Deep Representation Learning With Part Loss for Person Re-Identification

Hantao Yao, Shiliang Zhang, *Member, IEEE*, Richang Hong, *Member, IEEE*, Yongdong Zhang, *Senior Member, IEEE*, Changsheng Xu, *Fellow, IEEE*, and Qi Tian, *Fellow, IEEE*

***Abstract*—Learning discriminative representations for unseen person images is critical for person re-identification (ReID). Most of the current approaches learn deep representations in classification tasks, which essentially minimize the empirical classification risk on the training set. As shown in our experiments, such representations easily get over-fitted on a discriminative human body part on the training set. To gain the discriminative power on unseen person images, we propose a deep representation learning procedure named part loss network, to minimize both the empirical classification risk on training person images and the representation learning risk on unseen person images. The representation learning risk is evaluated by the proposed part loss, which automatically detects human body parts and computes the person classification loss on each part separately. Compared with traditional global classification loss, simultaneously considering part loss enforces the deep network to learn representations for different body parts and gain the discriminative power on unseen persons. Experimental results on three person ReID datasets, i.e., Market1501, CUHK03, and VIPeR, show that our representation outperforms existing deep representations.**

***Index Terms*—Person re-identification, representation learning, part loss networks, convolutional neural networks.**

I. Introduction

PERSON Re-Identification (ReID) targets to identify a probe person appeared under multiple cameras. More specifically, person ReID can be regarded as a zero-shot learning problem, because the training and test sets do not share any person in common. As illustrated in Fig. 1, person images taken by different cameras could also be easily affected by variances of camera viewpoint, human pose, illumination, occlusion, *etc*. Consequently, person ReID is a challenging problem.

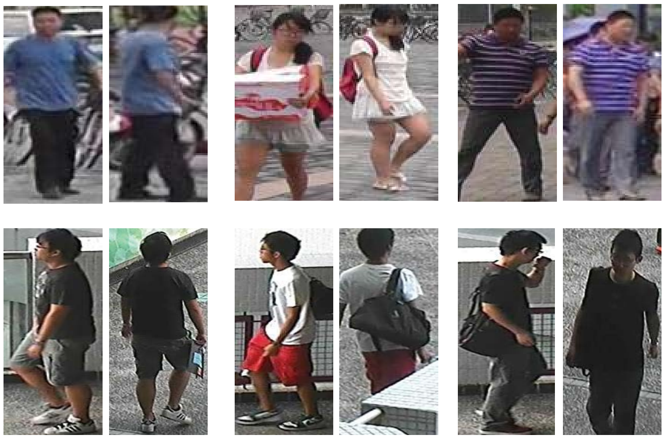


Fig. 1. Example images of three persons from Market1501 (first row) and CUHK03 (second row), respectively. The subtle differences among different persons and the large variance among images of the same person make person ReID challenging.

Existing approaches conquer this challenge by either seeking discriminative metrics [1]–[11], or generating discriminative features [12]–[19]. Inspired by the success of Convolutional Neural Network (CNN) in large-scale visual classification [20], lots of approaches have been proposed to generate representations based on CNN [19], [21]–[30]. For example, several works [22], [31], [32] employ deep classification model to learn representations. More detailed reviews on deep learning based person ReID will be given in Sec. II.

Notwithstanding the success of these approaches, we argue that the representations learned by current deep classification models are not optimal for zero-shot learning problems like person ReID. Most of current deep classification models learn representations by minimizing the classification loss on the training set. Differently, the optimal representation of person ReID is expected to maximize the discriminative power to unseen person images. Different optimization objectives make current deep representations perform promisingly on

Manuscript received April 29, 2018; revised November 8, 2018 and December 26, 2018; accepted December 26, 2018. Date of publication January 10, 2019; date of current version March 21, 2019. This work was supported in part by the National Postdoctoral Programme for Innovative Talents, in part by the National Nature Science Foundation of China under Grants 61525206, 61532009, 61721004, U1705262, 61572050, 91538111, 61620106009, and 61532009, and in part by the Key Research Program of Frontier Sciences, CAS, under Grant QYZDJ-SSW-JSC039. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Liang Wang. *(Corresponding author: Hantao Yao.)*

H. Yao is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: hantao.yao@nlpr.ia.ac.cn).

S. Zhang is with the School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: slzhang.jdl@pku.edu.cn).

R. Hong is with the Department of Computer Science and Technology, University of Technology, Hefei 230009, China (e-mail: hongrc@hfut.edu.cn).

Y. Zhang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhyd@ict.ac.cn).

C. Xu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Department of Artificial Intelligence, University of the Chinese Academy of Sciences, Beijing 100049, China (e-mail: csxu@nlpr.ia.ac.cn).

Q. Tian is with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249-1604 USA (e-mail: qi.tian@utsa.edu).

Digital Object Identifier 10.1109/TIP.2019.2891888

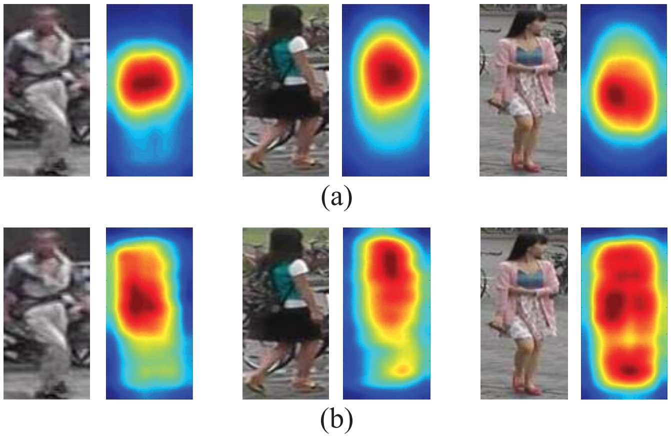


Fig. 2. Saliency maps of CNN learned in traditional classification network (a), and part loss networks (PL-Net) (b). The salient region reveals the body part that the CNN representation focuses on. Representations of our PL-Net are more discriminative to different parts.

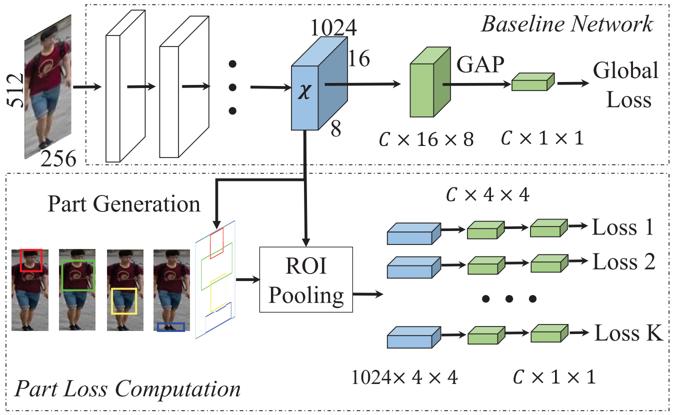


Fig. 3. Overview of Part Loss Network (PL-Net), which is composed of a baseline network and a part loss computation extension. "GAP" denotes the Global Average Pooling. Given an input image, we firstly extract its feature maps $\mathcal{X}$, then compute the global loss and person part loss based on $\mathcal{X}$. The person part loss is computed on K parts generated with an unsupervised method.

traditional classification tasks, but might be not optimal to depict and distinguish unseen person images.

Observations from our experiments are consistent with the above discussions. As shown in Fig. 2(a), the representations generated by deep classification model mainly focus on one body region, *i.e.*, the upper body, and ignore the other body parts. This is reasonable because, guided by the classification loss minimization, deep network tends to select the most discriminative features from the training set and ignore the others, *e.g.*, the upper body conveys more distinct clothing cues than the other parts. However, the other parts like head, lower body, and foot, are potential to be meaningful to describe the unseen persons. Ignoring such parts essentially increases the *risk of representation learning* for unseen persons.

