

Part-based Deep Hashing for Large-scale Person Re-identification*

Fuqing Zhu, Xiangwei Kong, *Member, IEEE*, Liang Zheng, *Member, IEEE*, Haiyan Fu, *Member, IEEE*, Qi Tian, *Fellow, IEEE*,

Abstract—Large-scale is a trend in person re-identification (re-id). It is important that real-time search be performed in a large gallery. While previous methods mostly focus on discriminative learning, this paper makes the attempt in integrating deep learning and hashing into one framework to evaluate the efficiency and accuracy for large-scale person re-id. We integrate spatial information for discriminative visual representation by partitioning the pedestrian image into horizontal parts. Specifically, Part-based Deep Hashing (PDH) is proposed, in which batches of triplet samples are employed as the input of the deep hashing architecture. Each triplet sample contains two pedestrian images (or parts) with the same identity and one pedestrian image (or part) of the different identity. A triplet loss function is employed with a constraint that the Hamming distance of pedestrian images (or parts) with the same identity is smaller than ones with the different identity. In the experiment, we show that the proposed Part-based Deep Hashing method yields very competitive re-id accuracy on the large-scale Market-1501 and Market-1501+500K datasets.

Index Terms—Deep Learning, hashing, part-based, large-scale person re-identification.

I. INTRODUCTION

THIS paper focuses on large-scale person re-identification (re-id), which has received increasing attention in automated surveillance for its potential applications in human retrieval, cross-camera tracking and anomaly detection. Given a pedestrian image, person re-id aims to match in a cross-camera database for the bounding boxes that contain the same person. Matching cross scenarios is challenging due to the varieties of lighting, pose and view point.

Person re-id lies in between image classification [1], [2], [3] and retrieval [4], [5], which has made detailed discussion in [6]. Previous person re-id works [7], [8], [9], [10] usually

take advantage of both image classification and retrieval. This work considers two issues in large-scale person re-id: efficiency and CNN models for effective descriptors. On the one hand, computational efficiency has been a concern in person re-id works. Some state-of-the-art methods employ brute-force feature matching strategies [11], [12], which obtain good matching rate. However, these methods suffer from low computational efficiency in large-scale applications. Motivated by [13], [10], we view person re-id as a special task of image retrieval. Both tasks share the same target: finding the images containing the same object/pedestrian as the query [10]. A reasonable choice to address the above efficiency problem of large-scale person re-id therefore involves the usage of image retrieval techniques. Hashing, known for fast Approximate Nearest Neighbor (ANN) search, is a good candidate in our solution kit. The main idea of hashing method is to construct a series of hash functions to map the visual feature of image into a binary feature vector so that visually similar images are mapped into similar binary codes. Recently, hashing methods based on deep neural networks [14], [15], [16], [17], [18], [19] obtain higher accuracy than traditional hashing methods. However, to our knowledge, there are few works employing hashing to address large-scale person re-id.

On the other hand, the Convolutional Neural Network (CNN) has demonstrated its effectiveness in improving accuracy of person re-id [20], [21], [7], [9]. The Siamese CNN model uses training image pair as input and a binary classification loss is used to determine if they belong to the same ID. This cross-image representation is effective in capturing the relationship between the two images and addressing horizontal displacement problem. For the conventional classification based CNN model, Zheng *et al.* [9] propose to learn an ID-discriminative embedding to discriminate between pedestrians in the testing set. These methods, while achieving impressive person re-id accuracy, do not address the efficiency issue either, because they typically use the Euclidean or Cosine distance for similarity calculation which is time-consuming under large galleries and high feature dimensions. Currently the largest person re-id dataset Market-1501 [10] contains 32,668 annotated bounding boxes, plus a distractor set of 500K images. It poses the scaling problem for person re-id methods. This paper therefore investigates how to balance re-id effectiveness and efficiency.

The approach we pursue in this work, as mentioned above, is motivated by hashing and CNN, which takes into account the efficiency and accuracy, respectively. A triplet loss based supervised Deep Hashing framework is employed to address

This work was supported in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (NSFC) under Grant 71421001, in part by the National Natural Science Foundation of China (NSFC) under Grant 61502073 and Grant 61429201, in part by the Open Projects Program of National Laboratory of Pattern Recognition under Grant 201407349, and in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290 and the Faculty Research Gift Awards by NEC Laboratories of America and Blippar. (Corresponding author: Xiangwei Kong)

F. Zhu, X. Kong and H. Fu are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: fuqingzhu@mail.dlut.edu.cn; kongxw@dlut.edu.cn; fuhuy@dlut.edu.cn).

L. Zheng is with the University of Technology Sydney, NSW, Australia. (e-mail: liangzheng06@gmail.com).

Q. Tian is with the University of Texas at San Antonio, Texas 78249-1604, USA (e-mail: qitian@cs.utsa.edu).

*Citation for this paper: F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based Deep Hashing for Large-scale Person Re-identification," *IEEE Transactions on Image Processing*, DOI: 10.1109/TIP.2017.2695101, 2017.

the efficiency of large-scale person re-id. The triplet deep neural networks [22], [23], [24], which have been used in face recognition [23] and fine-grained image similarity models [24], learn discriminative embeddings by imposing a relative distance constraint. The relative distance constraint aims to minimize the distance between positive pairs, while pushing away the negative pairs. This constraint is flexible comparing with restricting the distances of positive or negative pairs in an absolute range. Moreover, the spatial information of pedestrian image is beneficial for higher person re-id accuracy, because the local parts of pedestrians provide more precise matching strategy compared with using the entire pedestrian images. The part-based trick is useful for improve the accuracy in face verification, such as DeepID [25] and DeepID2 [26]. In DeepID [25], the face image is converted into ten parts which are global regions taken from the weakly aligned faces and local regions centered around the five facial landmarks, respectively. However, the part partitioning strategies of DeepID is not suitable for ensuring the efficiency of large-scale person re-id. For simplicity, in this paper we just partition the entire pedestrian image into horizontal 3 or 4 parts without any semantic alignment strategy. Our work gives two aspects of improvement on the basis of triplet-based deep neural network works [22], [24] for large-scale person re-id. First, in the intermediate layers of CNN, a hash layer is designed to make the output of network suitable for binarization. Second, the proposed network is composed by several sub-network branches for individual parts, and each sub-network branch is a triplet-based deep network. From the above consideration, we propose a **Part-based Deep Hashing (PDH)** method for large-scale person re-id. Our goal is to generate a binary representation for each pedestrian image using the deep CNN, which 1) is effective in discriminate different identities, 2) integrates spatial constraint, and 3) improves efficiency for the large-scale pedestrian gallery in terms of both memory and speed. Our code will be available at the website <https://sites.google.com/site/fqzhu001>.

Different from most previous works on person re-id, this paper focuses on hashing methods on Market-1501 dataset and its associating distractor set with 500K images. To our best knowledge, there is only one published paper which utilizes deep hashing for person re-id [18] on CUHK 03 [7], a dataset having only 100 identities in each gallery split. We show that our method yields effective yet efficient person re-id performance compared to several competing methods. The main contributions of this paper are listed below.

- Among the first attempts, we employ hashing to improve the efficiency for large-scale person re-id. While several previous works [18] only use small datasets, this paper reports large-scale evaluation results on the largest Market-1501 and Market-1501+500K datasets, such gaining more insights into the hashing task. The binary hash codes achieve fast matching for large-scale person re-id, which addresses the problem of computational and storage efficiency.
- A part-based model is integrated into the deep hashing framework to increase the discriminative ability of visual

matching. The performance increases significantly compared with the baseline.

The rest of the paper is organized as follows. In section II, we review related work briefly. The proposed PDH method will be described in section III. In section IV, extensive results are presented on Market-1501 and Market-1501+500K datasets. Finally, we conclude the paper in section V.

Note: this work was done in late 2015 when we were trying the triplet loss network to learn embeddings. Later, we turn to the identification models and obtain more competitive results. Interested readers can also refer to our works on the identification models [27], [9], [6], [28], [29], [30].

II. RELATED WORK

This paper considers the efficiency and accuracy of large-scale person re-id via deep hashing method. So we briefly review the methods of person re-id using both hand-crafted and deeply-learned features, and hashing methods.

