Deep Representation Learning with Part Loss for Person Re-Identification

Hantao Yao, Shiliang Zhang, *Member, IEEE*, Yongdong Zhang, *Member, IEEE*, Jintao Li, Qi Tian, *Fellow, IEEE*

***Abstract*—Learning discriminative representations for unseen person images is critical for person Re-Identification (ReID). Most of 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 commonly focus on several body parts discriminative to the training set, rather than the entire human body. Inspired by the structural risk minimization principle in SVM, we revise the traditional deep representation learning procedure to minimize both the empirical classification risk and the representation learning risk. The representation learning risk is evaluated by the proposed part loss, which automatically generates several parts for an image, and computes the person classification loss on each part separately. Compared with traditional global classification loss, simultaneously considering multiple part loss enforces the deep network to focus on the entire human body and learn discriminative representations for different parts. Experimental results on three datasets, *i.e.*, Market1501, CUHK03, VIPeR, show that our representation outperforms the existing deep representations.**

***Index Terms*—Person Re-Identification, Representation Learning**

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. 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.

Existing approaches conquer this challenge by either seeking discriminative metrics [1], [2], [3], [4], [5], [6], [7], [8], [2], [9], [10], [11], or generating discriminative features [12], [13], [14], [15], [16], [17], [18], [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], [22], [23], [24], [25], [26], [27], [28], [29], [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.

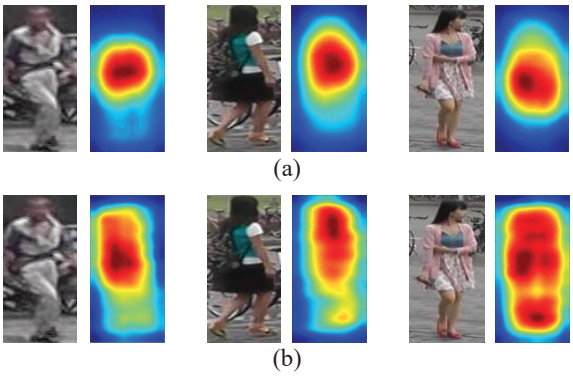

Fig. 1. Saliency maps of CNN learned in traditional classification task (a), and our method (b). The salient region reveals the body part that the CNN representation focuses on. Representations of our method are more discriminative to different parts.

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 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 Figure 1(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 for the training set and ignores the others, *e.g.*, the upper body conveys more distinct clothing cues than the other parts. However, the other parts like head, down-body, and foot, are potential to be meaningful to describe the other unseen persons. Ignoring such parts essentially increases the *risk of representation learning* for unseen data.

The above observations motivate us to study more reason-

Hantao Yao is with Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, and also with the University of the Chinese Academy of Sciences, Beijing 100049, China, Email: yaohantao@ict.ac.cn

Shiliang Zhang is with School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China, Email: slzhang.jdl@pku.edu.cn

Yongdong Zhang, is with Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, and also with the Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048, China, Email: zhyd@ict.ac.cn

Jintao Li is with Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, Email: jtli@ict.ac.cn

Qi Tian is with Department of Computer Science University of Texas at San Antonio, San Antonio, USA, Email: qitian@cs.utsa.edu

able deep representations for person ReID. We are inspired by the structural risk minimization principle in SVM [33], which imposes more strict constraint by maximizing the classification margin. Similarly, we enforce the network to learn better representation with extra representation learning risk minimization. Specifically, the representation learning risk is evaluated by the proposed part loss. The part loss 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 Figure 1(b), minimizing the person part loss guides the deep network to learn more discriminative representations for different body parts.

We propose a network structure that can be optimized accordingly. As shown in Figure 2, this network 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 our approach learns more reasonable representations and gets promising performance in comparison with state-of-the-arts. It also should be noted that, our approach 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. We analyzed the reasons why such representations are not optimal for person ReID. Representation learning risk and part loss are proposed to learn more reasonable deep representations. The proposed method is simple but shows promising performance in comparison with the state-of-the-arts. To our best knowledge, this work is an original effort on improving person ReID by considering the representation learning risk for unseen person images.

II. Related Work

In the past several years, a lot of deep learning-based methods have been proposed for person ReID. Those methods could be summarized into three categories according to their network structures, *i.e.,* classification network, siamese network, and triplet network, respectively. We briefly review these works in the following paragraphs.