The above observations motivate us to study more reliable deep representations for person ReID. We are inspired by the structural risk minimization principle in SVM [33], which imposes strict constraint by maximizing the classification margin. Similarly, we enforce the network to learn better representation with extra representation learning risk minimization constraint. Specifically, the representation learning risk is evaluated by the proposed part loss, which automatically generates K parts for an image, and computes the person classification loss on each part separately. In other words, the network is trained to focus on every body part and learn representations for each of them. As illustrated in Fig. 2(b), minimizing the person part loss guides the deep network to learn discriminative representations for different body parts, hence avoids overfitting on a specific body part and decreases the representation learning risk for unseen persons.

We propose Part Loss Network (PL-Net) structure that can be optimized accordingly. As shown in Fig. 3, PL-Net is composed of a baseline network and an extension to compute the person part loss. It is trained to simultaneously minimize the part loss and the global classification loss. Experiments on three public datasets, *i.e.,* Market1501, CUHK03, VIPeR show PL-Net learns more reliable representations and achieves promising performance compared with state-of-the-arts. It also should be noted that, PL-Net is easy to repeat because it only has one important parameter to tune, *i.e.*, the number of generated parts K.

Most of previous person ReID works directly train deep classification models to extract image representations. To our best knowledge, this work is an original effort discussing the reasons why such representations are not optimal for person ReID. Representation learning risk and part loss are hence proposed to learn more reliable deep representations to depict unseen person images. The proposed PL-Net is simple but achieves promising performance. It may inspire future research on zero-shot learning for person ReID.

II. Related Work

In the past several years, a lot of deep learning-based methods have been proposed for person ReID. This section briefly reviews those works.

The promising performance of CNN on large-scale ImageNet classification indicates that the classification network extracts discriminative image features. Therefore, several works [22], [31], [32], [34] employ the classification network fine-tuned on target datasets as the feature extractor for person ReID. For example, Xiao *et al.* [22] propose a novel dropout strategy to train a classification model with multiple datasets jointly. Zheng *et al.* [31] extract features with deep classification network to perform person ReID. Wu *et al.* [32] combine the hand-crafted histogram features and Convolutional Neural Network (CNN) features to fine-tune the classification network.

The classification network commonly needs a lot of training samples for fine-tuning. This conflicts with the fact that, most of current person ReID datasets are small-scale. The siamese network takes a pair of images as input, and is trained to verify the similarity between those two images. Therefore, pair-wise verification is also a good choice for person ReID. There exist several works [24], [25], [35]–[38] that use siamese network to test whether the two input images contain the same person. Ahmed *et al.* [24] employ the siamese network to infer the description and a corresponding similarity metric simultaneously. Shi *et al.* [38] replace the Euclidean distance metric with Mahalanobis distance metric in the siamese network. Yi *et al.* [37] jointly learn the color feature, texture descriptor, and distance metric in a siamese deep neural network. Zheng *et al.* [25] propose another network by jointly considering the objective functions of classification and similarity learning. Varior *et al.* [36] combine the LSTM and siamese network architecture for person ReID. Wu *et al.* [35] propose a verification network to simultaneously learn high-level features and a corresponding similarity metric for person ReID.

The siamese network is trained with known pair-wise similarity, which could be too strict and hard to collect. Therefore, some researchers study to train the network with relative similarity among three images, named as triplet. Some works [26], [39], [40] employ the triplet networks to learn the discriminative description for person ReID. Cheng *et al.* [39] propose a multi-channel parts-based CNN model for person ReID. Liu *et al.* [40] propose an end-to-end Comparative Attention Network to generate image description. Su *et al.* [26] 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.

Body part provides extra local details of a pedestrian. Therefore, many works are proposed to fuse body part representations for person ReID [27], [41]–[44]. Some works use person parts generated by existing methods, such as Convolutional Pose Machines (CPM) [45]. For example, Zhao *et al.* [27] and Su *et al.* [41] firstly extract human body parts with fourteen body joints, then fuse features extracted on body parts with global feature. Some other works apply the attention strategy to automatically generate local parts. For example, Li *et al.* [42] employ Spatial Transform Network (STN) [46] for part localization, and propose Multli-Scale Context-Aware Network to infer representations on the generated local parts.

Our work also uses body parts during feature learning, but with the motivation different from feature fusion. Our motivation is that, the deep features trained with classification loss tend to over-fit to one body region, thus are not discriminative to identify unseen persons. We hence extract body parts, design the part loss, and guide the network to learn a more reasonable representation for unseen persons as illustrated in Fig. 2(b). As shown by our experimental results, part loss substantially boosts the performance of global body feature, showing that our performance is improved without fusing part features. Compared with the work from Zhao *et al.* [27] and Su *et al.* [41], PL-Net generates body parts without requiring any pose estimation models or manual annotations on body key-points. The Spindlenet [27] generates body parts using pose estimation model proposed in [45], which needs additional human landmarks annotations.

MSCAN [42] also automatically detects body parts, but with a different strategy, *i.e.*, the Spatial Transformer Network (STN). To help STN localize body parts, the authors proposed three prior constraints. Differently, our part is detected by the activations in the feature maps and does not need any prior constraint. Our motivation and part detection strategy highlight our differences with existing works using body part for person ReID.

III. Methodology

A. Formulation

Given a probe person image I_q, person ReID targets to return images containing the identical person in I_q from a gallery set G. We denote the gallery set as $G = \{I_i\}, i \in [1, m]$, where m is the total number of person images. Person ReID can be tackled by learning a discriminative feature representation $\mathbf{f}$ for each person image from a training set T. Therefore, the probe image can be identified by matching its $\mathbf{f}_q$ against the gallery images.

Suppose the training set contains n labeled images from C persons, we denote the training set as $T = \{I_i, y_i\}, i \in [1, n], y_i \in [1, C]$, where I_i is the i-th image and y_i is its person ID label. Note that, person ReID assumes the training and gallery sets contain distinct persons. Therefore, person ReID can be regarded as a zero-shot learning problem, *i.e.*, the ID of probe person is not included in the training set.

Currently, some methods [22], [25], [32] fine-tune a classification-based CNN to generate the feature representation. The feature representation learning can be formulated as updating the CNN network parameter θ by minimizing the empirical classification risk of representation $\mathbf{f}$ on T through back prorogation. We denote the empirical classification risk on T as,

$$\mathcal{J} = \frac{1}{n}[\sum_{i=1}^{n} L^g(\hat{y}_i)], \tag{1}$$

where $\hat{y}_i$ is the predicted classification score for the i-th training sample, and $L^g(\cdot)$ computes the classification loss for each training image. We use the superscript ^{g} to denote it is computed on the global image. The predicted classification score $\hat{y}_i$ can be formulated as, *i.e.*,

$$\hat{y}_i = \mathbf{W}\mathbf{f}_i + b, \tag{2}$$

where $\mathbf{W}$ denotes the parameter of the classifier in CNN, *e.g.*, the weighting matrix in the fully connected layer.

Given a new image I_q, its representation $\mathbf{f}_q$ is hence extracted by CNN with the updated parameter θ, *i.e.*,

$$\mathbf{f}_q = \mathbf{CNN}_\theta(I_q). \tag{3}$$

It can be inferred from Eq. (1) and Eq. (2) that, to improve the discriminative power of $\mathbf{f}_i$ during training, a possible way is to restrict the classification ability of $\mathbf{W}$. In another word, a weaker $\mathbf{W}$ would enforce the network to learn more discriminative $\mathbf{f}_i$ to minimize the classification error. This motivates us to introduce a baseline CNN network with weak classifiers. Details of this network will be given in Sec. III-B

It also can be inferred from Eq. (1) that, minimizing the empirical classification risk on T results in a discriminative representation $\mathbf{f}$ for classifying the seen categories in T. For example in Fig. 2(a), the learned representations focus on discriminative parts for training set. However, such representations lack the ability to describe other parts like head, lower-body, and foot which could be meaningful to distinguish an unseen person. Therefore, more parts should be depicted by the network to minimize the risk of representation learning for unseen data.

