work in Section 6.

2. Related Work

2.1. Overview on Person Retrieval

For the attribute-based person retrieval, modern methods are usually developed with the Deep Convolutional Neural Network (DCNN) [32, 14, 16], Recurrent Neural Network (RNN) [38], and attention model [23] *etc*. And some of them improve the performance through exploring the human pose information [16] or the context information among attributes [38, 23].

For the image-based person retrieval, representation learning and metric learning are the two core research contents. To learn robust representation, many efforts mainly focus on alleviating the problem of misalignment across different viewpoints. In [33, 37], the person image is partitioned into several rigid horizontal stripes to learn the local features. Some other work learns the aligned deep features using human parts [15, 46], pose [31], mask [30, 13], or even dense semantic alignment of the human body [45]. For the metric learning, to keep the features of the same identity closer while push the features of different identities further apart, modern methods [3, 37] usually utilize both the classification loss and ranking loss to optimize the network.

Though the performances of these methods have achieved much progress on current public datasets, most of them only consider the gallery set with predefined bounding boxes. While these bounding boxes are usually generated by person detectors in the real applications. Meanwhile, few of the above methods take the impacts of large-scale gallery set with lots of distractions into account.

The task of person search, which is first proposed in [41], turns to consider the person detection and ReID as a coherent system. Some work [39, 22] jointly learns detection and identification features in an end-to-end manner; others [49, 1] optimize the detection and ReID separately. Different from the work of common person ReID and person search, open-world person ReID [20, 50, 53, 19] studies the person verification problem, in which not all the query identities are appeared in the gallery set. Though the tasks of person search and open-world ReID are close to the real applications, neither of them proposes a proper evaluation metric when only partial samples in the large-scale gallery set are annotated.

2.2. Related Benchmarks And Challenges

Benchmarks. The large-scale annotated datasets with standard evaluation metrics play an important role to develop robust algorithms. APiS [52], PETA [4] and PA-100k [23] are three popular datasets for attribute recognition. They are collected with different number of samples and kinds of attributes. CUHK03[18], Market1501 [47] and DukeMTMC-reID [28] are the three most popular datasets for person ReID. They are collected with different number of identities across multiple cameras in various scenes. However, all of these datasets only consider the samples with pre-defined bounding boxes. CUHK-SYSU [39] and PRW [49] are two datasets for person search, in which both the person location and identity are annotated. Thus, the effects of person detection can be considered when developing and evaluating the algorithms for person ReID. However, it is much labor-intensive to annotate such large-scale datasets.

Challenges. The related challenges also promote the researchers to improve the performances of corresponding algorithms. The challenge of Semantic Person Retrieval in Surveillance Using Soft Biometrics [7] is run in AVSS 2018, whose purpose is to locate the interesting target with the semantic attributes. It only released a small size of data to the participants. The person search track in WIDER Challenge [25] is another challenge for person retrieval held in ECCV 2018, which is to search a given portrait from candidate movies. While this challenge is not for the surveillance scenarios. Moreover, neither of the two challenges considers the situation with large-scale gallery set.

3. Proposed Benchmark

3.1. Dataset

A subset of samples is selected from the richly annotated pedestrian (RAP) dataset [17] to construct the proposed benchmark which includes more than 68 thousands pedestrian images annotated with 72 fine-grade attributes and 2,589 identities (IDs). It is named RAP-LSPRC in this paper. Some basic information is listed in Table 1. And several samples in RAP-LSPRC are shown in Fig. 1.

Moreover, 39,278 complete frames corresponding to the test sets for attribute recognition and person ReID are also provided to execute system-level evaluation where a person detection algorithm is firstly perform to construct a large-scale gallery set. Thus, lots of detection results in the gallery set would act as distractions.

Attribute				Re-Identification		Other Information	
#Training Samples	#Validation Samples	#Test Samples	#Attri.	#Training IDs (samples)	#Test IDs (samples)	Resolution ($w \times h$)	#Cam.
33268	8317	25986	72	1295 (13178)	1294 (13460)	from 33×81 to 415×583	25

Table 1. Basic information of RAP-LSPRC. *Attri.* is short for *Attributes* and *Cam.* is short for *Cameras*.