A. Hand-crafted Methods for Person Re-identification

The previous mainstream works in person re-id typically focus on visual feature representation [11], [13], [31] and distance metric learning [32], [33], [34]. On feature representation, Ma *et al.* [35] utilize Gabor filters and Covariance descriptors to deal with illumination changes and background variations, while Bazzani *et al.* [36] design a Symmetry-Driven Accumulation of Local Features (SDALF) descriptor. Inspired by recent advanced Bag-of-Words (BOW) model in large-scale image retrieval field, Zheng *et al.* [10] propose an unsupervised BOW based descriptor. By generating a codebook on training data, each pedestrian image is represented as a histogram based on visual words. Li *et al.* [37] learn a cross-view dictionaries based on SIFT and color histogram to obtain an effective patch-level feature across different views for person re-id. Ma *et al.* [38] use Fisher Vector (FV) to encode local feature descriptors for patches to improve the performance of person re-id. Liao *et al.* [8] propose a method for building a descriptor which was invariant to illumination and viewpoint changes. Zhao *et al.* [11] propose a method which assigned different weights to rare colors on the basis of salience information among pedestrian images. However, traditional fixed hand-crafted visual features may not optimally represent the visual content of images. That means a pair of semantically similar pedestrian images may not have feature vectors with relatively small Euclidean distance. In the work of distance metric learning methods for person re-id, the classic RankSVM [34], [32] and boosting [33] methods are widely used. B. Prosser *et al.* [34] solve person re-id task as a ranking problem using RankSVM to learn similarity parameters. The method of KISSME [39] and EIML [40] are effective metric learning methods which have been shown in [41].

B. Deeply-learned Methods for Person Re-identification

Recently the state-of-the-art methods in person re-id have been dominated with deep learning models. The main advantage is that the CNN framework can either optimize the feature

representation alone [9] or simultaneously learn features and distance metrics [7]. Li *et al.* [7] propose a filter pairing neural network (FPNN) by a patch matching layer and a maxout-grouping layer. The patch matching layer is used to learn the displacement of horizontal stripes in across-view images, while the maxout-grouping layer is used to boost the robustness of patch matching. Ahmed *et al.* [20] design an improved deep neural network by adding a special layer to learn the cross-image representation via computing the neighborhood distance between two input images. The softmax classifier is added on the learned cross-image representation for person re-id. Yi *et al.* [42] employ the Siamese architecture which consists of two sub-networks. Each sub-network processes one image independently and the final representations of images are connected to evaluate similarity by a special layer. The deep networks are trained by preserving the similarity of the two images. The author evaluates the performance on VIPER [43] and PRID-2011 [44] datasets. However, the VIPER and PRID-2011 are both comparatively small datasets. E. Ustinova *et al.* [45] utilize bilinear pooling method based on Bilinear CNN for person re-id, which is implemented over multi-region for extracting more useful descriptors in the two large datasets CUHK 03 [7] and Market-1501 [10]. Chen *et al.* [21] design a deep ranking framework to formulate the person re-id task. The image pair is converted into a holistic image horizontally firstly, then feeds these images into CNN to learn the representations. Finally the ranking loss is used to ensure that positive matched image pair is more similar than negative matched image pair. Wang *et al.* [46] design a joint learning deep CNN framework, in which the matching of single-image representation and the classification of cross-image representation are jointly optimized for pursuing better matching accuracy with moderate computational cost. Since single-image representation is efficient in matching, while cross-image representation is effective in modeling the relationship between probe image and gallery image, the fusion of two representation losses together is utilized the advantages of both these representations. Xiao *et al.* [47] propose a pipeline for learning generic and robust deep feature representations from multiple domains with CNN, in which the Domain Guided Dropout algorithm is utilized to improve the feature learning procedure.

C. Review of Hashing Methods

The field of fast Approximate Nearest Neighbor (ANN) search has been greatly advanced due to the development of hashing technique, especially those based on deep CNN. For the non-deep hashing methods, the hash code generation process has two stages. First, the image is represented by a vector of hand-crafted visual features (such as Gist descriptor). Then, separate projection or quantization step is used to generate hash codes. Unsupervised and supervised hashing are two main streams, such as Spectral Hashing (SH) [48], Iterative Quantization (ITQ) [49], Semi-supervised Hashing (SSH) [50], Minimal Loss Hashing (MLH) [51], Robust Discrete Spectral Hashing (RDSH) [52], Zero-shot Hashing (ZSH) [53] and Kernel Supervised Hashing (KSH) [54]. However,

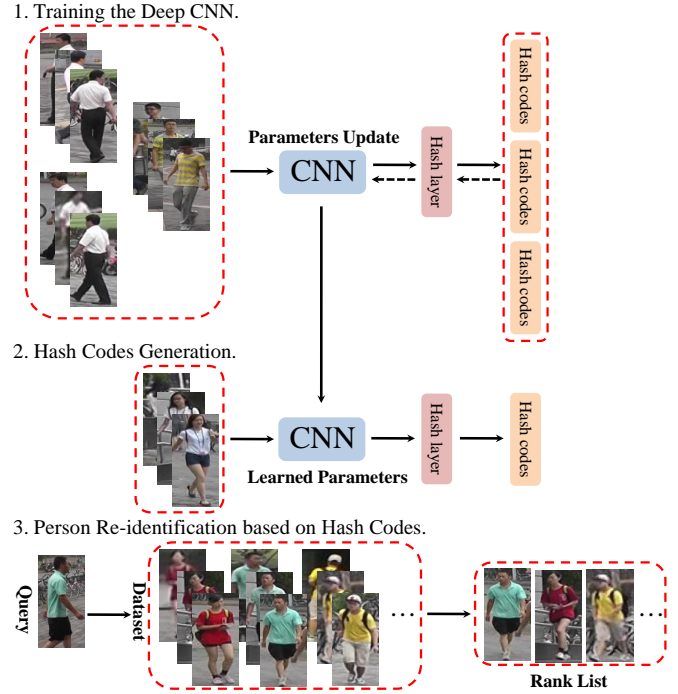


Fig. 1. The overview of baseline triplet-based deep hashing model for large-scale Person Re-id. The deep hashing model consists of three main modules. The first module is training the deep CNN model. In the second module, the trained CNN model is used to generate hash codes for pedestrian images of query and testing set. The final module is to retrieve similar images based on Hamming distance of hash codes between the samples of query and testing set.

hashing methods based on hand-crafted features may not be effective in dealing with the complex semantic structure of images, thus producing sub-optimal hash codes.

The deep hashing method maps the input raw images to hash codes directly, which learns feature representation and the mapping from the feature to hash codes jointly. Xia *et al.* [14] propose a supervised deep hashing method CNNH, in which the learning process is decomposed into a stage of learning approximate hash codes from similarity matrix, followed by a stage of simultaneously learning hashing functions and image representations based on the learned approximate hash codes. Zhao *et al.* [15] propose a Deep Semantic Ranking Hashing (DSRH) method to employ multi-level semantic ranking supervision information to learn hashing function, which preserves the semantic similarity between multi-label images. Lai *et al.* [16] develop a “one-stage” supervised hashing framework by a well designed deep architecture. The deep neural network employs the shared sub-network which makes feature learning and hash coding process simultaneously. Lin *et al.* [17] propose a point-wise supervised deep hashing method by adding a latent layer in the CNN for fast image retrieval. Zhang *et al.* [18] propose a novel supervised bit-scalable deep hashing method for image retrieval and person re-id. By designing an element-wise layer, the hash codes can be obtained bit-scalable, which is more flexible to special task when need different length of hash codes.

It is true that deep hashing has been employed in image retrieval, and that part-based method is a common technique

to improve re-id performance. However, both techniques have rarely been evaluated in person re-id and hashing tasks, respectively, especially in large-scale settings. Our work departs from previous person re-id works. We apply such simple yet effective techniques on the Market-1501 and Market-1501+500K datasets, and provide insights on how re-id performance (efficiency and accuracy) can be improved on the large-scale settings.