Therefore, we propose to consider the representation learning risk, which tends to make the CNN network learn discriminative representation for each part of the human body. We denote the representation of each body part as $\mathbf{f}^k, k \in [1, K]$, where K is the total number of parts. The representation learning risk $\mathcal{P}$ can be formulated as,

$$\mathcal{P} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}[\sum_{i=1}^{n} L^p(\hat{y}_i^k)], \tag{4}$$

where $L^p(\cdot)$ computes the part loss, *i.e.*, the classification loss on each part. $\hat{y}_i^k$ is the predicted person classification score for the i-th training sample by the representation of k-th part. $\hat{y}_i^k$ is computed with,

$$\hat{y}_i^k = \mathbf{W}^k\mathbf{f}_i^k + b^k, \tag{5}$$

where $\mathbf{W}^k$ denotes the classifier for the representation of the k-th part.

The representation learning risk monitors the network and enforces it to learn discriminative representation for each part. It shares a certain similarity with the structural risk minimization principle in SVM [33], which also imposes more strict constraints to enforce the classifier to learn better discriminative power.

The final part loss networks (PL-Net) model could be inferred by minimizing the empirical classification risk and the representation learning risk simultaneously, *i.e.*,

$$\theta = \arg\min(\mathcal{J} + \mathcal{P}). \tag{6}$$

During the training stage, we set $\mathcal{J}$ and $\mathcal{P}$ with the same weight. This is equivalent to setting the weight for $\mathcal{J}$ as 1, and the weight for each local part loss as $1/K$, where K is the number of body parts. In the following parts, we proceed to introduce the Part Loss Network(PL-Net) and the computation of part loss.

B. Part Loss Network

PL-Net is extended from our baseline network. Our baseline network replaces the fully-connected classifier with a convolutional layer and a Global Average Pooling (GAP) layer. As shown in Fig. 3, the convolutional layer directly generates C activation maps explicitly corresponding to C classes. Then GAP generates the classification score for each class, *i.e.*,

$$s_c = \frac{1}{W \times H}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathcal{C}_c(h, w), \tag{7}$$

where s_c is the average response of the c-th activation map $\mathcal{C}_c$ with size $W \times H$, and $\mathcal{C}_c(h, w)$ denotes the activation value on the location (h, w) on $\mathcal{C}_c$. s_c is hence regarded as the classification score for the c-th class.

This modified baseline network uses a Convolutional + GAP structure to generate classification scores. Compared with traditional GAP with Fully Connected layer (FC) classifier structure, our Convolutional + GAP structure generates spatial activation maps explicitly corresponding to object categories, *i.e.*, each class explicitly corresponds to one activation map. As a consequence, the activation map of each class could be visualized to show the cues learned by the deep model, *e.g.*, showing the most discriminative regions of an image. Moreover, our baseline network could preserve the spatial activations on feature maps, which is useful for our part localization.

According to Eq. (6), our representation is learned to minimize both the empirical classification risk and the representation learning risk. We thus extend the baseline network accordingly to make it can be optimized by these two types of supervisions. The overall PL-Net is shown in Fig. 3. Specifically, PL-Net processes the input image and generates feature maps. During training, it computes a person part loss and a global loss. The global loss is computed by the baseline network to minimize the empirical classification risk. We denote the feature maps of the last convolutional layer before the classification module as $\mathcal{X} \in \mathcal{R}^{Z\times H\times W}$. For example, $Z = 1024$, $H = 16$, $W = 8$ when we input a 512×256 sized image into the baseline network modified from GoogleNet [47]. After obtaining $\mathcal{X}$, the global loss is calculated as,

$$L^g(\hat{y}_i) = -\sum_{c=1}^{C} 1\{y_i = c\}\log\frac{e^{\hat{y}_i}}{\sum_{l=1}^{C} e^{\hat{y}_l}}. \tag{8}$$

The part loss is computed on each automatically generated part to minimize the representation learning risk. The network first generates K person parts based on $\mathcal{X}$ in an unsupervised way. Then part loss is computed on each part. The following part gives details of the unsupervised part generation and part loss computation.

C. Person Part Loss Computation

Person parts can be extracted by various methods. For instance, detection models could be trained with part annotations to detect part locations. However, those methods [27], [30] require extra annotations that are hard to collect. We thus propose an unsupervised part generation algorithm that can be optimized together with the representation learning procedure.

Wei *et al.* [48] show that simply average pooling the feature maps of convolutional layers generates a saliency map. The saliency essentially denotes the "focused" regions by the neural network. Fig. 4 shows several feature maps generated by a CNN trained in the classification task. It can be observed that, different feature maps have different spatial attention locations and activation values. For example, the lower part of the body has substantially stronger activations. As illustrated in Fig. 4, for the CNN trained in the classification task, simply average pooling all of those feature maps gathers together the activations on discriminative region and suppresses the activations of other regions.

Although the responses on different parts are seriously imbalanced, they still provide cues of different part locations. By clustering feature maps based on the locations of their maximum responses, we could collect feature maps depicting different body parts. Individually pooling those feature map clusters indicates the part locations. As shown in Fig. 4 and Fig. 5, the four saliency maps focus on head, upper body,

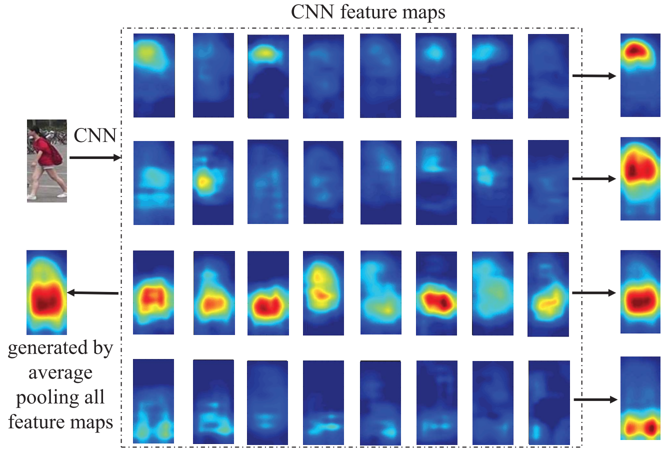


Fig. 4. Examples of CNN feature maps and generated saliency maps. The saliency map generated on all feature maps focuses on one part and suppresses the activations on other parts. The four saliency maps on the right side are generated by average pooling four types of feature maps based spatial attention consistency respectively. They clearly indicate different part locations.

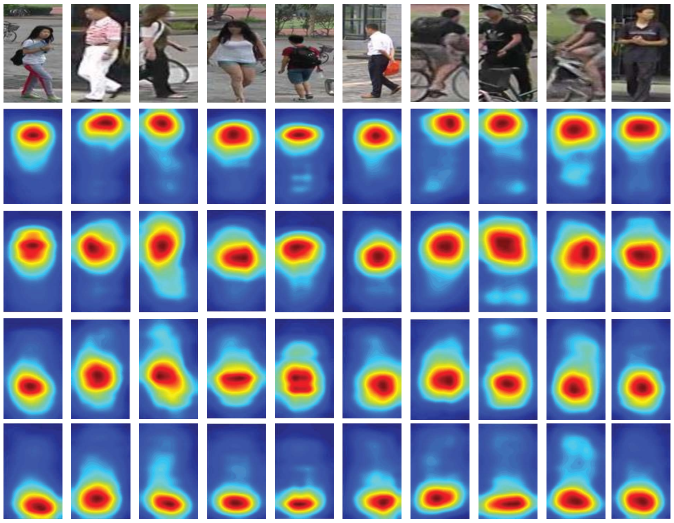


Fig. 5. Examples of generated saliency maps on four parts across different persons. The first row is the original image, and the other four rows illustrate the saliency maps on four parts, *i.e.*, head, upper body, lower body, and foot, respectively.

lower body, and foot, respectively. This might be because the head, lower body, and foot are also provide discriminative visual cues for distinguishing persons, thus CNN still learns certain neurons to depict them.