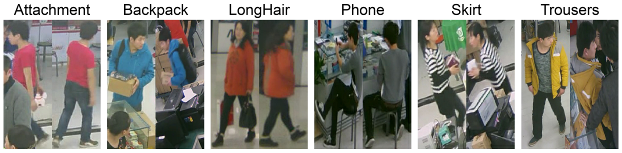


Figure 1. Image samples in RAP-LSPRC. It shows some samples with different IDs and attributes under various camera viewpoints, body part occlusions, human poses, and image qualities.

3.2. Tasks

Based on the different querying conditions, two kinds of tasks, *i.e.*, attribute-based person retrieval (PR-A) and image-based person retrieval (PR-ID), are considered in the proposed benchmark.

3.2.1 Task of PR-A

The task of PR-A is to search the interesting person with one or multiple attributes. To complete this task, attribute recognition is utilized to obtain the likelihoods of predefined attributes which are further used to calculate the ranking list. PR-A consists of two stages, *i.e.*, PR-A-RAP and PR-A-SYS, based on the two kinds of evaluation metrics.

PR-A-RAP. A list of query conditions are generated with different number of attributes (ranging from 1 to 4). Under different query conditions, the samples in the gallery set are ranked with the likelihoods of desirable attributes. And the ranking results are further evaluated using the ground truth which is provided in the RAP-LSPRC dataset.

PR-A-SYS. Different pipelines (algorithm combinations) are generated through combing the algorithms for attribute recognition with a specific person detector. The complete frames, which are corresponding to the gallery set, are parsed with these pipelines. Inspired by the work [6] and [26], a question-answering (QA) paradigm is designed to evaluate the performance of person retrieval system. As shown in Table 2, more than 5 million polar (binary) queries (whether a person with the specified attribute(s)) are generated based on the ground truth of the test set in RAP-LSPRC. The performances of the QA results reflect the performances of the submitted algorithms integrated into the person retrieval system.

3.2.2 Task of PR-ID

The task of PR-ID is to search the interesting person through measuring the similarities between its ReID features and those of the samples in gallery set. The Euclidean distance is restricted to measure the similarities. And then these similarities are used to generate the ranking list. Similar to the task of PR-A, it also consists of two stages, *i.e.*, PR-ID-RAP and PR-ID-SYS, based on the two kinds of evaluation metrics.

PR-ID-RAP. A subset of samples in the test set is selected as the query set; the others are selected as gallery set. For each image in the query set, the similarities with all the images in the gallery set are calculated based on the ReID features. Then, sort all the images in the gallery set descending with the similarities to generate the ranking list. The ranking results are also evaluated using the ground truth which is provided in the RAP-LSPRC dataset.

PR-ID-SYS. Similar with PR-A-SYS, the algorithms for ReID are combined with a specific person detector to generate different pipelines. The complete frames corresponding to both the query and gallery set in RAP-LSPRC are parsed with these pipelines. Then, the kNN-search is conducted for all the detected persons to get their top-k (k is set to 100 in this paper) neighbours using ReID features. A relationship graph is constructed with the results of kNN-search, where each node represents a person, and the edge reflects if the two nodes belong to the same ID. Meanwhile, more than 17 million polar queries (whether the two persons with the same ID or not) are generated based on the ground truth of the test set in RAP-LSPRC (see Table 2). They are used to measure the performance of the person retrieval system.

3.3. Evaluation Metrics

Two kinds of evaluations, *i.e.*, algorithm-level evaluation and system-level evaluation, are adopted in this benchmark.

3.3.1 Algorithm-Level Evaluation

Given a query image or condition with different attributes, a ranking list is generated with the similarities/likelihoods which are calculated using the trained models. Thus, the algorithm-level evaluation is to measure the ranking results with the ground truth of test set in the RAP-LSPRC dataset. The mean average precision (mAP) is utilized to evaluate the performances of the two kinds of tasks. For a query, the average precision (AP) is defined in Eq. 1.

$$AP = \frac{\sum_{r=1}^{N} P(r) \times \mathbb{1}(r)}{K} \tag{1}$$

where N is the number of samples in the gallery set; $P(r)$ is the precision at rank r; $\mathbb{1}(r)$ is an indicator function that represents whether the sample at rank r is related to current query; and K is the number of related samples. The mAP is further calculated through averaging all the values of AP under different query conditions. It is worth noting that only the samples locating in the different cameras with the query image are considered in Eq. 1.