III. PROPOSED APPROACH

The task of person re-id is to match relevant pedestrian images for a query in the cross-camera scenario. Due to the variation of pedestrian in different scenarios, the spatial information is important for enhancing the discriminative ability of image representation. This is the motivation of integrating part-based model into the baseline triplet-based deep hashing framework, so that more discriminative hash codes can be generated. First, an overview of the baseline triplet-based deep Convolutional Neural Network (CNN) hashing framework for person re-id is illustrated in Fig. 1. The triplet-based deep CNN hashing framework to generate the binary hash codes for pedestrian images based on the CaffeNet [55], where a hash layer is well designed to ensure the compact binary output. In the training phase, a triplet-based loss function is employed for learning optimal parameters of the deep CNN model. Second, the proposed PDH method is implemented on the basis of the triplet-based deep hashing framework, which is illustrated in Fig. 3. For the part subsets of pedestrian images at the same corresponding location, we train a separate network for each part subset and obtain a series of optimal part-based deep CNN models. In this way, the corresponding parts of testing pedestrian image are processed by a series of trained part-based deep CNN models. The final representation of pedestrian image is the concatenation of each part result. The learned hash codes will be directly used for person re-ID without any feature selection [56], [57] process. Third, due to each identity has multiple query images in a single camera, multiple query images are merged into a single query for a further accuracy improvement of large-scale person re-id.

A. Baseline Triplet-based Deep Hashing Model

We employ the triplet-based deep hashing method to solve the efficiency problem of large-scale person re-id. The baseline method of triplet-based deep hashing is an end-to-end framework which jointly optimizes the image feature representation and hashing function learning, *i.e.* the input of framework is raw pixels of pedestrian images, while the output is hash codes. For the task of person re-id, the aim is to obtain the hash code of pedestrian image by the trained deep hashing model. How to train a discriminative deep neural network that can preserve the similarity of samples is critical. We briefly describe the training process of triplet-based deep CNN hashing model.

Each training sample is associated with an identity label. The principle of learning optimal deep neural network is formulated to ensure the Hamming distance of hash codes small for same identity samples. Meanwhile, the Hamming

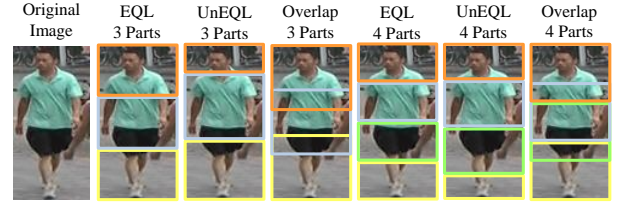


Fig. 2. Different part partitioning strategies. “EQL” and “UnEQL” represent dividing the image equally and unequally, respectively.

distance of binary hash codes should be large for different identity samples. The triplet-based input form is suitable for learning the parameters of deep neural network. Each triplet input includes three pedestrian images, in which one of them is anchor. The other two images are the same and different identity samples with the anchor, respectively.

Let I_i be anchor. I_i^+ and I_i^- are the same and different identity samples with the anchor, respectively. Let the hash code representation of image I_i represent as $H(I_i)$, which is the response of hash layer. The hash layer follows fully connected layer ($FC7$).

$$H(I_i) = \text{sign}(\mathbf{w}^T FC7(I_i)), \quad (1)$$

where \mathbf{w} denotes weights in the hash layer and $\text{sign}(v)$ returns 1 if $v > 0$ and 0 otherwise. According to this criterion, the objective function is

$$\min_{\mathbf{w}} \sum_{i=1}^N (\|H(I_i) - H(I_i^+)\|_H - \|H(I_i) - H(I_i^-)\|_H), \quad (2)$$

where \mathbf{W} denotes weights of each layer. The weights updating of each layer is achieved by a triplet-based loss function which is defined by

$$\begin{aligned} \mathcal{L}(H(I_i), H(I_i^+), H(I_i^-)) \\ = \max(0, 1 - (\|H(I_i) - H(I_i^-)\|_H \\ - \|H(I_i) - H(I_i^+)\|_H)), \end{aligned} \quad (3)$$

where $H(I_i), H(I_i^+), H(I_i^-) \in \{0, 1\}^q$ and $\|\cdot\|_H$ represents the Hamming distance. The loss function (3) is not differentiable due to the $\|\cdot\|_H$ of (3) and the sign function of (1). To facilitate the optimization, a relaxation trick on (3) is utilized to replace the Hamming distance with the L_2 norm. In addition, we replace the sign function of (1) with sigmoid function. Let $f_H(I_i)$ represent the relaxation of $H(I_i)$.

$$f_H(I_i) = \text{sigmoid}(\mathbf{w}^T FC7(I_i)), \quad (4)$$

where sigmoid function is defined as:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}. \quad (5)$$

The sigmoid function can restrict the output value $f_H(I_i)$ in the range $[0, 1]$. The modified loss function is

$$\begin{aligned} \mathcal{L}(f_H(I_i), f_H(I_i^+), f_H(I_i^-)) \\ = \max(0, 1 - (\|f_H(I_i) - f_H(I_i^-)\|_2 \\ - \|f_H(I_i) - f_H(I_i^+)\|_2)) \end{aligned} \quad (6)$$

where $f_H(I_i), f_H(I_i^+), f_H(I_i^-) \in [0, 1]^q$.

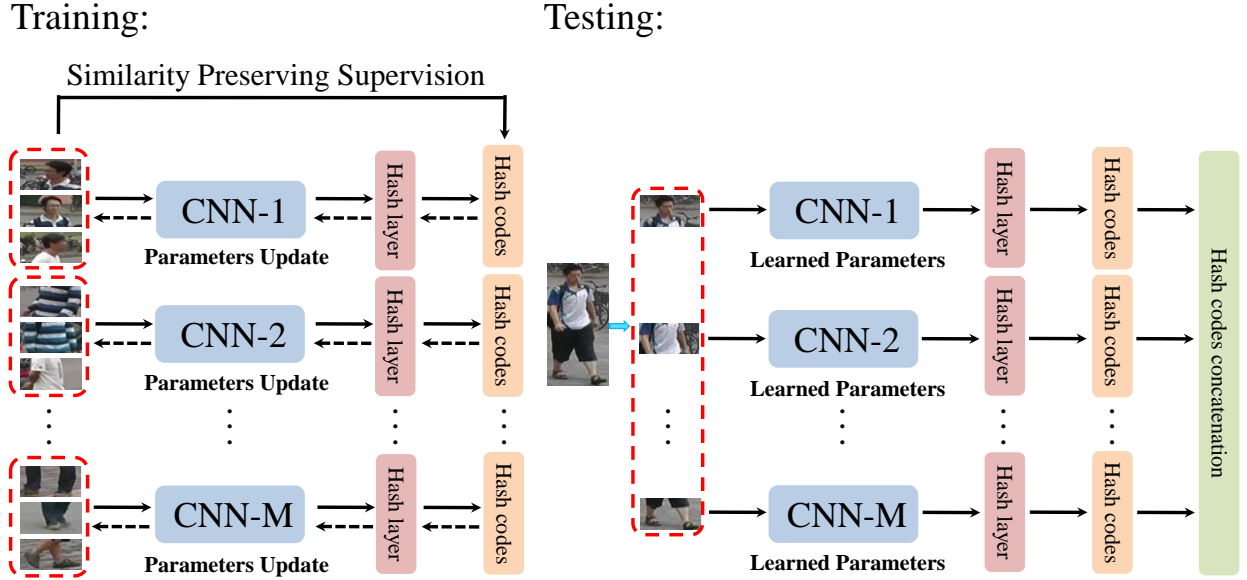


Fig. 3. The framework of the proposed Part-based Deep Hashing (PDH) model.

TABLE I

THE SIZE OF PARTS OF VARIOUS REGION PARTITIONING METHODS. “EQL” AND “UNEQL” REPRESENT DIVIDING THE IMAGE EQUALLY AND UNEQUALLY, RESPECTIVELY.