The above observation motivates our unsupervised part generation. Assume that we have got the feature map $\mathcal{X}$, we first compute the position of maximum activation on each feature map, denoted as $(h_z, w_z), z \in [1, Z]$,

$$(h_z, w_z) = \arg\max_{h,w} \mathcal{X}_z(h, w), \tag{9}$$

where $\mathcal{X}_z(h, w)$ is the activation value on location (h, w) in the z-th channel of $\mathcal{X}$.

As illustrated in Fig. 6, the feature maps are divided into K groups according to their maximum activation locations for part generation. Feature maps in each group are expected to show similar maximum activation locations. This is implemented by clustering their maximum activation locations into K clusters using L2 distance and K-means clustering. From each feature map group, we generate one part bounding box. Specifically, we average pooling the feature maps in each group and apply the max-min normalization to produce a saliency map. A threshold, *e.g.*, 0.5, is set to turn each saliency map into a binary image. For each binary image, we treat its minimum enclosing rectangle as the part bounding box. Examples of generated parts are shown in Fig. 8. The parameter K decides the number of generated body parts. It will be studied in our experiments.

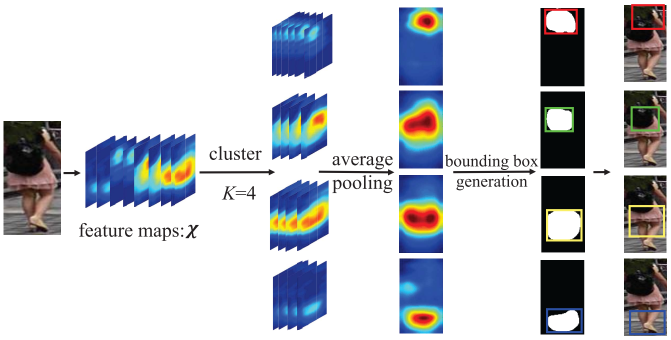


Fig. 6. Illustration of the procedure for unsupervised person part generation.

After obtaining the part bounding boxes, we proceed to compute the part loss. Inspired by Fast R-CNN [49], we employ the RoI pooling to convert the responses of $\mathcal{X}$ inside each part bounding box into a new feature map $\mathcal{X}^k \in \mathcal{R}^{Z\times H'\times W'}$ with a fixed spatial size, *e.g.,* $H' = W' = 4$ in this work. Based on those feature maps, we compute the part loss $L^p(\cdot)$ for k-th part with a similar procedure of global loss computation, *i.e.*,

$$L^p(\hat{y}_l^k) = -\sum_{c=1}^{C} 1\{y_i = c\} \log \frac{e^{\hat{y}_i^k}}{\sum_{l=1}^{C} e^{\hat{y}_l^k}}. \tag{10}$$

Similar to the notations in Eq. (4), $\hat{y}_i^k$ is the predicted person classification score of the i-th training sample based on the representation of its k-th part.

We next discuss the rationality of our part localization strategy. The intuition behind this strategy is that, CNN training enforces convolutional kernels to capture discriminative parts for person classification. As a result, the discriminative body parts are activated on feature maps by CNN kernels, *e.g.*, as shown in Fig. 4, the lower body is activated because it provides more distinctive cues for person classification. In other words, the activation on a local part would be stable, if there are well-trained CNN kernels to depict it. Our part loss essentially minimizes the training error on each part, this generates well-trained CNN kernels to depict each part, and in-turn boosts the stability of activations on different body parts. Therefore, the accuracy of our unsupervised part generation is related with the representation learning performance.

The generated parts are updated on each iteration of network training. As more discriminative part features are learned, more stable part locations can be detected. For example in Fig. 4, if more neurons are trained to depict parts like head and foot during representation learning, more feature maps would focus on these parts. This in turn improves the feature maps clustering and results in more accurate bounding boxes for head and foot. Existing works [48], [50]–[52] also demonstrate that feature map could be utilized to localize parts. The validity of our part localization strategy will be tested in our experiments.

Another property of PL-Net is that, it is an end-to-end differentiable network. For the part loss computation, PL-Net firstly generates the bounding box for each local part, then employs RoI pooling [49] to convert the response inside each part bounding box into a new feature map. During the back propagation, the gradient of RoI pooling is only computed for the features inside the bounding boxes rather than the whole original feature map, making the body part branches in PL-Net also differentiable.

D. Person ReID

On the testing phase, we extract feature representation from the trained neural network for person ReID. We use the feature maps $\mathcal{X}$ to generate the global and part representations for similarity computation.

Given a person image I, we firstly resize it to the size of 512×256, then fed it into network to obtain the feature maps $\mathcal{X}$. We hence compute the global representation $\mathbf{f}^{(g)}$ with Eq. (11),

$$\mathbf{f}^{(g)} = [f_1, ..., f_z, ...f_Z], \tag{11}$$

$$f_z = \frac{1}{W \times H} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathcal{X}_z(h, w). \tag{12}$$

For the part representation, we obtain the feature maps after RoI pooling for each part, denoted as $\mathcal{X}^k \in \mathcal{R}^{Z \times 4 \times 4}$, $k \in [1, K]$. For each $\mathcal{X}^k$, we calculate the part description $\mathbf{f}^k$ in similar way with Eq. (11). The final representation is the concatenation of global and part representations, *i.e.*,

$$\mathbf{f} = [\mathbf{f}^{(g)}, \mathbf{f}^1, ..., \mathbf{f}^K]. \tag{13}$$

IV. Experiments

A. Datasets

We verify the proposed part loss networks (PL-Net) on three datasets: VIPeR [53], CUHK03 [19], and Market1501 [54]. VIPeR [53] contains 632 identities appeared under two cameras. For each identity, there is one image for each camera. The dataset is split randomly into equal halves and cross camera search is performed to evaluate the algorithms.

CUHK03 [19] consists of 14,097 cropped images from 1,467 identities. For each identity, images are captured from two cameras and there are about 5 images for each view. Two ways are used to produce the cropped images, *i.e.*, human annotation and detection by Deformable Part Model (DPM) [55]. Our evaluation is based on the human annotated images. We use the standard experimental setting [19] to select 1,367 identities for training, and the rest 100 for testing.

Market1501 [54] contains 32,668 images from 1,501 identities, and each image is annotated with a bounding box detected by DPM. Each identity is captured by at most six cameras. We use the standard training, testing, and query split provided by Zheng *et al.* [54]. The Rank-1, Rank-5, Rank-10 accuracies are evaluated for VIPeR and CUHK03. For Market1501, we report the Rank-1 accuracy and mean Average Precision (mAP).

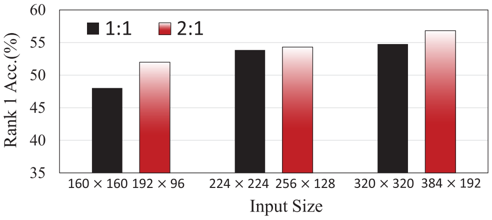


Fig. 7. Effects of the hight-width ration of input image to baseline network performance on CUHK03.

B. Implementation Details

We use Caffe [56] to implement and train the part loss networks (PL-Net). The baseline network is modified from second version of GoogLeNet [57]. Following the *inception5b/output* layer, an 1×1 convolutional layer with the output of $\mathcal{C}$ channels is used to generate the category confidence map. For the training, we use the pre-trained model introduced in [58] to initialize the PL-Net, and use a step strategy with mini-batch Stochastic Gradient Descent (SGD) to train the neural networks on Tesla K80 GPU. Parameters like the maximum number of iterations, learning rate, step size, and gamma are set as 50,000, 0.001, 2500, and 0.75, respectively. For the person images, we first resize their size to 512×256, and then feed their into the PL-Net for training and testing.

