LARGE-SCALE PERSON RE-IDENTIFICATION AS RETRIEVAL

Hantao Yao[1,2], *Shiliang Zhang*[3], *Dongming Zhang*[1], *Yongdong Zhang*[1,2], *Jintao Li*[1], *Yu Wang*[4], *Qi Tian*[5]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China
[4] National Computer Network Emergency Response Technical Team/Coordination Center of China
[5] Department of Computer Science University of Texas at San Antonio, San Antonio, USA
{yaohantao,dmzhang,zhyd,jtli}@ict.ac.cn, slzhang.jdl@pku.edu.cn, slimzczy@163.com,qi.tian@cs.utsa.edu

ABSTRACT

This paper targets to bring together the research efforts on two fields that are growing actively in the past few years: multi-camera person Re-Identification (ReID) and large-scale image retrieval. We demonstrate that the essentials of image retrieval and person ReID are the same, *i.e.*, measuring the similarity between images. However, person ReID requires more discriminative and robust features to identify the subtle differences of different persons and overcome the large variance among images of the same person. Specifically, we propose a coarse-to-fine (C2F) framework and a Convolutional Neural Network structure named as Conv-Net to tackle the large-scale person ReID as an image retrieval task. Given a query person image, the C2F firstly employ Conv-Net to extract a compact descriptor and perform the coarse-level search. A robust descriptor conveying more spatial cues is hence extracted to perform the fine-level search. Extensive experimental results show that the proposed method outperforms existing methods on two public datasets. Further, the evaluation on a large-scale Person-520K dataset demonstrates that our work is significantly more efficient than existing works, *e.g.*, only needs 180ms to identify a query person from 520K images.

Index Terms— Person Re-identification (ReID), Large-Scale Person Retrieval, Person-520K

1. INTRODUCTION

Person Re-Identification (ReID) targets to identify the re-appearing persons taken by multiple cameras. It is potential to address the challenging issues like intelligent surveillance video storage and analysis, as well as to explore the promising applications on public security. One key step of person ReID is to match the query person image to database images and return all and only the images of the same person. Therefore, person ReID is essentially similar to the image retrieval task. However, as shown in Figure 1, person images taken by different cameras can be easily affected by variances of camera viewpoint, human pose, illumination, occlusion, *etc*., making person ReID a significantly more challenging problem.

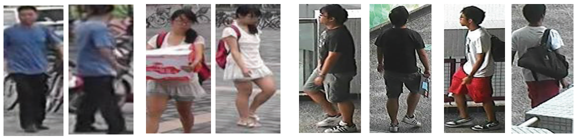

Fig. 1. Example images of four persons from Market1501 (first two) and CUHK03 (last two). The subtle differences among different persons and the large variance among images of the same person make person ReID challenging.

Although person ReID is challenging, its performance has been significantly improved by generating robust descriptions [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], learning discriminative distance metrics [3], and training powerful classifiers [5]. For instance, the Rank-1 accuracy on CUHK03 [11] has been pushed from 19.9%[11] to 73.1% [3]. However, most of existing algorithms are designed and evaluated on small-scale datasets, *e.g.,* CHUK03 consists of 14,096 images from 1,467 identities, the currently largest Market1501 [12] consists of 36,036 images from 1,501 identities. Working on those datasets, most of existing methods concatenate on improving the ReID accuracy, and pay less attention to the algorithm efficiency and generalization ability. A more detailed review of current ReID works will be presented in Sec. 2.

Because the surveillance video database commonly contains thousands of hours of videos and millions of person images, it is appealing to design a scalable person ReID system with both high efficiency and generalization ability. Considering the essentials of image retrieval and person ReID are the

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, and also supported in part by the NSFC(61525206,61572050,61672495, 61429201, 91538111, 61620106009, 61273247) and the National Key Research and Development Plan of China (2016YFB0801203).