Region Partitions	Size
EQL 3 Parts	42×64; 42×64; 42×64;
UneQL 3 Parts	24×64; 56×64; 48×64;
Overlap 3 Parts	56×64; 56×64; 56×64;
EQL 4 Parts	32×64; 32×64; 32×64; 32×64;
UneQL 4 Parts	28×64; 40×64; 40×64; 20×64;
Overlap 4 Parts	48×64; 48×64; 48×64; 48×64;

In this way, the variant of triplet loss becomes a convex optimization problem. If the condition $1 - (\|f_H(I_i) - f_H(I_i^-)\|_2 - \|f_H(I_i) - f_H(I_i^+)\|_2) > 0$ is satisfied, their gradient values are as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial f_H(I_i)} &= 2(f_H(I_i^-) - f_H(I_i^+)) \\ \frac{\partial \mathcal{L}}{\partial f_H(I_i^+)} &= 2(f_H(I_i^+) - f_H(I_i)) \\ \frac{\partial \mathcal{L}}{\partial f_H(I_i^-)} &= 2(f_H(I_i^-) - f_H(I_i)). \end{aligned} \quad (7)$$

These gradient values can be fed into the deep CNN by the back propagation algorithm to update the parameters of each layer.

After the deep neural network model is trained, the new input pedestrian image I_j in query and testing set can be evaluated to generate hash code in the testing phase. The final binary representation of each image I_j is $\hat{H}(I_j)$, which is operated by simple quantization:

$$\hat{H}(I_j) = \text{sign}(f_H(I_j) - 0.5). \quad (8)$$

B. The Proposed Part-based Deep Hashing Model

Due to the intensely variation of pedestrian in cross-camera scenarios, the spatial information of the pedestrian image is

significant for enhancing the discriminative ability. A logical idea is to utilize the local part instead of the entire image to train the deep model. According to the consistency of person spatial information, we briefly make 6 part partitioning variants, which are listed in Table I. The direction of region partition is along with horizontal and from top to bottom. The examples of different part partitioning are shown in Fig. 2. The size of parts of various region partitioning methods is shown in Table I. We can train deep hashing model for each part separately instead of entire image. However, we do not know which part of the pedestrian image is more beneficial for training the deep hashing model. A simple strategy is to combine the results of each part with a uniform standard. To avoid complex calculation, we just divide the pedestrian image into a few parts. The number of the part for a pedestrian image and the trained deep CNN models is consistent. The architecture of proposed PDH method is shown in Fig. 3, which is on the basis of baseline triplet deep hashing model. The PDH method is as follows:

In the Training Phase, first, the training pedestrian image I_i is divided into a few parts. *i.e.* $I_i = \{I_{i,k}\}, k = 1, \dots, M$, where $I_{i,k}$ is the k -th part of pedestrian image I_i and M is the number of parts of one image.

Then, the same locations of pedestrian images constitute a specific part-based subset. The number of training samples is N . The total number of subset is M . The k -th subset is denoted as:

$$\text{subset} = \{I_{i,k}\}, i = 1, \dots, N. \quad (9)$$

Finally, for each subset, we train the deep CNN model using the samples of subset, and obtain the learned parameters of

each layers. The loss function is as follows:

$$\begin{aligned} \mathcal{L} & \left(f_H(I_{i,k}), f_H(I_{i,k}^+), f_H(I_{i,k}^-) \right) \\ & = \max(0, 1 - (\|f_H(I_{i,k}) - f_H(I_{i,k}^-)\|_2 \\ & \quad - \|f_H(I_{i,k}) - f_H(I_{i,k}^+)\|_2)) \end{aligned} \quad (10)$$

where $f_H(I_{i,k}), f_H(I_{i,k}^+), f_H(I_{i,k}^-) \in [0, 1]^q, k = 1, \dots, M$, and $f_H(I_{i,k}) = \text{sigmoid}(\mathbf{w}_k^T FC7(I_{i,k}))$. The training process of network is same as above baseline as shown in Section III-A. So a series of trained CNN models are obtained for corresponding to each part subset.

In the Testing Phase, first, the pedestrian images are also divided into several parts as same as the samples of training set.

Then, for the parts of new query and testing pedestrian image, we calculate the binary feature with the learned parameters of each layers. For the k -th part of pedestrian image I_j , the hash code is calculated as follows:

$$\hat{H}(I_{j,k}) = \text{sign}(f_H(I_{i,k}) - 0.5). \quad (11)$$

In this way, a group of hash codes is obtained for any parts of a single pedestrian image.

Finally, the hash codes of query and testing image I_j is represented by concatenating each part.

$$\hat{H}(I_j) = \text{concatenation}\{\hat{H}(I_{j,k})\}, k = 1, \dots, M. \quad (12)$$

In this way, we finish a hash codes conversion of local parts to global image. The new part-based hash codes can extract some rich and useful descriptors that retain the spatial information.

After the generation of hash codes for query and testing pedestrian image dataset, the person re-id is evaluated by calculating and sorting the Hamming distance between query and testing samples.

C. Multiple Queries

The motivation of multiple queries is that the intra-class variation of samples is taken into consideration. The strategy of multiple queries is to merge the query images which belong to same identity under a single camera into a single query for speed consideration. The method of multiple queries, which is more robust to pedestrian variations, has shown superior performance in image search [58] and person re-id [59]. We implement two pooling strategies, which are average pooling and max pooling, respectively. In average pooling, the feature vectors of multiple queries are pooled into one vector by averaged sum. In max pooling, the feature vectors of multiple queries are pooled into one vector by taking the maximum value in each dimension from all queries.

IV. EXPERIMENTS

In this section, we first describe the datasets and evaluation protocol. Then we evaluate the proposed PDH method and provide some comparisons with the state-of-the-art hashing and person re-id methods to demonstrate the effectiveness and efficiency of the PDH method.

TABLE II
BASELINE PERFORMANCE (RANK-1 ACCURACY (%) AND mAP (%))
WITH DIFFERENT HASH CODES LENGTHS ON MARKET-1501 AND
MARKET-1501+500K DATASETS.

Hash Codes Length	Market-1501		Market-1501+500K	
	r=1	mAP	r=1	mAP
128 bits	19.06	9.58	12.20	4.48
256 bits	21.91	10.75	14.82	5.33
512 bits	25.36	12.37	17.61	6.44
1,024 bits	25.24	11.95	16.63	5.91
2,048 bits	27.14	12.76	18.68	6.56

A. Datasets and Evaluation Protocol

This paper evaluates the performance of the proposed PDH method on the largest person re-id dataset: Market-1501 [10] and its associating distractor set with 500K images. The two datasets are denoted as: **Market-1501** and **Market-1501+500K**, respectively. The Market-1501 dataset contains 32,668 bounding boxes of 1,501 identities. There are 14.8 cross-camera ground truths for each query on average. The testing process is performed in a cross-camera mode. The distractor set contains 500K images which are treated as outliers besides the 32,668 bounding boxes of 1,501 identities. The Market-1501 is currently the largest person re-id dataset which is closer towards realistic situations than previous ones. We choose these two datasets due to their scales, for which effective retrieval methods are of great needs.

We adopt the Cumulated Matching Characteristics (CMC) curve and mean Average Precision (mAP) on Market-1501 and Market-1501+500K datasets. The CMC curve shows the probability that a query identity appears in the ranking lists of different sizes. The rank-1 accuracy (r=1) is shown when CMC curves are absent. The CMC is generally believed to focus on precision. In case of there is only one ground truth match for a given query, the precision and recall are the same. However, if multiple ground truths exist, the CMC curve is biased because recall is not considered. For Market-1501 and Market-1501+500K datasets, there are several cross-camera ground truths for each query. The mean Average Precision (mAP) is more suitable to evaluate the overall performance. The mAP considers both the precision and recall, thus providing a more comprehensive evaluation.