C. Performance of Baseline Network

Most of previous methods [25], [26] 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 should be larger than 1.0. Setting height-width ratio as 1.0 might be not an optimal choice for person ReID. We therefore evaluate the effect of input height-width ratio on CUHK03. To make a fair comparison, we ensure each compared input pair contain similar number of pixels. As shown in Fig. 7, setting the ratio as 2.0 gets higher performance than 1.0. The input ratio 2.0 performs better than 1.0 could be because most of persons in existing ReID datasets show standing or walking poses. If lots of persons are sitting or occluded in the dataset, this setting might be not optimal.

D. Performance of Learned Representations

1) Accuracy of Part Generation: One key component of our representation learning is the person part generation. Some examples of automatically detected body parts are illustrated in Fig. 8. As existing person ReID datasets do not provide part annotations, it is hard to quantify the results. To demonstrate that our generated parts are reasonable, we compare the representations learned by CNN trained with part loss using the *generated parts* and *fixed grid parts*, respectively. As shown on the left side of Fig. 9, we generate grid parts by equally dividing an image into horizontal stripes following [1] and [7]. In Fig. 9, the generated parts get substantially higher accuracy

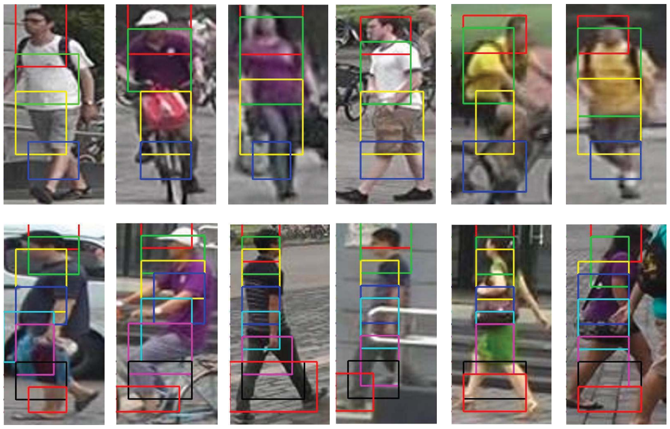


Fig. 8. Samples of generated part bounding boxes. The first and second row correspond to $K = 4$ and $K = 8$, respectively.

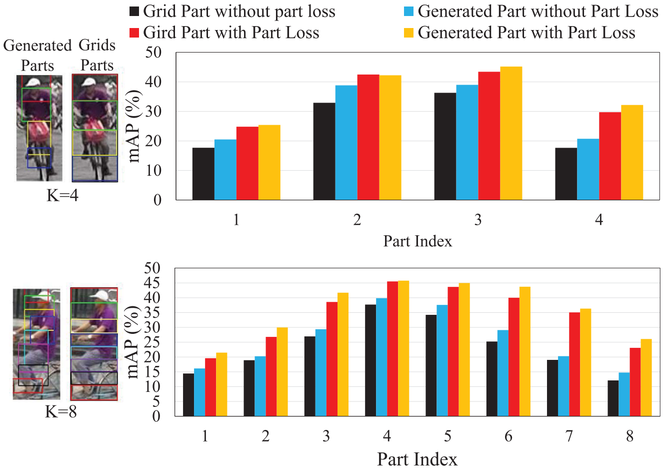


Fig. 9. Performance of grid parts and generated parts representation learned with and without part loss on Market1501.

TABLE I

PERFORMANCE OF FINAL REPRESENTATIONS LEARNED WITH OUR GENERATED PARTS *vs.* FIXED GRID PARTS WITH $K = 8$ ON MARKET1501

Part	mAP(%)	Rank-1 (%)
Grid Part	67.99	86.96
Generated Part	69.3	88.2

than the fixed grid parts for $K = 4$ and 8, respectively. This conclusion is reasonable, because the generated parts cover most of the human body and filter the clustered backgrounds. It also can be observed that, part representations extracted from the center parts of human body, *e.g.*, parts with index = 4 and 5 for $K = 8$, get higher accuracies. This might be because the center of human body generally presents more distinct clothing cues. Table I compares the final global-part representations learned with fixed grid parts and the generated parts, respectively. It is clear that, the generated parts perform substantially better.

2) Validity of Part Loss: This experiment shows that part loss helps to minimize the representation learning risk and improve the descriptive ability of CNN. We firstly show the effects of part loss computed with fixed grid parts. We equally divide an image into stripes, then learn part representations on them with and without part loss, respectively. The evaluation is performed on Market1501. Fig. 9 clearly shows that more discriminative part representations can be learned with part loss for $K = 4$ and 8, respectively.

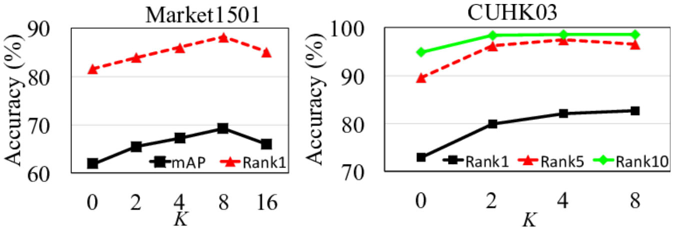


Fig. 10. Performance of final representation on Market1501 and CUHK03 with different K.

TABLE II

PERFORMANCE OF GLOBAL REPRESENTATION ON MARKET1501 WITH DIFFERENT K. $K = 0$ MEANS THE PART LOSS IS NOT CONSIDERED

K	0	2	4	8
mAP(%)	61.9	62.0	64.46	65.91
Rank-1 Acc.(%)	81.5	81.9	84	85.6

Besides using fixed grid part, we further perform experiments to show the validity of part loss computed on generated parts. Comparisons with similar settings are shown in Fig. 9, where part loss also constantly improves the performance. Those two experiments show that, part loss enforces the network to learn more discriminative representations for different body parts, thus better avoids overfitting and decreases the representation learning risk for unseen person images. As shown in Fig. 9, the performance of a single part is not high. However, their concatenation achieves promising performance in Fig. 10.

3) Performance of Global Representation: This experiment verifies the effects of part loss to the global representation. As shown in Fig. 3, the global representation is computed on $\mathcal{X}$, which is also affected by the part loss. Experimental results on Market1501 are shown in Table II, where $K = 0$ means no part is generated, thus part loss is not considered. From Table II, we could observe that part loss also boosts the global representation, *e.g.,* the mAP and Rank-1 accuracy constantly increase with larger K. This phenomenon can be explained by the saliency maps in Fig. 2 (b), which shows the global representation learned with part loss focuses on larger body regions. We thus conclude that, part loss also boosts the global representation to focus on more body parts.

4) Performance of Final Representation: K is the only parameter for part loss. We thus test the performance of the final representation with different K. As shown in Fig. 10, the increasing of K does not constantly improve the performance. For example, setting large $K = 16$ could descrease the performance on datasets Market1501, and setting large $K = 8$ achieves the similar performance to $K = 4$. The reason might be that too large K results in small part regions, which could be sensitive to misalignment errors and human pose variances. In this work, we set $K = 8$ in the following experiments.