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

same, we treat person ReID as a retrieval task, which is easier to extend to large-scale data. To make the retrieval system discriminative enough to the subtle differences among different persons, we propose a coarse-to-fine retrieval framework to progressively narrow down the search space and identify the query person, respectively. As shown in Figure 2, the framework consists of two components: coarse-level search and fine-level search. The coarse-level search returns the Top-M similar images using a compact descriptor extracted by our proposed CNN structure, *i.e.*, Conv-Net. The returned Top-M images are then reranked with a more robust descriptor.

The key component in our framework is the Conv-Net, which extracts discriminative deep feature. Inspired by the promising performance of CNN [13] in image recognition, CNN has been used in many person ReID works [1, 2, 3, 4]. Our work differs from them in the following aspects: 1) the proposed Conv-Net is easier to train and interpret, as well as could better avoid overfitting. Experiments show its advantages over existing CNN models. 2) Existing works commonly perform person ReID by linearly matching deep features between query and database images. We perform ReID in a coarse-to-fine retrieval framework, which considers offline indexing to ensure both the efficiency and accuracy.

To evaluate the performance of our work on large-scale person ReID task, we construct a Person-520K dataset containing about 520K person images. Experiments on Person-520K and other public benchmark datasets validate the promising accuracy and efficiency of our work.

The major contributions of this work are summarized into following: 1) We reveal person ReID can be solved in a corase-to-fine retrieval framework. This benefits person ReID systems with significantly better scalability and efficiency. 2) We present the Conv-Net, which is easy to train and implement, and extracts promising deep features. 3) The large-scale Person-520K dataset is construct to benefit future research on large-scale person ReID.

2. RELATED WORK

This work is closely related with deep learning based person ReID. In this section, we summarize existing works into three categories according to the network structure, *i.e.*, classification network, siamese network, and triplet network.

CNN [13, 14, 15] has shown promising performance on Large Scale Visual Recognition Challenge (ILSVRC). Several works [1, 2] directly fine-tune the classification network to extract image description for person ReID. As most of person ReID datasets do not contain enough samples to train the deep model, some efficient training strategies are proposed. For instance, Xiao *et al.* [1] propose a novel dropout strategy to efficiently trained the network on small-scale datasets.

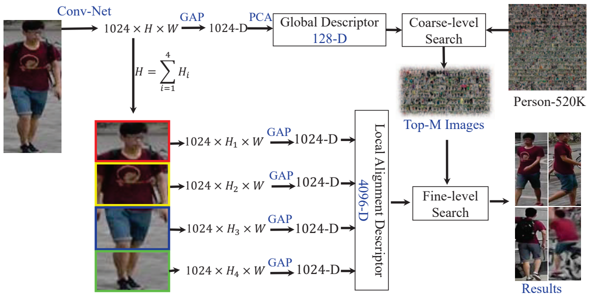

Fig. 2. The coarse-to-fine retrieval framework for large-scale person Re-identification.

Directly applying classification network in person ReID shows several shortcomings. For example, 1) training a classification network needs a lot of training samples, and 2) the number of neurons in the classification layer is related with the number of identities in the training set. Aiming to overcome these disadvantages, several works [3, 4, 5, 6, 7] employ siamese network to infer the image description. The siamese network takes a pair of images as input, and is trained to predict the similarity between two images. Yi *et al.*[7] propose a siamese convolutional network for distance metric learning. Zheng *et al.* [3] propose another network by jointly considering the objective functions of classification and similarity learning. Varior *et al* [6] combine the LSTM and siamese network architecture for person ReID.

The siamese network is trained with known pair-wise simialrity, which is too strict and hard to collect in some cases. Therefore, some researchers study to train the network with relative similarity among three images, named as triplet. Some works [8, 9, 10] employ the triplet networks to learn the discriminative description for person ReID. Cheng *et al.* [9] propose a multi-channel parts-based CNN model for person ReID. Liu *et al.* [10] propose an end-to-end Comparative Attention Network to generate image description. Su *et al.* [8] propose a semi-supervised network trained by triplet loss to learn human semantic attributes. The learned human attributes are treated as a discriminative mid-level feature for person ReID.