B. Experimental Results

1) *Performance of the Baseline Method:* We evaluate the baseline deep hashing model (described in Section III-A) trained by the entire pedestrian images. We observe from Table II that the baseline produces a relatively low accuracy on Market-1501 and Market-1501+500K datasets. Hash codes with various lengths are tested on the two datasets. It is shown from the results that longer hash codes generally yield higher re-id accuracy. The increase is most evident for shorter hash codes. For hash codes of more than 512 bits, re-id accuracy remains stable or witnesses some slight decrease. **As a trade-off between efficiency and accuracy, we use the 512 bits hash codes for each part-based deep CNN model in the following experiments.**

TABLE III

RANK-1 ACCURACY (%) AND MAP (%) FOR PART-BASED MODEL ON MARKET-1501 AND MARKET-1501+500K. “MQ” REPRESENTS MULTIPLE QUERIES. THE “AVG” AND “MAX” DENOTE AVERAGE AND MAX POOLING, RESPECTIVELY.

Methods	Market-1501						Market-1501+500K					
	Single Query		MQ avg		MQ max		Single Query		MQ avg		MQ max	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
Entire	27.14	12.76	32.54	15.39	30.17	14.66	18.68	6.56	21.91	7.82	19.69	7.31
3 Parts	43.05	21.80	49.52	26.89	46.70	25.25	32.28	13.21	39.34	17.26	36.46	15.72
4 Parts	47.24	24.94	57.13	31.03	54.39	29.74	37.23	16.38	45.72	21.26	43.53	19.93
5 Parts	46.23	23.72	53.86	29.46	51.07	28.12	37.08	15.44	44.63	20.61	41.48	19.10

TABLE IV

RANK-1 ACCURACY (%) AND MAP (%) FOR DIFFERENT REGION PARTITIONING STRATEGIES. “EQL” AND “UNEQL” DENOTE PARTITIONING IMAGES EQUALLY AND UNEQUALLY, RESPECTIVELY.

Partitions	Market-1501						Market-1501+500K					
	Single Query		MQ avg		MQ max		Single Query		MQ avg		MQ max	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
EQL 3 Parts	43.05	21.80	49.52	26.89	46.70	25.25	32.28	13.21	39.34	17.26	36.46	15.72
UnEQL 3 Parts	36.72	18.56	46.41	24.18	43.08	22.38	26.34	10.69	36.34	15.36	31.21	13.38
Overlap 3 Parts	47.36	25.47	53.36	30.29	50.86	28.54	37.47	16.19	42.70	19.96	39.61	18.40
EQL 4 Parts	47.24	24.94	57.13	31.03	54.39	29.74	37.23	16.38	45.72	21.26	43.53	19.93
UnEQL 4 Parts	46.17	24.31	54.69	30.48	51.57	29.16	35.99	15.47	44.80	21.29	42.04	20.02
Overlap 4 Parts	47.89	26.06	56.80	31.67	53.83	30.40	38.39	16.82	45.64	21.16	41.83	19.99

TABLE V

RANK-1 ACCURACY (%) AND MAP (%) OF EACH INDIVIDUAL PART AND CONCATENATION FOR DIFFERENT REGION PARTITIONING STRATEGIES. “EQL” AND “UNEQL” DENOTE PARTITIONING IMAGES EQUALLY AND UNEQUALLY, RESPECTIVELY.

Methods	Market-1501						Market-1501+500K					
	EQL 4 Parts		UnEQL 4 Parts		Overlap 4 Parts		EQL 4 Parts		UnEQL 4 Parts		Overlap 4 Parts	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
CNN-1 (Part 1)	6.83	3.21	5.82	2.75	11.43	4.89	3.33	1.07	2.58	0.87	6.65	1.81
CNN-2 (Part 2)	11.49	4.83	14.55	5.98	19.12	8.35	6.41	1.77	8.70	2.43	11.64	3.71
CNN-3 (Part 3)	10.66	5.13	7.54	3.59	19.69	9.15	5.85	1.97	4.04	1.33	11.52	4.06
CNN-4 (Part 4)	3.36	1.45	1.93	0.99	5.70	2.76	1.84	0.44	1.01	0.24	3.15	0.97
Concatenation	47.24	24.94	46.17	24.31	47.89	26.06	37.23	16.38	35.99	15.47	38.39	16.82

2) *Impact of Part Integration:* In Table III, we evaluate the impact of part integration on re-id accuracy, with a comparison with the baseline on the two re-id datasets. The entire pedestrian image is partitioned into several equal parts horizontally. We observe from Table III that when partitioning into 4 parts, mAP increases from 12.76% to 24.94% (+12.18%), and an even larger improvement can be seen from rank-1 accuracy, from 27.14% to 47.24% (+20.10%) on Market-1501 dataset. On Market-1501+500K dataset, mAP increases from 6.56% to 16.38% (+9.82%) with 4 parts, and for rank-1 accuracy, from 18.68% to 37.23% (+18.55%). This illustrates the effectiveness of the part integration over the baseline method. Moreover, we find that using more parts typically produces higher re-id performance, but again, the improvement tends to saturate after 4 parts.

We then evaluate multiple queries on the two re-id datasets. The experimental results demonstrate that the usage of multiple queries improves 4%~7% in mAP and 3%~10% in rank-1 accuracy. Moreover, multiple queries by average pooling is slightly superior to max pooling. The performance of part-based model increases significantly compared with the original general deep hashing model. These results demonstrate the effectiveness of part-based model and multiple queries for large-scale person re-id.

3) *Comparison of Different Part Partitioning Strategies:* The Section III-B describes 6 part partitioning variants. Specifi-

cally, the three types of part partitioning strategies are evaluated, including “Equally”, “Unequally” and “Overlap”. The height of the original pedestrian image is 128. The direction of region partition is along with horizontal. The partition details are listed in Table I.

In Table IV, we provide a comparison among these partitioning strategies. Results suggest that generating parts with overlap is an effective way of training the CNN model, probably because the overlaps provide some complementary information between two adjacent parts. We observe from Table IV that when using “Overlap 4 parts”, rank-1 accuracy increases from 47.24% to 47.89% (+0.65%), and an even larger improvement can be seen from mAP, from 24.94% to 26.06% (+1.12%) on Market-1501 dataset. On Market-1501+500K dataset, rank-1 accuracy increases from 16.38% to 16.82% (+0.44%) with 4 parts, and for mAP, from 37.23% to 38.39% (+1.16%). Meanwhile, the unequal part partition is inferior to equal parts, especially on the results of 3 parts. We speculate that the non-uniform operation separates some parts which have specific semantic meanings.

In order to further investigate the role of different individual parts, we evaluate the re-id performance of individual parts, and compare it with the concatenation of all parts. The hash code of each part is generated by the training CNN model at the corresponding regions. We observe from Table V that each individual CNN model produces a low accuracy on Market-

TABLE VI
RANK-1 ACCURACY (%) AND MAP (%) COMPARING WHETHER TO SHARE WEIGHTS AMONG THE PART SUB-NETWORKS. “EQL” AND “UNEQL”
DENOTE PARTITIONING IMAGES EQUALLY AND UNEQUALLY, RESPECTIVELY.

Methods	Market-1501						Market-1501+500K					
	EQL 4 Parts		UnEQL 4 Parts		Overlap 4 Parts		EQL 4 Parts		UnEQL 4 Parts		Overlap 4 Parts	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
share weights	41.42	19.57	33.97	16.13	42.96	21.70	32.39	12.40	25.21	9.36	34.09	13.74
not share weights	47.24	24.94	46.17	24.31	47.89	26.06	37.23	16.38	35.99	15.47	38.39	16.82

TABLE VII
RANK-1 ACCURACY (%) AND MAP (%) COMPARISON WITH THE STATE-OF-THE-ART HASHING METHODS.