3.3.2 System-Level Evaluation

For the system-level evaluations, the Faster-RCNN [27] is chosen as the default person detector to combine with the algorithms for attribute recognition or person ReID, in which the threshold of Intersection over Union (IoU) is set to 0.5 (IoU>0.5). Thus, a large-scale gallery set including 243,353 samples are collected, in which only partial samples are annotated and lots of distractions are included. To evaluate the person retrieval system with such a gallery set, we adopt a QA paradigm which is inspired by two work [6, 26] on visual Turing test for computer vision systems. Thus, millions of polar queries, as shown in Table 2, are designed based on the annotation information of RAP-LSPRC dataset, where the *positive query* means the correct answer is *yes*; and the *negative query* means the correct answer is *no*. Meanwhile, the *non-annotated query* is generated with the non-labeled samples so that has no ground truth answer.

PR-A-SYS			PR-ID-SYS		
# Positive Queries	# Negative Queries	# Non-annotated Queries	# Positive Queries	# Negative Queries	# Non-annotated Queries
354,700	4,972,430	44,797,420	491,518	17,094,390	1,540,523,194

Table 2. Number of queries used in PR-A-SYS and PR-ID-SYS.

PR-A-SYS. Eq. 2 and Eq. 3 show how to evaluate the performance of PR-A-SYS quantitatively, in which the F_1 score is calculated as follows:

$$prec_i = \frac{N_{tp}^i}{N_{tp}^i + N_{fp}^i}, \quad recall_i = \frac{N_{tp}^i}{N_p^i} \tag{2}$$

$$F_{1i} = \frac{2 * prec_i * recall_i}{prec_i + recall_i}, \quad F_1 = \frac{1}{n_c}\sum_{i=1}^{n_c} F_{1i} \tag{3}$$

where i represents the ith query condition; n_c is the total number of query conditions; N_{tp}^i, N_{fp}^i and N_p^i are the number of true positives, the number of false positives and the number of positives (the designed queries whose correct answers are yes) respectively under the ith query condition.

PR-ID-SYS. The F_1 score is also utilized to evaluate the performance of PR-ID-SYS as shown in Eq. 4 and Eq. 5:

$$prec = \frac{N_{tp}}{N_{tp} + N_{fp}}, \quad recall = \frac{N_{tp}}{N_p} \tag{4}$$

$$F_1 = \frac{2 * prec * recall}{prec + recall} \tag{5}$$

where N_{tp}, N_{fp} and N_p are the total number of true positives, the total number of false positives and the total number of positives respectively. As only partial samples are annotated in the gallery set, we determine the true positives and false positives as follows:

$$f(Q_{qg}) = \begin{cases} tp, & \text{if } Q_{qg} \text{ is positive and } R_q(g) \leq r_p \\ fp, & \text{if } Q_{qg} \text{ is negative and } R_q(g) \leq r_n \end{cases} \tag{6}$$

$$s.t. \;\; r_n < r_p$$

where Q_{qg} is a query (a pair of persons), in which “q” and “g” represent the person images selected from the query set and gallery set in RAP-LSPRC respectively; tp represents the true positive; fp represents the false positive; $R_q(g)$ represents the rank of “image g” in the ranking list of “image q”; r_p and r_n are the two thresholds ($r_p = 100$ and $r_n = 10$) that are obtained empirically.

4. Methods

The Large-Scale Person Retrieval Challenge (LSPRC) is held based on the proposed benchmark, to which more than fifteen valid algorithms are submitted. In this paper, eight representative algorithms of them are selected to have a comprehensive study.

All the submissions for the LSPRC are restricted to train without using extra data. And the data in RAP-LSPRC for one task can NOT be utilized to another as well.

4.1. Attribute Recognition Methods

Attribute recognition is the core technique for the task of PR-A. This section summarizes the top-3 submissions for this task (see Table. 3). From the submissions, we can find that all of them are based on deep architecture with different backbone networks and loss functions. Meanwhile, we can also find that the attention mechanism and learning the context information among different attributes are two popular methods to improve the performance.