Our Conv-Net can be categorized into the classification network. Existing classification networks used for person ReID commonly employ fully-connected layers to implement classifiers. As the person ReID dataset contains less images, the huge parameters contained in fully-connected layer make the network easily overfitted during training. Our Conv-Net replaces the fully-connected classifier with a convolutional classifier, which contains significantly less parameters, and maintains the spatial cues on each feature map. As illustrated in our experiments, Conv-Net is easier to train and interpret, as well as could better avoid overfitting.

3. PROPOSED APPROACH

3.1. Framework

Image retrieval and person ReID could be uniformly formulated: given the query image I_q and N database images

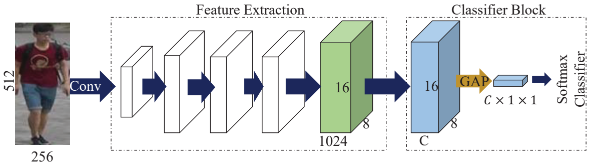

Fig. 3. The structure of Conv-Net, where GAP denotes Global Average Pooling.

$\mathcal{I} = \{(I_1, y_1), (I_2, y_2), ..., (I_N, y_N)\}$, where I_n and y_n are the image and label, respectively. The goal is to rerank those N images based on the similarities with query I_q. The sorted ranklist of N image can be denoted as $\{r_1, r_2, ..., r_N\}$, where r_i is the sorted index of image I_i. The objective function of those two tasks could be summarized as Eq. (1), *i.e.*,

$$\min \sum_{i=1}^{N} r_i f(y^q, y_i), f(y^q, y_i) = \begin{cases} 1 & y^q = y_i \\ 0 & y^q \neq y_i \end{cases} \quad (1)$$

where y^q is the label for query image I_q.

To chase a reasonable trade-off between retrieval accuracy and efficiency, we propose a coarse-to-fine retrieval framework shown in Figure 2. Before proceeding to the online retrieval, we first train a Conv-Net to extract a global descriptor $\mathbf{f}^g$ and local descriptor $\mathbf{f}^l$ for each person image. We expect the global descriptor $\mathbf{f}^g$ to quickly narrow down the search space and the local descriptor to convey enough discriminative power.

The online retrieval consists of two stages: coarse-level search and fine-level search. Given a query image I_q, we firstly extract its global descriptor $\mathbf{f}_q^g$ and local descriptor $\mathbf{f}_q^l$ using Conv-Net. Then, we calculate the Euclidean distance between the query global descriptor $\mathbf{f}_q^g$ and all gallery descriptors to obtain the Top-M, *e.g.,* M=500, similar images. Next, we utilize descriptor $\mathbf{f}_q^l$ to perform a fine-level search on the M images. In the following, we first describe the Conv-Net, then proceed to introduce the feature extraction and online retrieval.

3.2. Conv-Net

Existing networks such as AlexNet[13], VGGNet[14], and GoogleNet[16] could be regarded as a combination of feature extraction and classifier learning. Given an image, these networks firstly use several convolutional layers or fully-connected layers to generate the description, and then infer a fully-connected classifier, *e.g.,* the output of layer fc_7 in AlexNet can be viewed as the inferred description, and layer fc_8 could be regarded as the classifier. The learned features by CNN have shown promising performance in various vision tasks. Although using fully-connected layers to infer the classifier is effective for classification tasks, the huge parameters introduced by such layers could make the network hard to tune and prone to overfitting.

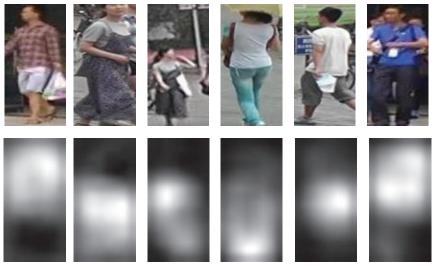

Fig. 4. Example confidence maps generated by Conv-Net. The first row shows the original images, and the second row shows their feature maps with the maximum average response values, *i.e.*, the confidence maps.