Methods	Market-1501									
	128 bits		256 bits		512 bits		1,024 bits		2,048 bits	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
SH-CNN [48]	34.35	16.26	36.08	19.50	39.86	21.75	42.65	23.28	44.23	23.73
USPLH-CNN [60]	33.31	16.42	37.62	18.35	38.42	18.92	39.04	18.84	39.93	18.82
SpH-CNN [61]	35.33	16.65	37.96	19.99	41.38	22.56	44.26	24.04	44.62	24.54
DSH-CNN [62]	33.11	16.17	38.09	19.21	41.95	21.22	43.50	22.15	44.86	22.70
KSH-CNN [54]	41.33	20.62	43.55	23.40	44.23	24.41	45.27	24.90	46.13	25.01
SDH-CNN [63]	35.69	17.82	39.01	20.59	40.41	21.93	41.81	22.28	43.56	23.19
Zhang <i>et al.</i> [18]	15.50	8.50	18.38	9.48	22.24	11.07	21.91	10.47	23.43	11.29
Lin <i>et al.</i> [17]	8.91	4.89	18.65	10.01	28.98	16.39	41.12	24.14	49.79	30.29
Our PDH method	36.31	19.59	42.07	22.43	44.60	24.30	49.58	26.09	47.89	26.06

Methods	Market-1501+500K									
	128 bits		256 bits		512 bits		1,024 bits		2,048 bits	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
SH-CNN [48]	14.05	6.03	16.98	7.30	22.26	9.05	25.38	10.51	28.89	11.48
USPLH-CNN [60]	12.56	4.91	15.32	6.03	15.80	6.46	16.48	6.59	17.19	6.71
SpH-CNN [61]	17.96	6.80	23.40	9.17	26.51	10.65	28.30	11.60	29.60	12.08
DSH-CNN [62]	10.51	3.40	14.16	4.92	14.96	5.32	15.02	5.17	18.53	6.25
KSH-CNN [54]	30.94	11.86	35.39	13.99	36.28	14.88	37.62	15.17	37.44	15.37
SDH-CNN [63]	23.75	8.61	26.75	10.34	29.75	11.72	31.86	12.45	30.79	12.25
Zhang <i>et al.</i> [18]	9.71	3.65	11.19	4.25	14.49	5.20	13.66	4.70	14.52	5.24
Lin <i>et al.</i> [17]	5.34	1.99	10.90	4.41	18.85	7.83	28.92	13.15	37.41	18.26
Our PDH method	27.05	11.58	31.80	13.43	34.17	15.04	39.34	16.77	38.39	16.82

1501 and Market-1501+500K datasets, especially the CNN-1 and CNN-4 models. However, after the concatenation of hash codes for all the parts, the re-id accuracy is improved dramatically. The experimental results thus demonstrate that the part partitioning is effective in the proposed method.

The sub-networks proposed in this paper do not share weights, because the body parts are different in nature. To illustrate this point, we conduct experiments comparing whether to share weights among the part sub-networks. We train the CNN models using the weight-sharing network for the parts and provide some experimental results and comparisons in Table VI. It can be observed that the accuracy of weight-sharing network can be over 6% lower than training the sub-networks independently, which validating our assumption.

4) *Comparison with the State-of-the-art Hashing Methods:* In this section, we compare the proposed PDH method with some state-of-the-art hashing methods on Market-1501 and Market-1501+500K datasets. The compared hashing methods include Spectral Hashing (SH) [48], Unsupervised Sequential Projection Learning Hashing (USPLH) [60], Spherical Hashing (SpH) [61], Density Sensitive Hashing (DSH) [62], Kernel Supervised Hashing (KSH) [54], Supervised Discrete Hashing (SDH) [63] and two deep hashing methods [18], [17]. The first four methods are unsupervised and the others are supervised hashing methods. For the two comparison deep hashing methods [18] and [17], which are Siamese and identification CNN model, respectively. We use the image

pixels as input directly and implement it based on the Caffe [64] framework for deep hashing. The conventional non-deep hashing methods are evaluated based on the 4,096-D FC7 features in CaffeNet [55] pre-trained on the ImageNet [65] dataset and fine-tuned on the training set of the Market-1501 dataset for fair comparison. This feature is also called ID-discriminative Embedding (IDE) in [9].

Table VII summarizes the results of the state-of-the-art hashing methods with different code lengths on Market-1501 and Market-1501+500K datasets. Fig. 4 shows the CMC curve comparison of different hashing methods at 2,048 bits code length. First, it is evident that when longer hash codes are used, the rank-1 accuracy and mAP increase significantly. Second, compared with unsupervised hashing methods, the conventional non-deep supervised hashing methods (KSH and SDH) generally achieve better performance. Third, the bit-scalable deep hashing method [18] produces a relatively low accuracy, similar to the baseline in Section IV-B1. For [18], using 1,024 bits is inferior to using 512 bits in both rank-1 accuracy and mAP as shown in Table VII. We also notice a similar trend for the baseline method in Table II. In fact, it is common that the retrieval accuracy becomes saturated as the hash code grows longer, so after 512 bits, there might be some small fluctuations in the accuracy. Fourth, comparing with [17], we show that [17] produces a superior mAP at 2,048 bits. However, the rank-1 accuracy and mAP of [17] decline significantly with the decrease of the hash codes length, and

TABLE VIII
RANK-1 ACCURACY (%) AND MAP (%) COMPARISON WITH OTHER DEEP HASHING METHODS FOR DIFFERENT REGION PARTITIONING STRATEGIES.
“EQL” AND “UNEQL” DENOTE PARTITIONING IMAGES EQUALLY AND UNEQUALLY, RESPECTIVELY.

Methods	Market-1501						Market-1501+500K					
	EQL 4 Parts		UnEQL 4 Parts		Overlap 4 Parts		EQL 4 Parts		UnEQL 4 Parts		Overlap 4 Parts	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
Zhang <i>et al.</i> [18]	38.90	20.14	40.08	19.86	41.63	21.91	29.78	12.17	29.45	11.54	30.73	13.05
Lin <i>et al.</i> [17]	48.60	26.82	41.89	22.25	49.55	28.25	36.37	16.53	29.48	12.79	37.35	17.56
Our PDH Method	47.24	24.94	46.17	24.31	47.89	26.06	37.23	16.38	35.99	15.47	38.39	16.82

is inferior to our PDH method in these cases.

Compared with these hashing methods, the proposed PDH method produces a competitive performance *w.r.t.* rank-1 accuracy, mAP, and the CMC curve when 2,048 bits are used. Specifically, our method achieves **rank-1 accuracy = 47.89% and mAP = 26.06% on Market-1501 dataset, rank-1 accuracy = 38.39% and mAP = 16.82% on Market-1501+500K dataset**, respectively.

In order to further study the scalability of our part integration, we evaluate the part integration on above two deep hashing methods [18], [17] with a comparison of the proposed PDH method. We use the 512 bits hash vectors for each part-based deep CNN model. From the results in Table VIII, the accuracy of part integration on two deep hashing methods [18], [17] have increased their baselines (as shown in Table VII) by a large margin.

The experimental results demonstrate that our PDH method produce a competitive performance for large-scale person re-id. Moreover, the part integration has superior scalability on other deep hashing methods.

5) *Comparison with the State-of-the-art Person Re-id Methods*: We first compare with the Bag-of-Words (BOW) descriptor [10]. We only list the best result in [10]. As can be seen in Table IX, the proposed PDH method brings decent improvement of benchmark in both rank-1 accuracy and mAP. In addition, we compare with some existing metric learning methods based on BOW descriptor. The metric learning methods include LMNN [66], ITML [67] and KISSME [39]. From the results in Table IX, it is clear that the proposed PDH method significantly outperforms the traditional pipeline approaches, which demonstrates the effectiveness of proposed PDH method.

Then we compare with some state-of-the-art person re-id methods based on deep learning, including Multi-region Bilinear Convolutional Neural Networks method [45], PersonNet method [68], Semi-supervised Deep Attribute Learning (SSDAL) method [69], Temporal Model Adaptation (TMA) method [70] and End-to-end Comparative Attention Network (CAN) method [71]. From the results in Table IX, it is clear that the proposed PDH method significantly outperforms most of deep learning based re-id methods in both rank-1 accuracy and mAP. Only the PDH (MQ avg) is slightly inferior to Multi-region Bilinear DML (MQ avg) [45] in mAP. Nevertheless, the advantage of our method lies in the binary signatures, which enable fast person re-id in large galleries. In summary, PDH yields competitive accuracy on Market-1501, but has the advantage of computational and storage efficiency.

TABLE IX
RANK-1 ACCURACY (%) AND MAP (%) COMPARISON WITH THE STATE-OF-THE-ART PERSON RE-ID METHODS. NOTE THAT WE ONLY COMPARE WITH THE “OLDER” METHODS AS THIS WORK WAS DONE IN LATE 2015.