Classification network: 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] employ the classification networks that are fine-tuned on target datasets as feature extractor for person ReID. For example, Xiao *etal* [22] propose a novel dropout strategy to train a classification model with multiple datasets jointly. Zheng *etal* [31] extract features with deep classification network to perform person ReID. Wu *etal* [32] combine the hand-crafted histogram features and Convolutional Neural Network (CNN) features to fine-tune the classification network.

Siamese 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 [25], [34], [24], [35], [36], [37] that use siamese network to test whether the two input images contain the same person. Ahmed *etal*[24] employ the siamese network to infer the description and a corresponding similarity metric simultaneously. Shi *etal*[37] replace the Euclidean distance metric with Mahalanobis distance metric in the siamese network. Yi *etal*[36] jointly learn the color feature, texture descriptor, and distance metric in a siamese deep neural network. Zheng *etal* [25] propose another network by jointly considering the objective functions of classification and similarity learning. Varior *etal* [35] combine the LSTM and siamese network architecture for person ReID. Wu *etal* [34] propose a verification network to simultaneously learn high-level features and a corresponding similarity metric for person ReID.

Triplet network: The siamese network is trained with known pair-wise simialrity, 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], [38], [39] employ the triplet networks to learn the discriminative description for person ReID. Cheng *et al.* [38] propose a multi-channel parts-based CNN model for person ReID. Liu *et al.* [39] 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.

By analyzing the difference between image classification and person ReID, we found that the representations learned by existing deep classification models are not optimal for person ReID. Therefore, we consider extra representation learning risk and person part loss for deep representation learning. Our work thus could be regarded as a research effort related to the classification network. However, the proposed part loss could also be leveraged in the other two categories of deep networks to boost their representation learning ability.

III. Methodology

A. Formulation

Given a probe person image I_q, person ReID targets to return the 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.

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 weaker classifiers. Details of this network would 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 $\mathbf{f}_i$ for classifying the seen categories. For example in Figure 1(a), the generated representations focus on discriminative parts that can easily distinguish the training set. However, such representations lack the ability to describe the other parts like head, down-body, and foot which could be meaningful to distinguish a person. These 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 by the representation of k-th part in the i-th training sample. $\hat{y}_i^k$ 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 k-th part.

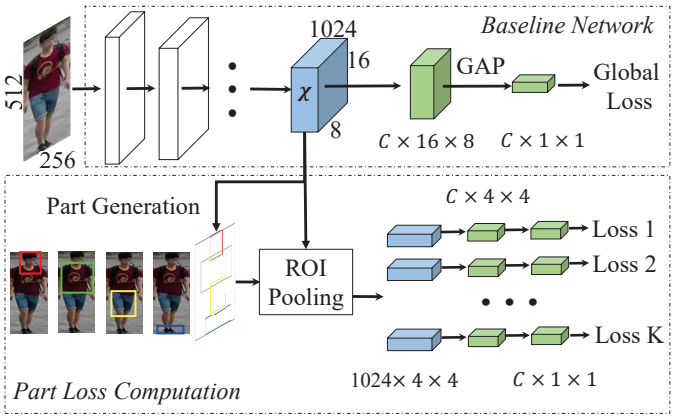

Fig. 2. Overview of our proposed network, which is composed of a baseline network and a part loss computation extension. "GAP" denotes the Global Average Pooling. Given an input image, we firstly extract its feature maps $\mathcal{X}$, then compute the global loss and person part loss based on $\mathcal{X}$. The person part loss is computed on K parts generated with an unsupervised method.

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 CNN model could be inferred by minimizing the empirical classification risk and the representation learning risk simultaneously, *i.e.*,

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

In the following parts, we proceed to introduce our proposed network and the computation of part loss.

B. Proposed Network

Most of the deep learning-based person ReID methods treat the Alexnet [20], GoogleNet [40], and Residual-50 [41] as the baseline network. Given an image, these networks firstly use several convolutional layers to generate the feature representation, then infer fully-connected layers for classification. Therefore, these networks essentially consist of feature extraction and classification modules.

As discussed in Sec. III-A, weaker classifiers should be used to improve the discriminative power of the learned representations. Moreover, the massive parameters in fully-connected layers may make the network prone to overfitting, especially for the relatively small-scale person ReID training sets.

We thus propose a simpler structure as our baseline network. Our baseline network replaces the fully-connected layers with a convolutional layer and a Global Average Pooling (GAP) layer. As shown in Figure 2, 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.