Inspired by [17], we propose a Conv-Net to generate the description for each image. Conv-Net is trained in a classification task by treating each identity in the training set as a category. As shown in Figure 3, the Conv-Net is produced by replacing the fully-connected classifier with a classifier block. The classifier block firstly employs convolutional operation to generate feature maps explicitly corresponding to each class, then uses Global Average Pooling (GAP) *i.e.*, Eq. (2), to generate the score for each class,

$$s_c = \frac{1}{W \times H} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathcal{C}_{c,h,w}, \quad (2)$$

where s_c is the average response of the c-th feature map $\mathcal{C}_c$, and $\mathcal{C}_{c,h,w}$ denote the activation value on the location (h, w) of c-th feature map .

As shown in Figure 3, Conv-Net firstly generates C feature maps corresponding to C classes, *e.g.,* the feature maps with size of $C \times 16 \times 8$ in Figure 3, then obtain a C-dim ($C \times 1 \times 1$) classification score vector with GAP. During training phase, the objective function guides the network to minimize the classification error, which increases the average response values of feature maps for the correct class, and suppresses the response values of the other feature maps. As a consequence, the average response of c-th ($c \in [1, C]$) feature map reasonably denotes the probability that the input image belongs to c-th class. The resulting feature maps could be viewed as category confidence maps. Some confidence map examples are shown in Figure 4.

Compared with existing network using fully-connected classifier, Conv-Net has following advantages: 1) Conv-Net generates an explicit confidence map for each class, *i.e.*, the confidence map denotes the spatial activation of the input image for each class. This makes the network easier to interpret. 2) The GAP operation does not involve any extra parameters, thus could better avoid overfitting. 3) The confidence map reveals the discriminative regions of the input image, thus can be applied for foreground extraction, which can be useful for cases where the input person image is not precisely aligned. As shown in Figure 4, the confidence maps of Conv-Net focus

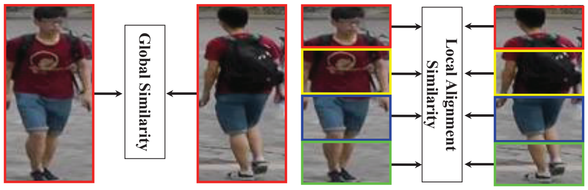

Fig. 5. Illustration of global and local descriptions.

on the most discriminative regions. In the following, we introduce how to use the trained Conv-Net to generate the global descriptor $\mathbf{f}^g$ and local descriptor $\mathbf{f}^l$.

3.3. Feature Extraction and Online Retrieval

After training Conv-Net, we discard its classifier block, and then use the outputs of the convolutional layer as deep features. This makes the extracted features less dependent on the training set and show better generalization ability. Given a person image I, we firstly resize it to the size of 512×256, and then fed it into Conv-Net to obtain the activations of last convolutional layer, *i.e.*, the green layer in Figure 3. We denote this output as $\mathcal{X} \in R^{K \times H \times W}$, *e.g.,* K=1024, H=16, W=8 when using GoogleNet as base-network. Based on $\mathcal{X}$, we calculate the global descriptor $\mathbf{f}^g$ with Eq. (3),

$$\mathbf{f}^g = [f_1, f_2, ..., f_K]^\top, f_k = \frac{1}{W \times H} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathcal{X}_{k,h,w}, \quad (3)$$

where $\mathcal{X}_{k,h,w}$ is the activation value on location (k, h, w). The final global descriptor is generated by applying PCA to reduce the dimensionality of $\mathbf{f}^g$ to 128-D.