4.2. Person ReID Methods

Person ReID is the core technique for the task of PR-ID. This section shows an overview on the top-5 submissions (see Table. 4). The deep learning based methods dominate this field as well. The characteristics of the top-5 submissions can be summarized as follows: 1) the ResNet [8] and DenseNet [11] with various numbers of layers are the most two popular networks that are utilized as the backbone; 2) multi-model fusion is widely used to generate the final features through concatenating the output of each model; 3) MGN [37] is a popular framework that are utilized to capture both the global and local features; 4) the feature maps from different layers are considered in some of the submissions to capture more information of different levels; 5)

Submissions	Characteristics	Multi-model Fusion (# Models)	Backbone Network	Loss Function
1	1. Design an attention-mechanism guided network. 2. Input images are resized to 128×256. 3. Data augmentation: random rotation and cropping. 4. The initial learning rate is 0.01 which is updated with SGDR [24].	No	ResNet [8]	Focal loss [21]
2	1. Design a network including four separate branches, and each branch is responsible for partial attributes. 2. Three models are trained with three kinds of ways to data augmentations. 3. Nadam [5] is utilized as the optimizer. 4. The input images are resized to 224×224×3.	Yes (3)	DenseNet [11]	Cross-Entropy Loss
3	1. Propose a network which consists of three components: Main Net, View Prediction Net (VPN) [29] and Spatial Regularization Net (SRN) [51]. 2. The VPN is trained to classify the viewpoints (front, back and side). 3. The SRN is utilized to learn the spatial context information among different attributes. 4. Main Net predicts the attributes from three viewpoints respectively, in which the outputs are weighted by the results of VPN; and it is further aggregated with the results of SRN to generate the final predictions.	No	GoogleNet Inception [35]	Cross-Entropy Loss

Table 3. A summary on the top-3 submissions of task PR-A.

multiple losses are usually combined to train the network which could introduce more supervised information.

5. Experimental Results

5.1. Setups

For the task of PR-A, 205 query conditions are generated with 54 attributes which involve gender, age, clothing, action and *etc*. The number of attributes for one query condition is ranging from 1 to 4. And we choose 0.5 as the classification threshold.

For the task of PR-ID, 7,202 query person images are randomly selected from the test set as shown in Table 1. And to count the number of true positives and false positives in PR-ID-SYS, the values of r_p and r_n in Eq. 6 are set to 100 and 10 respectively. The two thresholds are determined based on a hypothesis that the numbers of positives in the large-scale gallery set for most query person images are far less than 100. To verify its validity, 100 query person images are randomly selected to check the ranking results with the detected gallery set whose size is more than 243k. The statistical result is shown in Table 5. We can find that the count of query images, whose number of positives is larger than 80 in top-100 , is only 7 (%7). It indicates the validity of the hypothesis.

# Positives	[0, 10]	(10, 50]	(50, 80]	(80, 100]
# Query Images	44	42	7	7

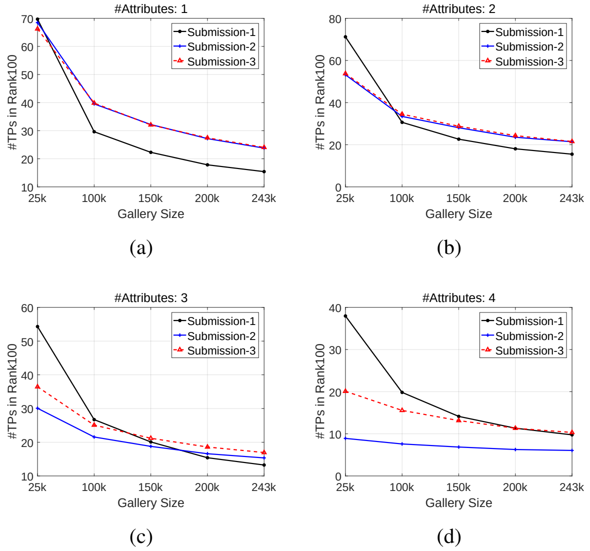
Table 5. The numbers of positives in the top-100 for 100 randomly selected query person images. And the average number is 23.74.

5.2. Results of PR-A

Methods	mAP (PR-A-RAP)	F_1 score (PR-A-SYS)
1	**0.4220**	**0.4135**
2	0.4107	0.2042
3	0.3512	0.2455

Table 6. The evaluation results of the top-3 submissions of PR-A.