Therefore, our baseline network uses one convolutional layer plus GAP as the classification module, which is more compact than fully connected layers. It is also substantially weaker and thus is potential to learn more discriminative representations. Our baseline network can be easily implemented by replacing the classification modules in available networks. The performance of our baseline network will be evaluated in Sec. IV-C.

According to Eq. (6), our representation is learned to minimize both the empirical classification risk and the representation learning risk. The empirical classification risk is evaluated by the classification loss on the training set. The representation learning risk is evaluated by counting the classification loss on each body part. We thus extend the baseline network accordingly to make it can be optimized by these two types of supervisions.

The overall network is shown in Figure 2. 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.

Specifically, the network processes the input image and generates feature maps. 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 [40]. 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 by counting its classification loss. 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 and extract 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.

Previous work [42] shows 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. Figure 3 shows several feature maps generated by a CNN trained in the classification task. It can be observed that, the lower part of the body has substantially stronger activations. There exist some feature maps responding to the other parts like head and upper body, but their responses

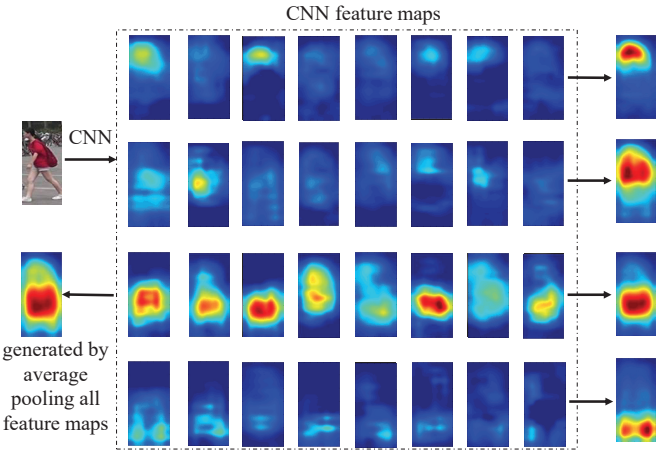

Fig. 3. Examples of CNN feature maps and generated saliency maps. The saliency map generated on all feature maps focuses on one part and suppresses the activations on other parts. The four saliency maps on the right side are generated by average pooling four types of feature maps, respectively. They clearly indicate different part locations.

are substantially weaker. As illustrated in Figure 3, 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 location of their maximum responses, we can gather together feature maps depicting different body parts. Pooling those feature maps indicates the part locations. As shown in Figure 3, the four saliency maps on the right side focus on head, upper body, lower body, and foot, respectively. This might be because the appearances of head, lower body, and foot are still various among different persons, thus CNN still learns certain neurons to depict them in classification.

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 $\mathcal{X}$. We then cluster those locations (h, w) into K groups using L2 distance. As the images in current person ReID datasets are cropped and coarsely aligned, we could simply perform clustering only according to the vertical location h.

After grouping all feature maps into K clusters, we generate one part bounding box from each cluster. Specifically, we average pooling the feature maps in each cluster and apply the max-min normalization to produce the saliency map. A threshold, *e.g.*, 0.5, is set to turn each saliency map into a binary image. In other words, we consider a pixel as the foreground if its value is larger than the threshold. For each binary image, we treat its minimum enclosing rectangle as the

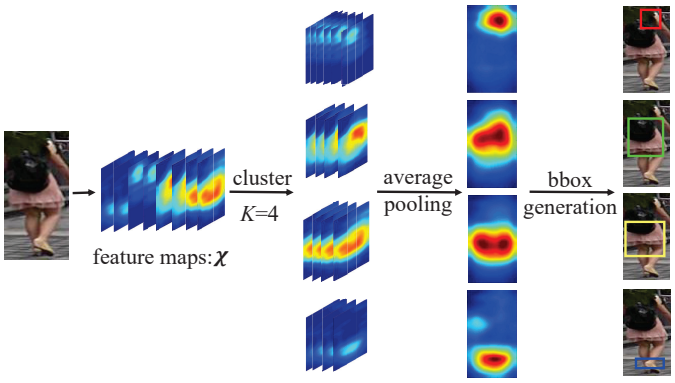

Fig. 4. Illustration of the procedure for unsupervised person part generation.

part bounding box. This procedure is illustrated in Figure 4. Examples of generated parts are shown in Figure 6.

After obtaining the part bounding boxes, we proceed to compute the part loss. Inspired by Fast R-CNN [43], 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.*,

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

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.

The generated parts are updated on each iteration of network training. It should be noted that, the accuracy of our unsupervised part generation is related with the representation learning performance. For example in Figure 3, 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 these parts. During this procedure, the part generation and representation learning can be jointly optimized.