The descriptor $\mathbf{f}^g$ describes the global appearance of the person image, but lacks the ability to describe the local parts, which are important to differentiate the subtle difference among persons. Because most person images show relatively stable spatial part configuration, *e.g.*, the head in the upper area of the image, the body in the middle area of the image, and legs in the lower area of the image, we divide the person images into four stripes and extract deep feature from each of them. As shown in Figure 5, we could generate an alignment local descriptor based on those features. Specifically, the alignment local descriptor $\mathbf{f}^l$ is calculated as,

$$\mathbf{f}^l = [\mathbf{f}^{l_1}; \mathbf{f}^{l_2}; \mathbf{f}^{l_3}; \mathbf{f}^{l_4}], \quad (4)$$

where

$$\mathbf{f}^{l_i} = [f_1, f_2, ..., f_K]^\top, \quad (5)$$

$$f_k = \frac{1}{\mathcal{H} \times W} \sum_{h=(i-1)*\mathcal{H}+1}^{i*\mathcal{H}} \sum_{w=1}^{W} \mathcal{X}_{k,h,w}, i \in [1, 4], \quad (6)$$

where $\mathcal{H} = \frac{H}{4}$. Note that, local descriptors are extracted from the activation $\mathcal{X}$ previously used in $\mathbf{f}^g$ extraction. Therefore, to extract $\mathbf{f}^g$ and $\mathbf{f}^l$ for an image, we only need to run Conv-Net for one time.

Table 1. Details of Person-520K dataset.

	Training		Testing		Query	
	Identity	Image	Identity	Image	Identity	Image
CUHK03	736	7,029	731	5,606	731	1,461
Market1501	751	12,936	750	19,732	750	3,368
Distractors	-	-	-	500K	-	-
Total	**1,487**	**19,965**	**1,481**	**520K**	**1,481**	**4,829**

Consequently, we extract two features $\mathbf{f}^g$ and $\mathbf{f}^l$ from each database image for off-line indexing. During online retrieval, $\mathbf{f}^g$ is first extracted from the query image to perform the coarse-level search, then followed by the fine-level search using $\mathbf{f}^l$. Note that, more efficient indexing strategies like Product Quantization [18] can be involved in our framework to further boost the search efficiency.

4. EXPERIMENTS

4.1. Person-520K Dataset

To test the accuracy and scalability of the proposed method, we introduce the large-scale Person-520K dataset. Because it is expensive to collect and annotate a large number of person IDs, we generate Person-520K from CUHK03 [11], and Market1501 [12] by adding a large number of retrieval distractors.

CUHK03 consists of 1,467 identities. The standard experimental setting [11] selects 1,367 identities for training, and the rest 100 for testing. Aiming to enlarge the testing set and verify the generalization ability of our algorithm, we randomly select 736 identities for training, and the rest 731 for testing. The images in CUHK03 are taken by two cameras. For each testing identity, we select one image from each camera as query and treat the others as database images. This results in 1,461 queries and 5,606 database images from CUHK03.

Market1501 contains 32,668 images of 1,501 identities, and each image is annotated with a bounding box. Each identity is captured by at most six cameras. We use the standard training, testing, and query split provided by the authors. By mixing the images from CUHK03 and Market1501 together and further adding 500K distracters, the final Person-520K contains about 520K database images, 20K training images, and 4,829 queries, respectively. Details of this dataset are summarized in Table 1. Note that, the training data and testing data do not share any common identity.

4.2. Implementation Details

We use Caffe to implement and train Conv-Net [1]. The Conv-Net is modified based on the networks described in [19], and is initialized with the model introduced in [20]. We use a step strategy to train Conv-Net on four Tesla K80 GPU. Parameters like maximum number of iterations, learning rate,

[1] The code and URL of person-520K dataset can be found from: https://github.com/coldrainyht/ICME2017

Table 2. Comparison between Conv-Net and other CNN models on Market1501.

Models	mAP(%)	Rank-1 (%)
AlexNet [13]	26.79	50.89
VGG16Net [14]	38.27	65.02
GoogleNet [16]	48.24	70.27
Res50Net [15]	51.48	73.69
Conv-Net	**54.86**	**76.36**

Table 3. Effect of various input sizes on Market1501.