Algorithm-Level Evaluation. The second column of Table 6 shows the evaluation results of PR-A-RAP about the top-3 submissions. We can find that the methods with complicated backbone networks (*e.g*. submission-1 with ResNet and submission-2 with DenseNet *vs*. submission-3 with GoogleNet) incline to generate superior performances (mAP) on the test set with predefined bounding boxes. And the data unbalance of different attributes is considered by using of focal loss in submission-1, which improves the performance further and makes it the first place.



(a)

(b)

(c)

(d)

Figure 2. Number of true positives in rank-100 under different gallery size. (a)∼(d) show the changing trends with different number of attributes.

System-Level Evaluation. The last column of Table 6 shows the evaluation results of PR-A-SYS about the top-3 submissions. The lower quality of the detected samples, such as the parts missing and position shifting in the bounding box, makes them much harder to recognize than the ideal samples. So the strategies, such as attention mechanism and learning the context information among different attributes, play the key role to improve the performance (F_1). It leads the results of submission-3 superior to that of submission-2. Meanwhile, the effect of different gallery sizes on the ranking results is also studied. We count the number of true positives in the rank-100. The statistical re-

Submissions	Characteristics	Multi-model Fusion (# Models)	Backbone Network	Loss Function
1	1. The basis of this method is MGN [37] with different backbone networks. 2. The main differences compared with the original MGN:1)the triplet loss is removed; 2)the Part-4 branch is introduced and the feature maps are split unevenly; 3) the layers of RELU, drop and fc are removed in the bottleneck. 3. Five models are trained with three partition strategies combined with different backbone networks; and the SGD is selected as the optimizer. 4. The input images are resized to 256×512.	Yes (5)	SEResNet152 [10], ResNet152 [8], DenseNet201 [11]	Softmax Loss
2	1. MGN [37], PCB [33] and DenseNet [11] are utilized to learn the features. 2. Three aspects are considered to improve the performance: 1)focal loss and adaptive data augmentation are introduced to alleviate the problem of data unbalance 2)a strategy named Bi-directional Feature Flow (BFF) is proposed to process the images with low quality. 3)the method of Aligned-ReID [44] is integrated into MGN to improve the feature alignment. 3. The input images are resized to 128×384.	Yes (5)	ResNet50 [8], ResNet101 [8], DenseNet129 [11], SEResNet50 [10]	Softmax Loss, Triplet Loss, Focal Loss [21]
3	1. MGN [37] and PCB [33] are used to learn the local features. 2. The method [42] is utilized to learn the global features. 3. The data is augmented through the method in [12].	Yes (4)	ResNet50 [8], ResNetXt50 [40], DenseNet121 [11], inception_ReseNet_v2 [34]	Triplet Loss, Quadruplet Loss [2], Classification Loss
4	1. The method of Multi-Branch Tree Embedding (MTE) [43] is proposed to improve the performance of feature learning. 2. The method of Hierarchical Deep Learning Feature (HDLF) [43] is adopted to fuse the features from multiple layers. 3. The method of Gaussian Difference Quadratic Subspace (GDQS) [43] learning is utilized to learn the discriminative subspace for the cascaded features efficiently.	Yes (3)	ResNet50 [8], MobileNet [9], DenseNet161 [11]	Softmax Loss
5	1. It is a MGN-like framework with four branches, and the feature maps are divided into different numbers of parts in each branch respectively. 2. Different to MGN[37], the weights of all the convolutional layers are shared across different branches. 3. The data is augmented with random cropping and flipping.	No	ResNet152 [8]	Cross-Entropy Loss, Focal Loss [21]

Table 4. A summary on the top-5 submissions of task PR-ID.

sults with different number of attributes are shown in Fig. 2. They are consistent with the expectations that the number of true positives decreases along with the increasing of gallery size as more distractions occupy the places in the rank-100. We can also find that the decline of submission-1 is more obvious than the other two methods. It is the higher recall (0.4889 for submission-1 *vs.* 0.1592 for submission-2 and 0.2045 for submission-3) and lower precision (0.3792 for submission-1 *vs.* 0.6080 for submission-2 and 0.4855 for submission-3) of submission-1 makes this method is prone to pull the distractions into rank-100.