5) Discussions on Part Loss: In this work, the part loss is computed with Eq. (10), *i.e.*, compute the ID classification

TABLE III

mAP ACHIEVED BY DIFFERENT WAYS OF PART LOSS COMPUTATION ON MARKET1501. "CONCAT." DENOTES PART LOSS COMPUTED WITH CONCATENATED PART FEATURES. "FINAL," "GLOBAL," "P-k" DENOTES THE FINAL, GLOBAL, AND k-TH PART REPRESENTATIONS. K IS SET AS 4

Methods	Final	Global	P-1	P-2	P-3	P-4
Concat.	64.72	63.36	21.80	38.55	37.78	19.39
Part Loss	67.17	64.46	25.43	42.24	45.19	32.19

TABLE IV

COMPARISON ON MARKET1501 WITH SINGLE QUERY

Methods	mAP(%)	Rank-1 (%)
LOMO+XQDA [1]	22.22	43.79
TMA [59]	22.31	47.92
DNS [11]	35.68	61.02
SSM [5]	68.80	82.21
GoogLeNet [47]	48.24	70.27
VGG16net [60]	38.27	65.02
Res50Net [61]	51.48	73.69
DLCNN(VGG16net) [25]	47.45	70.16
LSTM SCNN [36]	35.31	61.60
Gated SCNN [21]	39.55	65.88
SpindleNet [27]	-	76.9
MSCAN [42]	57.53	80.31
DLPAR [43]	63.4	81.0
P2S [62]	44.27	70.72
CADL [63]	55.58	80.85
PDC [41]	63.41	84.14
Baseline Network	61.9	81.5
Global Representation	65.9	85.6
Part Representation	69	88.0
PL-Net	**69.3**	**88.2**

error on each part separately. Another possible solution is first to concatenate part representations then compute the ID classification with the fused representations. Therefore, we have compared those two methods and summarize the results in Table III. As shown in the comparison, part loss computed with Eq. (10) performs better than the other solution, *e.g.,* 67.17% *vs* 64.72%. This might be because Eq. (10) better ensures the quality of each learned part representation, thus is more effective in decreasing the representation learning risk.

E. Comparison With State-of-the-Art

In this section, we compare the proposed method with existing ones on the Market1501, CUHK03, and VIPeR.

Table IV shows the comparison on Market1501 in the aspects of both mAP and Rank-1 accuracy. As shown in Table IV, the proposed method achieves the mAP of 69.3% and Rank-1 accuracy 88.2%, which both outperform the existing methods. From Table IV we could also see that, our baseline network has achieved competitive performance. This further demonstrates the validity of our convolutional classifier and the selected height-width ratio. By adding the part loss, the global and part representation achieve 4% and 7.1% improvements in mAP upon the baseline network, respectively. This makes the global and part representations already perform better than existing methods. By combining the global and part representations, the final representation further boosts the performance.

On CUHK03, the comparisons with existing methods are summarized in Table V. As shown in Table V, the global and part representations improve the baseline network by 8.1% and 9.85% on Rank-1 accuracy, respectively. The proposed PL-Net achieves 82.75%, 96.59%, and 98.59% for the for Rank-1, Rank-5, and Rank-10 accuracies, respectively. This substantially outperforms most of the compared methods. Note that, the SpindelNet [27] and PDC [41] are learned with extra human landmark annotations, thus leverages more detailed annotations than our method, and DLPAR [43] has a higher baseline performance, *e.g.,* 82.4% [43] vs 72.85% for our baseline.

TABLE V

COMPARISON WITH EXISTING METHODS ON CUHK03

Methods	Rank-1	Rank-5	Rank-10
DeepReID [19]	20.65	51.50	66.5
LSTM SCNN [36]	57.3	80.1	88.3
Gated SCNN [21]	61.8	88.1	92.6
DNS [11]	62.55	90.05	94.80
GOG [12]	67.3	91.0	96.0
DGD [22]	72.58	95.21	97.72
SSM [5]	76.6	94.6	98.0
SpindleNet [27]	88.5	97.8	98.6
MSCAN [42]	74.21	94.33	97.54
DLPAR [43]	85.4	97.6	**99.4**
MuDeep [64]	76.87	96.12	98.41
PDC [41]	**88.70**	**98.61**	99.24
Baseline Network	72.85	89.53	94.82
Global Representation	80.95	95.86	98.16
Local Representation	82.7	96.6	98.59
PL-Net	82.75	96.59	98.6

TABLE VI

COMPARISON WITH EXISTING METHODS ON VIPeR

Methods	Rank-1	Rank-5	Rank-10
DNS [11]	41.01	69.81	81.61
TMA [59]	48.19	87.65	93.54
GOG [12]	49.72	88.67	94.53
Null [11]	51.17	90.51	95.92
SCSP [65]	53.54	**91.49**	**96.65**
SSM [5]	53.73	91.49	96.08
DeepReID [19]	19.9	49.3	64.7
Gated Siamese [21]	37.8	66.9	77.4
LSTM Siamese [36]	42.4	68.7	79.4
SpindleNet [27]	53.8	74.1	83.2
MuDeep [64]	43.03	74.36	85.76
DLPAR [43]	48.7	74.7	85.1
PDC [41]	51.27	74.05	84.18
Baseline Network	34.81	61.71	72.47
Global Representation	44.30	69.30	79.11
Local Representation	44.94	72.47	80.70
PL-Net	47.47	72.47	80.70
PL-Net+LOMO [1]	**56.65**	82.59	89.87

The comparisons on VIPeR are summarized in Table VI. As VIPeR dataset contains fewer training images, it is hard to learn a robust deep representation. Therefore, deep learning-based methods [19], [21], [36], [41], [64] achieve lower performance than metric learning methods [5], [11], [12], [65]. As shown in Table VI, simply using the generated representation obtains the Rank-1 accuracy of 47.47%, which is lower than some metric learning methods [5], [11], [12], [65]. However, it outperforms most of recent deep learning based methods, *e.g.*, DeepReID [19], LSTM Siamese [36], Gated Siamese [21], and MuDeep [64]. Some recent deep learning based methods [27], [41], [43] perform better than ours. Note that, SpindelNet [27] and PDC [41] leverage extra annotations