Methods	Market-1501	
	r=1	mAP
BOW+HS [10]	47.25	21.88
BOW+LMNN [66]	34.00	15.66
BOW+ITML [67]	38.21	17.05
BOW+KISSME [39]	39.61	17.73
Bit-scalable Deep Hashing [18]	23.43	11.29
Multiregion Bilinear (Single Query) [45]	45.58	26.11
Multiregion Bilinear (MQ avg) [45]	56.59	32.26
Multiregion Bilinear (MQ max) [45]	53.62	30.76
PersonNet (Single Query) [68]	37.21	18.57
SSDAL (Single Query) [69]	39.40	19.60
SSDAL (MQ avg) [69]	48.10	25.40
SSDAL (MQ max) [69]	49.00	25.80
TMA (Single Query) [70]	47.92	22.31
End-to-end CAN (Single Query) [71]	48.24	24.43
Our PDH (Single Query)	47.89	26.06
Our PDH (MQ max)	53.83	30.40
Our PDH (MQ avg)	56.80	31.67

6) *Comparison of Total Coding Time with Different Person Re-id Methods During the Testing Phase*: We compare the total coding time of PDH with two existing methods, including the 5,600-D BOW descriptor based on Color Names [10] and the 4,096-D IDE descriptor [9] (FC7 features in CaffeNet [55] pretrained on the ImageNet [65] dataset and fine-tuned on the training set of Market-1501 dataset). The coding time during the testing phase is composed of three aspects: 1) feature extraction, 2) average search (distance calculation), and 3) sorting. On the one hand, the computation of Hamming distance is much faster than Euclidean distance. On the other hand, using the Bucket sorting algorithm, the sorting complexity of PDH is $O(n)$, which is much lower than the baseline sorting complexity $O(n * \log(n))$ for floating-point vectors.

Table X presents the feature extraction, distance calculation, sorting and total coding time (millisecond (ms)) of the three methods on Market-1501 and Market-1501+500K datasets. The evaluation is performed on a server with GTX 1080 GPU (8G memory), 2.60 GHz CPU and 128 GB memory. The feature extraction time of the proposed PDH method is 32.8 ms, which is slower than IDE features due to the multiple parts evaluation. However, in practice, we can extract the features of each part for an image in parallel, and accelerate the feature extraction process of PDH method. Therefore, the disadvantage in the feature extraction time could be reduced. The search time of the PDH method is 0.98 ms and 26.2 ms

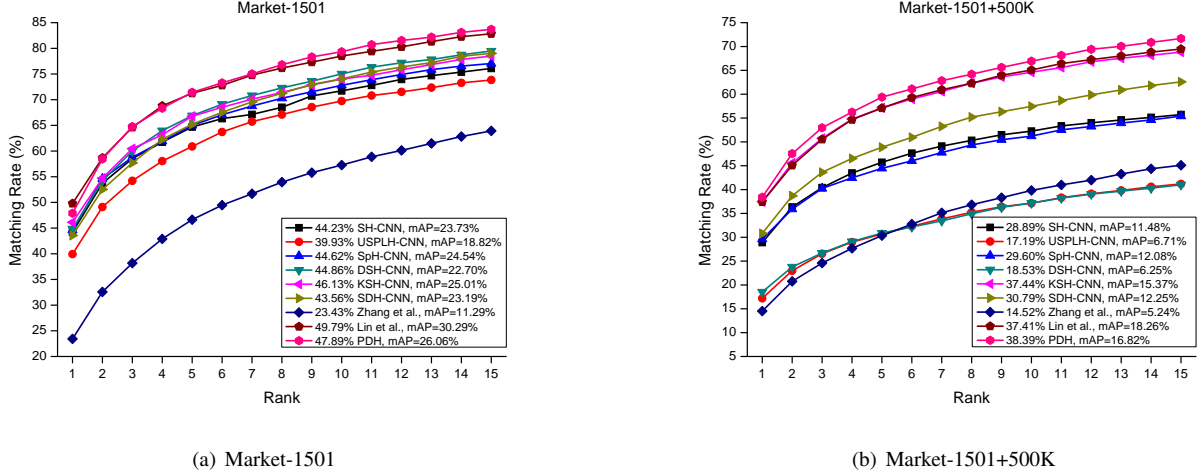


Fig. 4. CMC curves of the state-of-the-art hashing methods on Market-1501 and Market-1501+500K datasets.

TABLE X
FEATURE EXTRACTION, DISTANCE CALCULATION, SORTING AND TOTAL CODING TIME OF THE IDE, BOW FEATURES AND THE PROPOSED PDH ON MARKET-1501 AND MARKET-1501+500K (PARTITION BY “/”, MILLISECOND PER IMAGE) DATASETS.

Methods	Dim.	Data Type	Feature Extraction	Distance Calculation	Sorting	Total Coding Time
IDE (FC7) [9]	4,096	Float	8.3/8.3	97.9/2,470.8	3.5/134.5	109.7/2,613.6
BOW (CN) [10]	5,600	Float	264.3/264.3	139.9/3,587.9	4.9/156.1	409.1/4,008.3
Our PDH method	2,048	Bool	32.8/32.8	0.98/26.2	0.83/16.8	34.61/75.8

on the two re-id datasets, respectively. While, the sorting time of the PDH method is 0.83 ms and 16.8 ms, respectively, which is much faster than the other two float-point feature representations. From the total coding time comparison of three methods, the efficiency of the proposed PDH method should be justified. With the growth of the scale of person re-id datasets, binary representations will become more important.

V. CONCLUSIONS

In this paper, we employ the triplet-based deep hashing model and propose a **Part-based Deep Hashing (PDH)** framework for improving the efficiency and accuracy of large-scale person re-id, which generates hash codes for pedestrian images via a well designed part-based deep architecture. The part-based representation increases the discriminative ability of visual matching, and provides a significant improvement over the baseline. Multiple queries method is rewarding to improve the person re-id performance. The proposed PDH method demonstrates very competitive performance compared with state-of-the-art re-id methods on large-scale Market-1501 and Market-1501+500K datasets. There are several challenging directions along which we will extend this work. First, larger databases with millions of bounding boxes will be built which will fully show the strength of hashing methods. Second, more discriminative CNN models will be investigated to learn effective binary representations.

REFERENCES

- [1] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, “Image classification by cross-media active learning with privileged information,” *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016.
- [2] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, “Feature selection for multimedia analysis by sharing information among multiple tasks,” *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 661–669, 2013.
- [3] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, and C. Zhang, “Compound rank- k projections for bilinear analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1502–1513, 2016.
- [4] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, “Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval,” *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 437–446, 2008.
- [5] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, “A multimedia retrieval framework based on semi-supervised ranking and relevance feedback,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 723–742, 2012.
- [6] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” *arXiv preprint arXiv:1610.02984*, 2016.
- [7] W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in *Proc. CVPR*, 2014, pp. 152–159.
- [8] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *Proc. CVPR*, 2015, pp. 2197–2206.
- [9] L. Zheng, Z. Bie, Y. Sun, J. Wang, S. Wang, C. Su, and Q. Tian, “Mars: A video benchmark for large-scale person re-identification,” in *Proc. ECCV*, 2016, pp. 868–884.
- [10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proc. ICCV*, 2015, pp. 1116–1124.
- [11] R. Zhao, W. Ouyang, and X. Wang, “Person re-identification by salience matching,” in *Proc. ICCV*, 2013, pp. 2528–2535.
- [12] —, “Unsupervised salience learning for person re-identification,” in *Proc. CVPR*, 2013, pp. 3586–3593.
- [13] 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 *Proc. CVPR*, 2015, pp. 1741–1750.
- [14] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, “Supervised hashing for image retrieval via image representation learning,” in *Proc. AAAI*, 2014, pp. 2156–2162.
- [15] F. Zhao, Y. Huang, L. Wang, and T. Tan, “Deep semantic ranking based hashing for multi-label image retrieval,” in *Proc. CVPR*, 2015, pp. 1556–1564.