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 method on three datasets: VIPeR [44], CUHK03 [19], and Market1501 [45]. VIPeR [44] 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, the 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) [46]. 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 [45] 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 the authors in [45]. 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).

B. Implementation Details

We use Caffe [47] to implement and train the neural networks. The baseline network is modified from GoogleNet with BatchNormalization [48], and is initialized with the model introduced in [49]. We 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.

C. Performance of Baseline Network

We first evaluate the performance of our baseline network and compare it with CNN networks used in existing methods [22], [25]. We learn representations with our baseline network and the GoogleNet, respectively. Then the learned representations are used to match the probe image against the gallery images to perform person ReID. Experimental results with different network input sizes are summarized in Figure 5(a).

From the Figure 5(a), we could observe that the our baseline network achieves substantially higher accuracy than the GoogleNet for different input sizes. Note that, compared with GoogleNet, our baseline network only has different classifier. This verifies that our weak convolutional classifier

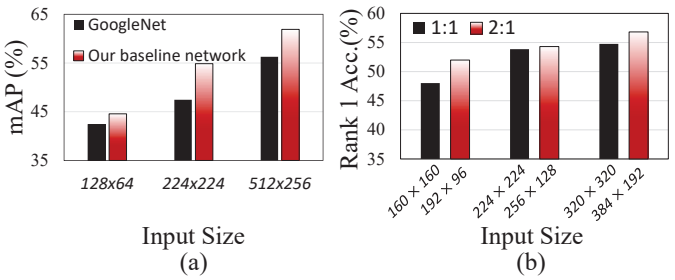

Fig. 5. (a) Performance comparison between representations learned by our baseline network and GoogleNet on Market1501. (b) Effects of the hight-width ration of input image to baseline network performance on CUHK03.

Fig. 6. Samples of generated part bounding boxes. The first and second row correspond to $K = 4$ and $K = 8$, respectively.

helps to learn more discriminative representation than the fully connected layers. Thus our network is more suitable for representation learning in person ReID.

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 Figure 5(b), setting the ratio as 2.0 gets better performance than 1.0. In the following experiments, we thus set the input ratio as 2.0.

D. Performance of Learned Representations

Accuracy of Part Generation: One key component of our representation learning is the person part generation. As existing person ReID datasets do not provide ground truth part annotations, it is hard to quantify the results. Figure 6 illustrates some bounding boxes of the generated part. As shown in Figure 6, the bounding boxes cover important body parts. For the case that K=4, the generated four parts coarsely cover the head, upper body, lower body, and legs, respectively. For the case that K=8, most of generated parts distribute on the human and cover more detailed parts.

To further demonstrate that our generated parts are reasonable, we compare the representations learned on the generated

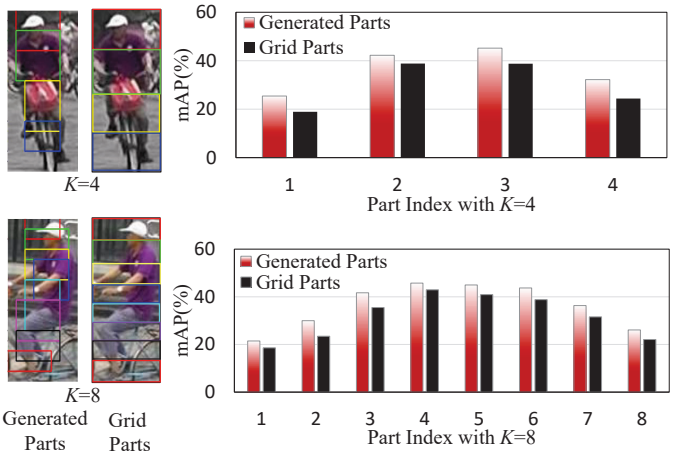

Fig. 7. Performance comparison of part representations learned with our generated parts and fixed grid parts.

parts and fixed grid parts. As shown on the left side of Figure 7, the grid parts are generated by equally dividing the image into horizontal stripes. As shown in Figure 7, the generated parts get substantially higher accuracy 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 of human body, *e.g.*, part index =4 and 5 for K=8 get higher accuracies. This is also reasonable, because center of human body contains more distinct clothing cues.