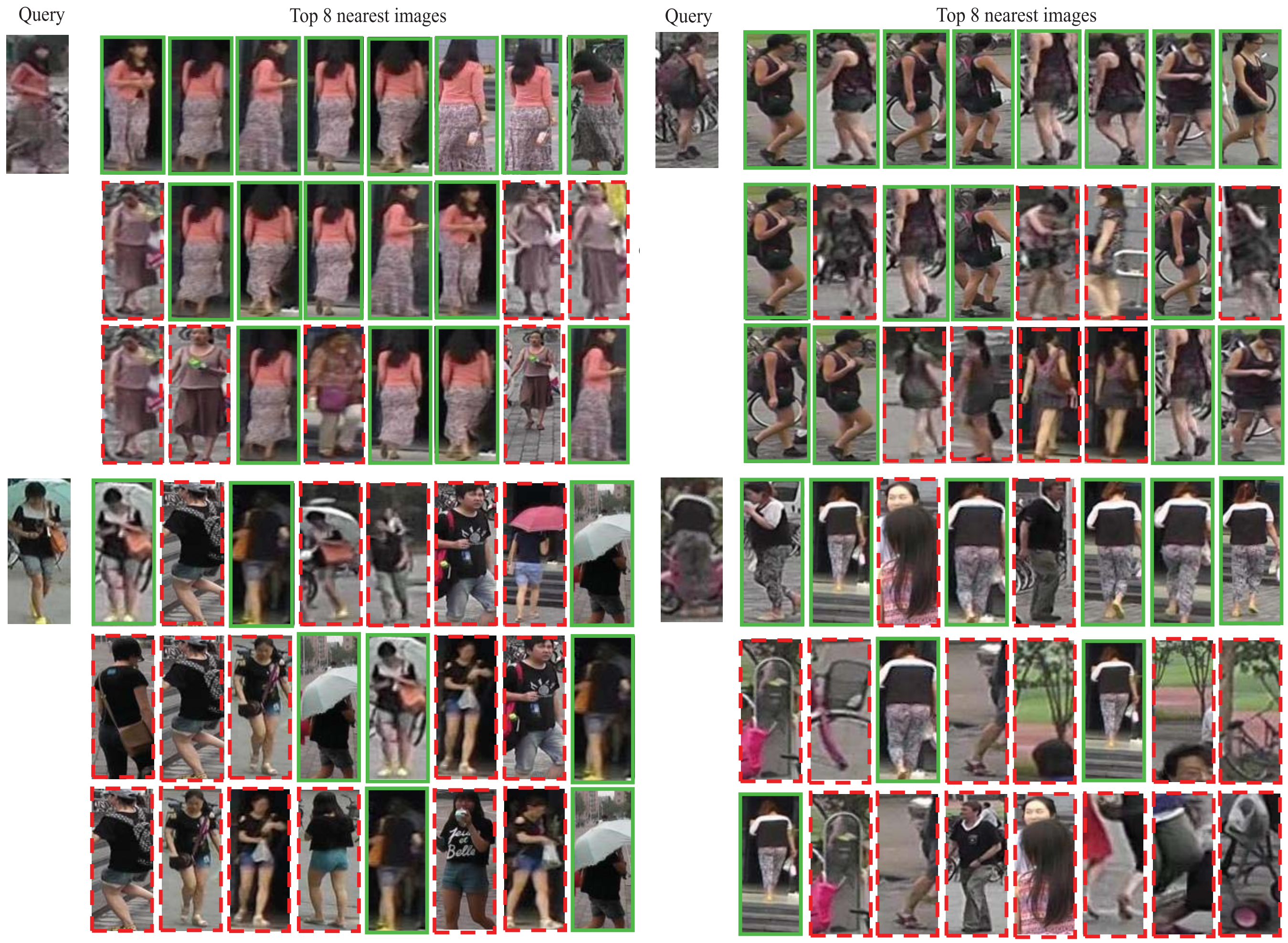


Fig. 11. Ilustration of the retrieval results on Market 1501. The green rectangle represents a true positive, and the red dash rectangle represents a negative positive. For each samples, the first, second, and third rows show the results for the final representation, global representation, and baseline network representations, respectively.

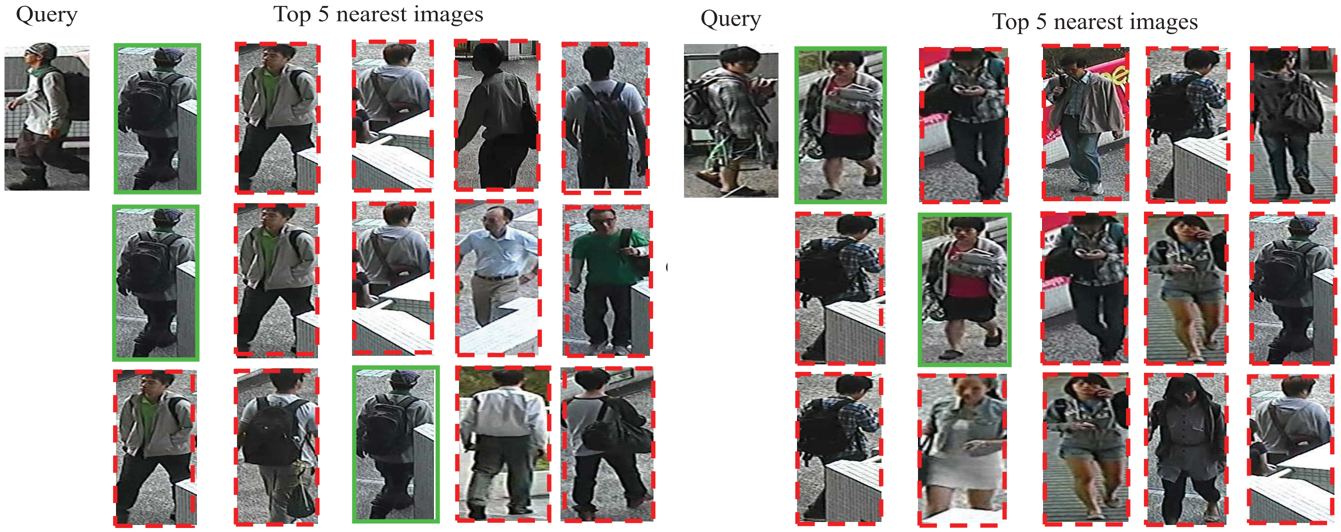


Fig. 12. Ilustration of the retrieval results on CUHK03. The green rectangle represents a true positive, and the red dash rectangle represents a negative positive. For each samples, the first, second, and third rows show the results for the final representation, global representation, and baseline network representations, respectively. Note that each query has one positive match in the gallery for CUHK03.

during training. Also, the training set of DLPAR [43] is larger than ours, *i.e.,* the combination of *CUHK03* and *VIPeR*. Our learned representation is capable of combining with other features to further boost the performance. By combining the traditional LOMO [1] feature, we improve the Rank-1 accuracy to 56.65%.

From the above comparisons, we summarize: 1) part loss computation could improve the discriminative ability for global and part representations, and 2) the combined final representation is learned only with person ID annotations but outperforms most of existing works on the three datasets.

V. Conclusions

This paper shows that, the traditional deep classification models are trained with empirical classification risk on the training set. This makes those deep models not optimal for representation learning in person ReID, which can be regarded as a zero-shot learning problem. We thus propose to minimize the representation learning risk to infer more discriminative representations for unseen person images. The person part loss is computed to evaluate the representation learning risk. Person part loss firstly generates K body parts in an unsupervised way, then optimizes the classification loss for each part separately. In this way, part loss network learns discriminative representations for different parts. Extensive experimental results on three public datasets demonstrate the advantages of our method. Our current body parts are extracted with an unsupervised strategy. Our current body parts are localized with an unsupervised strategy. Supervised part localization is expected to get more accurate part bounding boxes, hence would help to improve our performance. Our future work would also investigate leveraging supervised part localization strategy to further boost the performance of part loss.

References

[1] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.

[2] W.-S. Zheng, S. Gong, and T. Xiang, “Reidentification by relative distance comparison,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

[3] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, “Person re-identification by iterative re-weighted sparse ranking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, Aug. 2015.

[4] A. J. Ma, P. C. Yuen, and J. Li, “Domain transfer support vector ranking for person re-identification without target camera label information,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3567–3574.

[5] S. Bai, X. Bai, and Q. Tian, “Scalable person re-identification on supervised smoothed manifold,” in *Proc. CVPR*, 2017, pp. 3356–3365.

[6] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, “Local Fisher discriminant analysis for pedestrian re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3318–3325.

[7] F. Xiong, M. Gou, O. Camps, and M. Sznaier, “Person re-identification using kernel-based metric learning methods,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 441–448.

[8] C. Liu, C. C. Loy, S. Gong, and G. Wang, “POP: Person re-identification post-rank optimisation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 441–448.

[9] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, “Similarity learning on an explicit polynomial kernel feature map for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1565–1573.

[10] P. Peng *et al.*, “Unsupervised cross-dataset transfer learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1306–1315.

[11] L. Zhang, T. Xiang, and S. Gong, “Learning a discriminative null space for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.

[12] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, “Hierarchical Gaussian descriptor for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1363–1372.

[13] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, “Salient color names for person re-identification,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 536–551.

[14] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2360–2367.

[15] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 1–68.

[16] C. Liu, S. Gong, C. C. Loy, and X. Lin, “Person re-identification: What features are important?” in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 391–401.

[17] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.

[18] 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. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1741–1750.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] R. R. Varior, M. Haloi, and G. Wang, “Gated Siamese convolutional neural network architecture for human re-identification,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 791–808.

[22] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning deep feature representations with domain guided dropout for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1249–1258.

[23] R. Zhao, W. Ouyang, and X. Wang, “Learning mid-level filters for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 144–151.

[24] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3908–3916.

[25] Z. Zheng, L. Zheng, and Y. Yang, “A discriminatively learned CNN embedding for person reidentification,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, p. 13, 2017.

[26] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Deep attributes driven multi-camera person re-identification,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 475–491.