Scales	mAP(%)	Rank-1 (%)
224×224	54.86	76.36
128×64	44.57	68.27
256×128	53.50	75.89
384×192	56.04	77.64
512×256	61.9	81.5

stepsize, and gamma are set as 50,000, 0.01, 2500, and 0.75, respectively. All experiments are conducted on a server equipped with a Intel Xeon E5-2650 CPU.

4.3. Performance of Conv-Net

We first compare the person ReID performance of Conv-Net and other widely used CNN models in person ReID works [3, 11, 8], *i.e.*, GoogleNet, VGG16Net, and Res50Net. The related results on Market1501 are summarized in Table 2. As shown in Table 2, Conv-Net achieves both higher mAP and Rank-1 accuracy than the other four networks. For example, Conv-Net achieves the mAP of 54.86%, showing 6.62% improvement over the 48.24% of GoogleNet, *i.e.*, the base network of Conv-Net. It thus can be seen that, as Conv-Net involves less parameters by removing the fully connect layers, it achieves better performance.

In the previous experiment, we resize the input person image into 224×224, which is commonly used in image classification networks. However, the reasonable and natural height-width ratio of person images should be larger than 1.0. Using 224×224 as the input size might be not a optimal choice for Conv-Net. We therefore further evaluate the effects of various input sizes and height-width ratios. As shown in Table. 3, larger size plus ratio close to 2.0 gets the best performance. Therefore, we use 512×256 as the input size in following experiments.

4.4. Performance on Public Datasets

In this section, we evaluate the proposed retrieval framework on two public person ReID datasets: Market1501[12], and CUHK03[11].

Table 4 summarizes our comparison with existing methods on Market1501 in the aspects of both mAP and Rank-1 accuracy. As shown in Table 4, our coarse-level search achieves the 61.9% mAP and 81.5% Rank-1 accuracy, which both outperform the existing methods. The performance of coarse-level search clearly reveals the advantage of our Conv-Net in deep feature extraction. Because Conv-Net is easier to train, it can be leveraged in existing methods [1, 2, 3] to further boost their ReID performance. From Table 4, we could also observe that using the complete coarse-to-fine (C2F) framework further improves the ReID accuracy.

Table 4. Comparison on Market1501 (Single Query).

Methods	mAP(%)	Rank-1 (%)
DNS [21]	29.87	55.43
Gate Reid [22]	39.61	65.88
VGG16net [14]	38.27	65.02
DLCNN(VGG16net) [3]	47.45	70.16
Res50Net [15]	51.48	73.69
DLCNN(Res50Net) [3]	59.87	79.51
GoogleNet [16]	48.24	70.27
coarse-level	61.9	81.5
C2F	**64.6**	**84.64**

Table 5. Comparison on CUHK03 with labeled bounding boxes.

Methods	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
DeepReID [11]	19.9	49.3	64.7
DNS [21]	54.7	80.1	88.3
DLCNN(Res50net) [3]	66.1	90.1	**95.5**
MetricEmsemb [23]	62.1	89.1	94.3
coarse-level	72.85	89.53	94.82
C2F	**75.13**	**90.16**	94.92

We next show the comparison on CUHK03. CUHK03 provides two kinds of cropped images: labeled bounding boxes and bounding boxes detected by DPM. We test and report results on the labeled bounding boxes. This experiment uses the standard setting of CUHK03, *i.e.*, 1,467 identities are split into two parts: 1,367 training identities, and 100 testing identities and the experiments are repeated with 20 random splits. Table 5 summarizes the comparison with existing methods on accuracies of Rank-1, Rank-5, and Rank-10, respectively. As shown in Table 5, the proposed method significantly outperforms existing methods. Especially for the Rank-1 accuracy, our method achieves the accuracy of 75.13%, which is 9% higher than the best accuracy 66.1% of the four compared works.