Validity of Part Loss: This part tries to show that part loss helps to minimize the representation learning risk. Following previous works [1], [7], we equally divide the image into stripes, *i.e.*, fixed grid parts, then learn part representations on them with and without part loss, respectively. We compare the ReID performance of the learned part representations on Market1501. Figure 8 clearly shows that more discriminative part representations can be learned with part loss for K =4 and 8, respectively. This means that, part loss enforces the network to learn more discriminative representations for different body parts, thus decreases the representation learning risk for unseen person images.

Performance of Global Representation: As shown in Figure 2, the global representation is computed on $\mathcal{X}$, which is also affected by the part loss. We thus verify if the global representation could be enhanced by part loss. Experimental results on Market1501 are shown in Table I, where K=0 means no part is generated, thus part loss is not considered. From Table I, we could observe that part loss also boosts the global representation, *e.g.,* the mAP and Rank-1 accuracy constantly increase with larger K. The saliency maps in Figure 1 (b) reflect the focused parts by the global representation. They also demonstrate that part loss boosts the global representation to focus on more parts.

Performance of Final Representation: K is the only one parameter for part loss. We thus test the performance of the final representation with different K. As shown in Figure 9, the final representation performs better with larger K, which

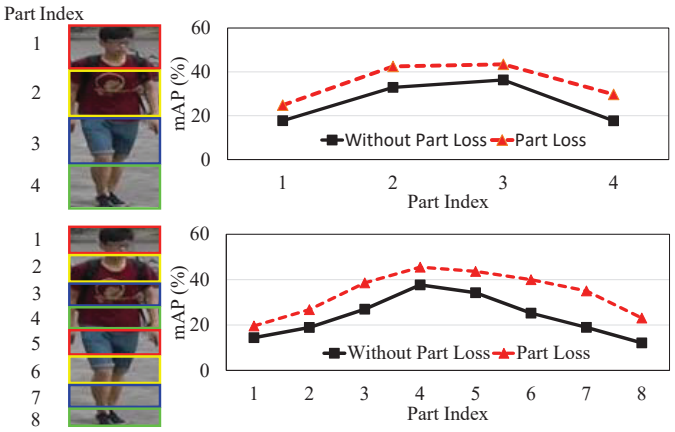

Fig. 8. Performance of part representation learned with and without part loss on Market1501. We use fixed grid parts in this experiment with K=4 and 8, respectively.

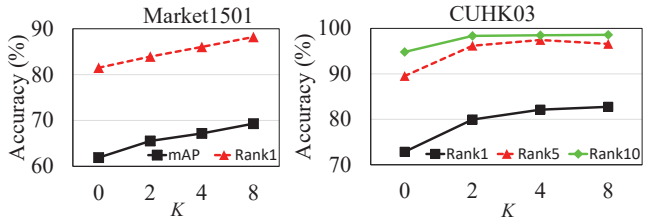

Fig. 9. Performance of final representation on Market1501 and CUHK03 with different K.

extracts more detailed parts. This is consistent with the observation in Table I. This also partially validates our part generation algorithm and part loss. Therefore, we set K=8 in the following experiments.

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 II shows the comparison on Market1501 in the aspects of both mAP and Rank-1 accuracy. As shown in Table II, the proposed method achieves the mAP of 69.3% and Rank-1 accuracy 88.2%, which both outperform the existing methods. From Table II 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.

TABLE I
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

TABLE II
COMPARISON ON MARKET1501 WITH SINGLE QUERY.

Methods	mAP(%)	Rank-1 (%)
LSTM Siamese [35]	35.31	61.60
DNS [11]	35.68	61.02
VGG16net [50]	38.27	65.02
Gated Siamese [21]	39.55	65.88
DLCNN(VGG16net) [25]	47.45	70.16
GoogleNet [40]	48.24	70.27
Res50Net [41]	51.48	73.69
SpindleNet [27]	-	76.9
DLCNN(Res50Net) [25]	59.87	79.51
SSM [5]	68.80	82.21
Baseline Network	61.9	81.5
Global Representation	65.9	85.6
Part Representation	69	88.0
Ours	**69.3**	**88.2**

On CUHK03, the comparisons with existing methods are summarized in Table III. As shown in Table III, our baseline network also obtains a promising performance, *e.g.,* 72.85%, 89.53%, and 94.82% for Rank-1, Rank-5, and Rank-10 accuracies, respectively. Our global and part representations improve the baseline network by 8.1% and 9.85% on Rank-1 accuracy, respectively. The final representation 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] is learned with extra human landmark annotations, thus leverages more detailed annotations than our method.