[27] H. Zhao *et al.*, “Spindle Net: Person re-identification with human body region guided feature decomposition and fusion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1077–1085.

[28] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, “A comprehensive study on cross-view gait based human identification with deep CNNs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2016.

[29] E. Ustinova, Y. Ganin, and V. Lempitsky, “Multi-region bilinear convolutional neural networks for person re-identification,” in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug./Sep. 2017, pp. 1–6.

[30] L. Zheng, Y. Huang, H. Lu, and Y. Yang. (2017). “Pose invariant embedding for deep person re-identification.” [Online]. Available: https://arxiv.org/abs/1701.07732

[31] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, “Person re-identification in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3346–3355.

[32] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, “An enhanced deep feature representation for person re-identification,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–8.

[33] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[34] H. Yao *et al.*, “Large-scale person re-identification as retrieval,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1440–1445.

[35] L. Wu, C. Shen, and A. van den Hengel. (2016). “PersonNet: Person re-identification with deep convolutional neural networks.” [Online]. Available: https://arxiv.org/abs/1601.07255

[36] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, “A Siamese long short-term memory architecture for human re-identification,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 135–153.

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 34–39.

[38] H. Shi *et al.*, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 732–748.

[39] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[40] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[41] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3980–3989.

[42] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 384–393.

[43] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3239–3248.

[44] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. ACM Multimedia Conf.*, 2017, pp. 420–428.

[45] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4724–4732.

[46] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[47] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[48] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017.

[49] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[50] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[51] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 842–850.

[52] H. Yao, S. Zhang, C. Yan, Y. Zhang, J. Li, and Q. Tian, "AutoBD: Automated bi-level description for scalable fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 10–23, Jan. 2018.

[53] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, vol. 3, no. 5, 2007, pp. 1–7.

[54] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[55] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[56] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[57] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[58] (2016). [Online]. Available: https://github.com/lim0606/caffe-googlenet-bn

[59] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 858–877.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2014, pp. 1–14.

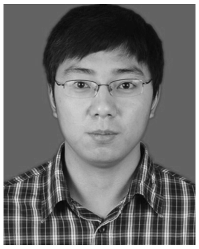
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[62] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5028–5037.

[63] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5771–5780.

[64] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5409–5418.

[65] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1268–1277.



Hantao Yao received the B.S. degree from Xidian University, Xi'an, China, in 2012, and the Ph.D. degree from the Institute of Computing Technology, University of Chinese Academy of Sciences, in 2018. He held a post-doctoral position with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests are computer vision and image retrieval. He was a recipient of the National Postdoctoral Program for Innovative Talents.

Shiliang Zhang received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2012.

He was a Post-Doctoral Scientist with the NEC Labs America and a Post-Doctoral Research Fellow with The University of Texas at San Antonio. He is currently a Tenure-Track Assistant Professor with the School of Electronic Engineering and Computer Science, Peking University. His research interests include large-scale image retrieval and computer vision for autonomous driving. He was a recipient of the National 1000 Youth Talents Plan of China, the NEC Laboratories America Spot Recognition Award, and the Microsoft Research Fellowship. He received the Outstanding Doctoral Dissertation Award from the Chinese Academy of Sciences and the Chinese Computer Federation, the President Scholarship by the Chinese Academy of Sciences, and the Top 10% Paper Award at IEEE MMSP 2011.

Richang Hong (M'12) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He was a Research Fellow with the School of Computing, National University of Singapore, from 2008 to 2010. He is currently a Professor with Hefei University of Technology, Hefei. He has co-authored over 100 publications in the areas of his research interests, which include multimedia content analysis and social media. He is a member of the ACM and the Executive Committee Member of the ACM SIGMM China Chapter. He received the Best Paper Award at the ACM Multimedia 2010 and the Best Paper Award at the ACM ICMR 2015. He was a recipient of the Honorable Mention of the IEEE TRANSACTIONS ON MULTIMEDIA Best Paper Award. He served as the Technical Program Chair for the PCM 2018 and the MMM 2016. He served as an Associate Editor for the *IEEE Multimedia Magazine*, *Information Sciences* (Elsevier), and *Signal Processing* (Elsevier).

Yongdong Zhang (M'08–SM'13) received the Ph.D. degree in electronics engineering from Tianjin University, Tianjin, China, in 2002. He is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored over 100 refereed journal and conference papers. His current research interests are in the fields of multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. He received best paper awards at PCM 2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate at ICME 2011. He serves as an Editorial Board Member for the *Multimedia Systems Journal* and the IEEE TRANSACTIONS ON MULTIMEDIA.

Changsheng Xu (M'97–SM'99–F'14) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and the Executive Director of the China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He has published over 200 refereed research papers in these areas and holds 30 granted/pending patents. He is a fellow of the IEEE and the IAPR, and an ACM Distinguished Scientist. He received the Best Associate Editor Award of the *ACM Transactions on Multimedia Computing, Communications and Applications* in 2012 and the Best Editorial Member Award of the *ACM/Springer Multimedia Systems Journal* in 2008. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the *ACM Transactions on Multimedia Computing, Communications and Applications*, and the *ACM/Springer Multimedia Systems Journal*. He served as the Program Chair for ACM Multimedia 2009. He has served as an associate editor, a guest editor, the general chair, the program chair, an area/track chair, a special session organizer, the session chair, and a TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops.

Qi Tian (F'15) received the B.E. degree in electronic engineering from Tsinghua University in 1992, the M.S. degree in electrical and computer engineering from Drexel University in 1996, and the Ph.D. degree in electrical and computer engineering from UIUC in 2002.

He joined NEC Laboratories of America, as a Visiting Professor, in 2003. In 2007, he joined the MIAS Center, University of Illinois at Urbana–Champaign (UIUC), as a Visiting Scholar. During 2008–2009, he was on one-year faculty leave from the Media Computing Group, Microsoft Research Asia, as a Lead Researcher. He was a Tenured Associate Professor from 2008 to 2012 and a Tenure-Track Assistant Professor from 2002 to 2008. He is currently a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA). He has published over 310 refereed journal and conference papers. He has co-authored the Best Paper in ACM International Conference on Multimedia Retrieval (ICMR) 2015, the Best Paper in PCM 2013, the Best Paper in MMM 2013, the Best Paper in ACM ICIMCS 2012, the Top 10% Paper Award in MMSP 2011, the Best Student Paper in ICASSP 2006, the Best Student Paper Candidate in ICME 2015, and the Best Paper Candidate in PCM 2007. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Dr. Tian has served as a Founding Member for the International Steering Committee for ACM ICMR from 2009 to 2014, an International Steering Committee Member for ACM MIR from 2006 to 2010, and a Best Paper Committee Member for ACM Multimedia 2009, ACM ICIMCS 2013, ICME 2006 and 2009, PCM 2012, and IEEE International Symposium on Multimedia 2011. He has been an ACM Multimedia Conference Review Committee Member since 2009. He will/has served as the General Chair for ACM Multimedia 2015, a Program Coordinator for ACM Multimedia 2009, and the program chair for various international conferences including ACM CIVR 2010, ACM ICMCS 2009, MMM 2010, IMAI 2007, VIP 2007 and 2008, and MIR 2005. He has also served various organization committees as the panel and tutorial chair, the publicity chair, the special session chair, and the track chair in numerous ACM and IEEE conferences such as ACM Multimedia, VCIP, PCM, CIVR, and ICME, and served as a TPC member for prestigious conferences such as ACM Multimedia, SIGIR, ICCV, and KDD. He received 2014 Research Achievement Award from the College of Science, UTSA. He was a recipient of the 2010 ACM Service Award. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Multimedia System Journal*, and on the Editorial Board of the *Journal of Multimedia* and the *Journal of Machine Vision and Applications*. He is the Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the *Journal of Computer Vision and Image Understanding*. He became a member of ACM in 2004.