- [16] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. CVPR*, 2015, pp. 3270–3278.
- [17] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proc. CVPR Workshops*, 2015, pp. 27–35.
- [18] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4766–4779, 2015.
- [19] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2469–2479, 2016.
- [20] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. CVPR*, 2015, pp. 3908–3916.
- [21] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *arXiv preprint arXiv:1505.06821*, 2015.
- [22] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. International Workshop on Similarity-Based Pattern Recognition*, 2015, pp. 84–92.
- [23] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.
- [24] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, 2014, pp. 1386–1393.
- [25] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. CVPR*, 2014, pp. 1891–1898.
- [26] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.
- [27] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. CVPR*, 2017.
- [28] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, 2017.
- [29] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.
- [30] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," *arXiv preprint arXiv:1703.05693*, 2017.
- [31] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. ICCV*, 2015, pp. 3739–3747.
- [32] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. CVPR*, 2014, pp. 144–151.
- [33] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proc. ICCV*, 2015, pp. 3200–3208.
- [34] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proc. BMVC*, 2010.
- [35] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *Proc. BMVC*, 2012.
- [36] L. Bazzani, M. Cristani, and V. Murino, "Sdalf: modeling human appearance with symmetry-driven accumulation of local features," in *Person Re-Identification*. Springer, 2014, pp. 43–69.
- [37] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *Proc. AAAI*, 2015, pp. 2155–2161.
- [38] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proc. ECCV*, 2012, pp. 413–422.
- [39] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. CVPR*, 2012, pp. 2288–2295.
- [40] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Proc. IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2012, pp. 203–208.
- [41] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznaï, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*. Springer, 2014, pp. 247–267.
- [42] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for practical person re-identification," *arXiv preprint arXiv:1407.4979*, 2014.
- [43] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008, pp. 262–275.
- [44] M. Hirzer, C. Beleznaï, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.
- [45] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multiregion bilinear convolutional neural networks for person re-identification," *arXiv preprint arXiv:1512.05300*, 2015.
- [46] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. CVPR*, 2016, pp. 1288–1296.
- [47] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. CVPR*, 2016, pp. 1249–1258.
- [48] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. NIPS*, 2008, pp. 1753–1760.
- [49] Y. Gong and S. Lazebnik, "Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval," in *Proc. CVPR*, 2011, pp. 2916–2929.
- [50] J. Wang, S. Kumar, and S. F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. CVPR*, 2010, pp. 3424–3431.
- [51] M. Norouzi, "Minimal loss hashing for compact binary codes," in *Proc. ICML*, 2011, pp. 353–360.
- [52] Y. Yang, F. Shen, H. T. Shen, H. Li, and X. Li, "Robust discrete spectral hashing for large-scale image semantic indexing," *IEEE Transactions on Big Data*, vol. 1, no. 4, pp. 162–171, 2015.
- [53] Y. Yang, W. Chen, Y. Luo, F. Shen, J. Shao, and H. T. Shen, "Zero-shot hashing via transferring supervised knowledge," in *Proc. ACM International Conference on Multimedia*, 2016, pp. 1286–1295.
- [54] W. Liu, J. Wang, R. Ji, Y. G. Jiang, and S. F. Chang, "Supervised hashing with kernels," in *Proc. CVPR*, 2012, pp. 2074–2081.
- [55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
- [56] X. Chang, F. Nie, Y. Yang, C. Zhang, and H. Huang, "Convex sparse pca for unsupervised feature learning," *ACM Transactions on Knowledge Discovery from Data*, vol. 11, no. 1, pp. 3:1–3:16, 2016.
- [57] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE Transactions on Neural Networks and Learning Systems*, 2016.
- [58] R. Arandjelovic and A. Zisserman, "Multiple queries for large scale specific object retrieval," in *Proc. BMVC*, 2012, pp. 1–11.
- [59] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. CVPR*, 2010, pp. 2360–2367.
- [60] J. Wang, S. Kumar, and S. F. Chang, "Sequential projection learning for hashing with compact codes," in *Proc. ICML*, 2010, pp. 1127–1134.
- [61] J. P. Heo, Y. Lee, J. He, S. F. Chang, and S. E. Yoon, "Spherical hashing," in *Proc. CVPR*, 2012, pp. 2957–2964.
- [62] Z. Jin, C. Li, Y. Lin, and D. Cai, "Density sensitive hashing," *IEEE Transactions on Cybernetics*, vol. 44, no. 8, pp. 1362–1371, 2014.
- [63] F. Shen, C. Shen, W. Liu, and H. Tao Shen, "Supervised discrete hashing," in *Proc. CVPR*, 2015, pp. 37–45.
- [64] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM International Conference on Multimedia*, 2014, pp. 675–678.
- [65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.
- [66] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, 2005, pp. 1473–1480.
- [67] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. ICML*, 2007, pp. 209–216.
- [68] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.
- [69] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. ECCV*, 2016, pp. 475–491.
- [70] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," *arXiv preprint arXiv:1607.07216*, 2016.
- [71] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *arXiv preprint arXiv:1606.04404*, 2016.



Fuqing Zhu received his B.E. and M.S. degree from Dalian Jiaotong University, China, in 2010 and 2013, respectively. Currently, he is seeking his Ph.D. degree in School of Information and Communication Engineering at Dalian University of Technology, China. His research interests include multimedia retrieval, image classification and person re-identification.



Xiangwei Kong received her Ph.D. degree in Management Science and Engineering from Dalian University of Technology, China, in 2003. From 2006 to 2007, she was a visiting scholar in Department of Computer Science at Purdue University, USA. From 2014 to 2015, she was a senior research scientist in Department of Computer Science at New York University, USA. She is currently a professor in School of Information and Communication Engineering, and the director of research center of multimedia information processing and security at

Dalian University of Technology, China. She has published 4 edited books and more than 185 research papers in refereed international journals and conferences in the areas of cross-modal retrieval, multimedia information security, knowledge mining and business intelligence.



Liang Zheng received the Ph.D. degree in Electronic Engineering from Tsinghua University, China, in 2015, and the B.E. degree in Life Science from Tsinghua University, China, in 2010. He was a postdoc researcher in University of Texas at San Antonio, USA. He is currently a postdoc researcher in Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia. His research interests include image retrieval, classification, and person re-identification.



Haiyan Fu received her Ph.D. degree from Dalian University of Technology, China, in 2014. She is currently an associate professor in School of Information and Communication Engineering at Dalian University of Technology, China. Her research interests are in the areas of image retrieval and computer vision.



Qi Tian (M'96-SM'03-F'16) received his Ph.D. in ECE from University of Illinois at Urbana-Champaign (UIUC) in 2002 and received his B.E. in Electronic Engineering from Tsinghua University in 1992 and M.S. in ECE from Drexel University in 1996, respectively. He is currently a Full Professor in the Department of Computer Science, the University of Texas at San Antonio (UTSA). He was a tenured Associate Professor from 2008-2012 and a tenure-track Assistant Professor from 2002-2008. During 2008-2009, he took one-year Faculty Leave at Microsoft Research Asia (MSRA) as Lead Researcher in the Media Computing Group.

Dr. Tian's research interests include multimedia information retrieval, computer vision, pattern recognition and bioinformatics and published over 350 refereed journal and conference papers. He was the co-author of a **Best Paper** in ACM ICMR 2015, a **Best Paper** in PCM 2013, a **Best Paper** in MMM 2013, a **Best Paper** in ACM ICIMCS 2012, a **Top 10% Paper Award** in MMSP 2011, a **Best Student Paper** in ICASSP 2006, and co-author of a **Best Student Paper Candidate** in ICME 2015, and a **Best Paper Candidate** in PCM 2007. Dr. Tian research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALS, CIAS, Akiira Media Systems, HP, Blippar and UTSA. He received 2014 Research Achievement Awards from College of Science, UTSA. He received **2016 UTSA Innovation Award** and **2010 ACM Service Award**. He is the associate editor of IEEE Transactions on Multimedia (TMM), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Multimedia System Journal (MMSJ), and in the Editorial Board of Journal of Multimedia (JMM) and Journal of Machine Vision and Applications (MVA). Dr. Tian is the Guest Editor of IEEE Transactions on Multimedia, Journal of Computer Vision and Image Understanding, etc.