The comparisons on VIPeR are summarized in Table IV. As shown in Table IV, simply using the generated representation obtains the Rank-1 accuracy of 47.47%. This outperforms most of deep learning based methods, *e.g.*, DeepReID [19], LSTM Siamese [35], and Gated Siamese [21], but is lower than some of existing methods [27], [51], [22]. This is partially because VIPeR contains fewer training images, thus it is hard to learn a robust deep representation. The SpindelNet [27] is also based on deep learning, but it leverages extra annotation during training. Our learned representation is capable to combine with other features to further boost the performance. By combining the traditional LOMO [1] feature, we improve the Rank-1 accuracy to 56.65%, which is higher than existing methods.

From the above comparisons, we could summarize the conclusions as : 1) our baseline network is effective for presentation learning in person ReID, 2) part loss improves the baseline network and results in more discriminative global and part representations, and 3) the combined final representation outperforms most of existing methods 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

TABLE III
COMPARISON WITH EXISTING METHODS ON CUHK03.

Methods	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
DeepReID [19]	20.65	51.50	66.5
LSTM Siamese [35]	57.3	80.1	88.3
Gated Siamese [21]	61.8	88.1	92.6
MetricEmsemb [52]	62.1	89.1	94.3
Null [11]	62.5	90.0	84.8
DNS [11]	62.55	90.05	94.80
DLCNN(Res50net) [25]	66.1	90.1	95.5
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
Baseline Network	72.85	89.53	94.82
Global Representation	80.95	95.86	98.16
Local Representation	82.7	96.6	98.59
Ours	82.75	96.59	**98.6**

TABLE IV
COMPARISON WITH EXISTING METHODS ON VIPER.

Methods	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
DeepReID [19]	19.9	49.3	64.7
Gated Siamese [21]	37.8	66.9	77.4
DNS [11]	41.01	69.81	81.61
LSTM Siamese [35]	42.4	68.7	79.4
TMA [53]	48.19	87.65	93.54
GOG [12]	49.72	88.67	94.53
Null [11]	51.17	90.51	95.92
SCSP [51]	53.54	**91.49**	**96.65**
SSM [5]	53.73	91.49	96.08
SpindleNet [27]	53.8	74.1	83.2
Baseline Network	34.81	61.71	72.47
Global Representation	44.30	69.30	79.11
Local Representation	44.94	72.47	80.70
Ours	47.47	72.47	80.70
Ours+LOMO [1]	**56.65**	82.59	89.87

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, the network learns discriminative representations for different parts. Extensive experimental results on three public datasets demonstrate the advantages of our method. This work explicitly infers parts based on given the parameter K. More optimal implicit ways to minimize the representation learning risk will be explored in our future work.

REFERENCES

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

[2] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *TPAMI*, vol. 35, no. 3, pp. 653–668, 2013.

[3] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *TPAMI*, vol. 37, no. 8, pp. 1629–1642, 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 *ICCV*, 2013.

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

[6] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013.

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

[8] C. Liu, C. Change Loy, S. Gong, and G. Wang, "Pop: Person re-identification post-rank optimisation," in *ICCV*, 2013.

[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 *CVPR*, 2015.

[10] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *CVPR*, 2016.

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

[12] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016.

[13] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014.

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

[15] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *Bmvc*, 2011.

[16] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *ECCV*, 2012.

[17] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[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 *CVPR*, 2015.

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

[20] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

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

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

[23] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014.

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

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

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

[27] H. Zhao, M. Tian, J. Shao, S. Sun, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature," in *CVPR*, 2017.

[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," *TPAMI*, 2016.

[29] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multiregion bilinear convolutional neural networks for person re-identification," *arXiv preprint arXiv:1512.05300*, 2015.

[30] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *arXiv preprint arXiv:1701.07732*, 2017.

[31] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," *arXiv preprint arXiv:1604.02531*, 2016.

[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 *WACV*, 2016.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

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

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

[37] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *ECCV*, 2016.

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

[39] 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.
[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[42] X.-S. Wei, J.-H. Luo, and J. Wu, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *arXiv preprint arXiv:1604.04994*, 2016.
[43] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.
[44] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETSW*, 2007.
[45] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.
[46] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.
[47] 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 *ACM MM*, 2014.
[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *NIPS*, 2015.
[49] "https://github.com/lim0606/caffe-googlenet-bn."
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2014.
[51] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016.
[52] S. Paisitkriangkrai, C. Shen, and A. den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015.
[53] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *ECCV*, 2016.