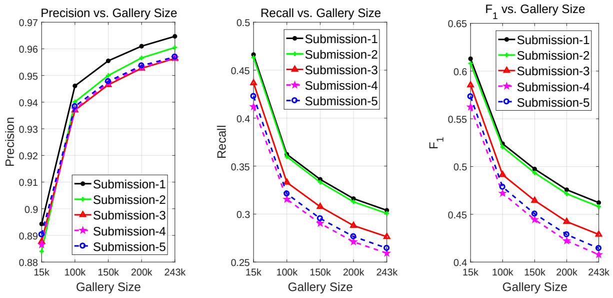
5.3. Results of PR-ID

Algorithm-Level Evaluation. The second column of Table 7 lists the evaluation results of PR-ID-RAP about the top-5 submissions. Except for the submission-5 in the 5th place, the multi-model fusion is utilized by all the other four submissions. And the number of models increases from the 4th place to the 1st place (3 to 5). Moreover, the backbone network becomes more complicated as higher ranking performance, *e.g.* SEResNet152, ResNet152 and DenseNet201 are used by submission-1 in the 1st place. The two folds indicate that the complicated frameworks cline to achieve better performance (mAP).

System-Level Evaluation. The last column of Table 7 lists the evaluation results of PR-ID-SYS about the top-5 submissions. It obtains the similar ranking with PR-ID-RAP except that the performance of submission-5 exceeds the result of submission-4. Comparing their frameworks, we can find that all of them are based on the MGN [37] except for submission-4. MGN can concentrate on both the global and local discriminative clues. It can improve the robustness of the extracted features especially when the input image with low quality, *e.g.* the bounding box image is generate with real person detectors. So submission-5 achieves superior performance than submission-4 in PR-ID-SYS. The effect of different gallery sizes on the performances of PR-ID-SYS is further studied. We can find from Fig. 3(b) that the values of recall decrease along with the

Methods	mAP (PR-ID-RAP)	F_1 score (PR-ID-SYS)
1	**0.7335**	**0.5286**
2	0.7087	0.5263
3	0.6421	0.5030
4	0.6059	0.4802
5	0.5933	0.4906

Table 7. The evaluation results of the top-5 submissions of PR-ID. We get the F_1 score with the gallery size is 83,622.



(a) (b) (c)

Figure 3. The values of precision, recall and F_1 score with different gallery size. Not only the backgrounds but also the person images with different identities to the query one consist of the distractions in the galley set.

enlarging of gallery set. It is mainly because more distractions occupy the places in top-r_p. Thus fewer true positives can appear at those positions. It is the same reason that fewer false positives can locate in the top-r_n which improve the value of precision (see Fig. 3(a)). However, the decline of the recall is more significant than the improvement of the precision. It results in the decline of F_1 scores finally.

5.4. Discussion

From the overview of the submissions (Sec. 4) and the experimental results, we can conclude as follows:

- The deep learning based methods dominate the field of person retrieval that the ResNet [8] and DenseNet [11] are the most two popular networks utilized as the backbone.
- The complicated frameworks, such as multi-model fusion with complex backbone networks, are tend to achieve superior performance.
- For the task of PR-A, the attention mechanism and learning the context information among different attributes are two effective methods to improve the performance, especially under the gallery set collected by person detectors.
- For the task of PR-ID, MGN [37] based frameworks can concentrate on both the global and local discriminative clues which contribute to generate superior performance especially under the system-level evaluation.
- As the introducing of distractions to the gallery set, it leads to heavily negative effect on the ranking results on both tasks. So the person retrieval is still a challenging issue under real scenarios.

6. Conclusion

This paper proposed a large-scale benchmark for person retrieval in real surveillance scenarios, in which the attribute recognition and person ReID are included. Different from the existing works, besides the common algorithm-level evaluation on closed set, the system-level evaluations are introduced to measure the algorithms combing with the coherent person detector, where the detected pedestrians constitute the large-scale gallery set and only some of them are annotated. Thus, it could reflect the performance under the real scenarios. We also give a comprehensive study for the impacts of various gallery sizes on the performance of the person retrieval system using the proposed metrics. The results indicate that person retrieval in real scenarios is still a intractable issue as lots of distractions exist in the gallery set. In future work, facing the large-scale gallery set, we will introduce a new evaluation mechanism based on "human-in-the-loop" to adopt users feedbacks to evaluate the system performance.