4.5. Performance on Person-520K

We finally evaluate the accuracy and efficiency of the proposed method on the large-scale Person-520K. The related results and comparisons are shown in Table 6. For AlexNet and GoogleNet, we use the outputs of layer fc_7 and $pool5/7 \times 7\_s1$, respectively as the baseline image descriptors. We apply L2-Normalization for each descriptor. In Table 6, the feature **f** denotes the concatenation of our global and local descriptors.

As shown in Table 6, the original 1024-D Conv-Net feature $\mathbf{f}^g$ achieves higher mAP and Rank-1 accuracy than features from AlexNet and GoogleNet. It also can be observed that, reducing the dimensionality of $\mathbf{f}^g$ with PCA does not

Table 6. The performance on Person-520K. "Time" denotes the retrieval time. Feature extraction takes about 50ms for Conv-Net on GPU.

Methods	Dim	Time(ms)	mAP(%)	Rank-1(%)
AlexNet	4,096	3932	17.13	33.46
GoogleNet	1,024	960	36.38	56.05
Conv-Net_$\mathbf{f}^g$	1,024	961	43.42	60.84
Conv-Net_$\mathbf{f}^g$	512	500	41.06	61.79
Conv-Net_$\mathbf{f}^g$	256	262	40.84	61.21
Conv-Net_$\mathbf{f}^g$	128	148	39.99	59.3
Conv-Net_$\mathbf{f}$	5,120	4746	46.95	64.60
coarse-level	128	150	39.99	59.3
C2F	-	180	46.74	64.58

significantly degrade its performance, *e.g.*, reducing to 128-D drops the mAP from 43.42% to 39.99%. Moreover, lower dimensionality significantly boosts the retrieval efficiency, *e.g.,* $150ms$ for 128-D *vs* $961ms$ for 1024-D. This shows the rationality of using 128-D $\mathbf{f}^g$ for coarse-level search.

By concatenating both $\mathbf{f}^g$ and $\mathbf{f}^l$, the $\mathbf{f}$ obtains the best accuracy at the cost of longest running-time. Comparing with the linear search strategy with descriptor $\mathbf{f}$, our coarse-to-fine framework achieves comparable mAP and Rank-1 accuracy. However, the efficiency of the coarse-to-fine retrieval framework is *28 times* faster speed. Therefore, we could conclude that our proposed framework is effective for person ReID and provides a reasonable trade-off between accuracy and efficiency.

5. CONCLUSION

We demonstrate that person ReID and image retrieval are essentially the same, therefore person ReID can be effectively tackled by leveraging the research efforts on large-scale image retrieval. To validate this idea, we propose a coarse-to-fine framework for large-scale person ReID. One key factor of the framework is the Conv-Net, which is used to generate a compact descriptor for the coarse-level search and a robust descriptor for the fine-level search, respectively. The proposed method not only achieves higher accuracy than existing works on two public datasets, but is also significantly more efficient for large-scale person ReID.

6. REFERENCES

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

[2] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," in *CVPR*, 2017.

[3] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person re-identification," *arXiv preprint arXiv:1611.05666*, 2016.

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

[5] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015, pp. 3908–3916.

[6] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *ECCV*, 2016, pp. 135–153.

[7] D. Yi, Z. Lei, and S.Z. Li, "Deep metric learning for practical person re-identification," in *ICPR*, 2014.

[8] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *ECCV*, 2016.

[9] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.

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

[11] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159.

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[13] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2014.

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

[16] 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*, 2015.

[17] M. Lin, C. Qiang, and S. Yan, "Network in network," in *ICLR*, 2014.

[18] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *TPAMI*, 2011.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *NIPS*, 2015.

[20] "https://github.com/lim0606/caffe-googlenet-bn," .

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

[22] R.R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016, pp. 791–808.

[23] S. Paisitkriangkrai, C. Shen, and A.van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015.