Acknowledgements. This work is sponsored by National Key Research and Development Program of China (2016YFB1001002) and CAS-AIR.

References

[1] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai. Person search via a mask-guided two-stream cnn model. In *ECCV*, pages 734–750, 2018. 1, 2

[2] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017. 6

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017. 2

[4] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM Multimedia*, pages 789–792, 2014. 1, 2

[5] T. Dozat. Incorporating nesterov momentum into adam. In *ICLR Workshop*, pages 2013–2016, 2016. 5

[6] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *PNAS*, 112(12):3618–3623, 2015. 3, 4

[7] M. Halstead, S. Denman, C. Fookes, Y. Tian, and M. S. Nixon. Semantic person retrieval in surveillance using soft biometrics: Avss 2018 challenge ii. In *AVSS*, pages 1–6, 2018. 2

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 5, 6, 7

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6

[10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 6

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 4, 5, 6, 7

[12] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang. Adversarially occluded samples for person re-identification. In *CVPR*, pages 5098–5107, 2018. 6

[13] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071, 2018. 2

[14] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, pages 111–115, 2015. 2

[15] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, pages 384–393, 2017. 2

[16] D. Li, X. Chen, Z. Zhang, and K. Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *ICME*, pages 1–6, 2018. 2

[17] D. Li, Z. Zhang, X. Chen, and K. Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE T-IP*, 28(4):1575–1590, 2019. 1, 2, 3

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2

[19] X. Li, A. Wu, and W.-S. Zheng. Adversarial open-world person re-identification. In *ECCV*, pages 280–296, 2018. 2

[20] S. Liao, Z. Mo, J. Zhu, Y. Hu, and S. Z. Li. Open-set person re-identification. *arXiv preprint arXiv:1408.0872*, 2014. 2

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 5, 6

[22] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. In *ICCV*, pages 493–501, 2017. 2

[23] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, pages 350–359, 2017. 1, 2

[24] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[25] C. C. Loy, D. Lin, W. Ouyang, Y. Xiong, S. Yang, Q. Huang, D. Zhou, W. Xia, Q. Li, P. Luo, et al. Wider face and pedestrian challenge 2018: Methods and results. *arXiv preprint arXiv:1902.06854*, 2019. 2

[26] H. Qi, T. Wu, M.-W. Lee, and S.-C. Zhu. A restricted visual turing test for deep scene and event understanding. *arXiv preprint arXiv:1512.01715*, 2015. 3, 4

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 4

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop*, pages 17–35, 2016. 1, 2

[29] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017. 5

[30] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, pages 1179–1188, 2018. 2

[31] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, pages 3960–3969, 2017. 2

[32] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *ICCV Workshops*, pages 87–95, 2015. 2

[33] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018. 2, 6

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 6

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015. 5

[36] P. L. Venetianer and H. Deng. Performance evaluation of an intelligent video surveillance system–a case study. *CVIU*, 114(11):1292–1302, 2010. 1

[37] G. Wang, Y. Yuan, X. Cheng, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Multimedia*, pages 274–282, 2018. 2, 4, 6, 7

[38] J. Wang, X. Zhu, S. Gong, and W. Li. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, pages 531–540, 2017. 2

[39] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3415–3424, 2017. 1, 2

[40] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 6

[41] Y. Xu, B. Ma, R. Huang, and L. Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *ACM Multimedia*, pages 937–940, 2014. 1, 2

[42] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106*, 2017. 6

[43] M. Zeng, C. Tian, and Z. Wu. Person re-identification with hierarchical deep learning feature and efficient xqda metric. In *ACM Multimedia*, pages 1838–1846. ACM, 2018. 6

[44] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 6

[45] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 2

[46] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3219–3228, 2017. 2

[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1, 2

[48] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[49] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017. 1, 2

[50] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE T-PAMI*, 38(3):591–606, 2016. 2

[51] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*, pages 5513–5522, 2017. 5

[52] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *ICCVW*, pages 331–338, 2013. 2

[53] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng. Fast open-world person re-identification. *IEEE T-IP*, 27(5):2286–2300, 2018. 2