IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 7, 2018, accepted September 15, 2018, date of publication September 19, 2018, date of current version October 17, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2871149

Learning Irregular Space Transformation for Person Re-Identification

YANWEI ZHENG[1], HAO SHENG[1], (Member, IEEE), YANG LIU[1], KAI LV[1], WEI KE[2], (Member, IEEE), AND ZHANG XIONG[1]

[1]State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[2]School of Public Administration, Macau Polytechnic Institute, Macau 999078, China

Corresponding author: Hao Sheng (shenghao@buaa.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2016QY01W0200, in part by the National Natural Science Foundation of China under Grant 61472019, in part by the Macao Science and Technology Development Fund under Grant 138/2016/A3, in part by the Open Fund of the State Key Laboratory of Software Development Environment under Grant SKLSDE-2017ZX-09, in part by the Project of Experimental Verification of the Basic Commonness and Key Technical Standards of the Industrial Internet Network Architecture, and in part by the Technology Innovation Fund of China Electronic Technology Group Corporation.

ABSTRACT Person re-identification (ReID) classifies the discriminative features of different people. Human perception usually depends on the minority of discriminative colors to classify targets, rather than the majority of mutual colors. ReID uses a small number of fixed cameras, which create a small account of similar backgrounds, leading to the majority of background pixels becoming non-discriminative (this is expanded in the feature map). This paper analyzes the distributions of feature maps to discover their different discriminative power. It also collects statistics that classify feature map values into individual ones and general ones according to the deviation of the mean value on each mini-batch. Finally, our findings introduce a learning irregular space transformation model in convolutional neural networks by enlarging the individual variance while reducing the general one to enhance the discrimination of features. We demonstrate our theories as valid on various public data sets, and achieve competitive results via quantitative evaluation.

INDEX TERMS Irregular space transformation, discriminative power enhancement, convolutional neural networks, person re-identification.

I. INTRODUCTION

Person re-identification (ReID) matches pedestrian images across camera views at different locations and time. ReID underpins many crucial applications in video surveillance, such as long-term cross-camera tracking, video retrieval, etc. It has been proven an academic challenge due to various illuminations, occlusions, viewpoints, background clutters, and image resolutions [1]. Therefore, ReID research becomes popular over the last few years, and a wide variety of features, experimental protocols, and evaluation metrics have been employed. Karanam *et al.* [2] implemented a unified code library that includes 11 feature extraction algorithms and 22 metric learning and ranking techniques.

Recently, many studies have adopted deep learning approaches to solve the ReID problem using three categories: (1) The classification network [3]–[5], which classifies images into person categories with convolutional neural networks (CNNs), and then extracts features to calculate and rank the similarity of images; (2) the siamese network [6]–[9], which is a method that takes two images as input and then generates either a similarity score between the two images or a classification of an image pair, which depicts either the same pedestrian or a group of different pedestrians. Its main focus is how to effectively merge the cross-corresponding pairs into one; (3) the triplet framework [10]–[12], which uses the input of three images – usually a matched image pair and a mismatched image – and outputs features by improving the loss function that minimizes the distance of the matched images while maximizing the mismatched images.

Human perception usually classifies targets by discriminative characters. If a particular color, e.g., red as shown in Fig. 1(a), is different from the others in some images when observing a person's appearance, we can distinguish people using this color – this is called a discriminative character. However, if the majority of people are dressed in red, as shown in Fig. 1(b), we need to choose another color as the discriminative color. Thus, in the

FIGURE 1. Discriminative color in the Market-1501 dataset. (a) Red is a discriminative color. (b) The majority of people are dressed in red. (c) The same viewpoint with a similar background.

end, the high discriminative color becomes the minority color.

Furthermore, there are only a few fixed cameras in person ReID. The fixed camera leads to similar backgrounds with an identical viewpoint, as shown in Fig. 1(c), while the small number of cameras causes a small number of different backgrounds. Thus, most of the background pixels are in the majority count and contain non-discriminative colors.

In this paper, we expand this point of view to the feature map. The discriminative power of the feature map is taken into account, and an irregular space transformation learning method is introduced to increase the discrimination. The main goal is to enhance the competitive power of particular values and increase the discriminative ability of the values that can protrude in the layer. To accomplish this, the feature map values are separated into two categories, where the points around the mean value are general data, which only carry a small amount of information. Their representation space is decreased to provide them with a greater probability of failure in the competition of pooling layer. The remaining values are individual points that represent the kernel feature, and the representation space is expanded to achieve better expression power.

The main contributions of this paper are threefold: (1) The concentration phenomenon and the discriminative power measurement of the feature maps, which enable us to analyze the distribution of the feature map values and distinguish them into individual and general values; (2) an approach that enhances this discrimination by compressing the representation space of the general points and expanding the individual points, which improves the discriminative power of retrieval ranking; and (3) the learning method that determines the border and transformation of individual and general data, which is utilized to achieve better accuracy in the existing models.

II. RELATED WORK

The Siamese network is a popular framework in person ReID that usually focuses on the concatenating of the cross-corresponding pairs when two sub-networks are concatenated into one network. Li *et al.* [6] used a full connection layer to connect two sub-networks. Ahmed *et al.* [8] designed a layer that captures the local relationships between the two input images based on mid-level features from each input image. Zhu *et al.* [13] computed both the element-wise absolute difference and multiplication of the CNN learning feature pair when two sub-networks merge. Wu *et al.* [14] used a layer to calculate neighborhood range differences, and develop an adaptive Root Mean-Square gradient descent algorithm.

A crucial step in classification is extracting discriminative features from the samples, which focuses on the correlation of patch pair matching during the initial stage of person ReID study. There are a few conventional manual distilling methods that are based on this idea [15]–[19]. These approaches are also adopted to build the Siamese networks. Yi *et al.* [7] segmented an image into top, middle and bottom parts, and learned features using a symmetry structure with two sub-networks that are connected by a cosine layer. Huang *et al.* [20] computed the cross-data, cross-map and cross-space differences between paired corresponding parts using different subnets. Zheng *et al.* [21] aligned pedestrians to a standard pose with a PoseBox structure, where the original image and the PoseBox are processed by two weight unshared subnets, and a pose estimation confidence is inputted before the two networks fully connect. Some asymmetrical Siamese architectures have also been developed. Li *et al.* [9] fed low- and high-resolution images to two subnets, whereas Wu *et al.* [22] fused a CNN and some different handcrafted features into one network. Li *et al.* [23] learned and localized deformable pedestrian parts, and this was then used to learn powerful features with the full body.

The triplet framework is another network that takes the correlation of whole images into account and uses the triplet loss function for training. Ding *et al.* [11] fed three images into the network, where two images belonged to one person while the third image did not belong to anyone. Then, a loss function was devised to make the L_2 distance in the feature space between the matched pair smaller than the mismatched pair in each triplet. Cheng *et al.* [10] designed another loss function to train the network models in order to make the distance between the matched pairs less than a predefined threshold and the mismatched pairs in the learned feature space. Wang *et al.* [24] fused different part features and used multiple classifiers to match the pedestrian.

The rank learning methods also adopt triplet networks. Wang *et al.* [12] used two sub-networks for a pair of input images, but two single-image representations and a cross one were calculated, whereas the triplet comparison objectives were combined to improve the matching performance. Liu *et al.* [25] focused on parts of person image pairs after reviewing them and adaptively comparing their appearance in triplet networks.

Some other architectures are regarded as a generalization of triplet networks. Wang *et al.* [26] replaced the image

pairs with ranking lists as training samples, and developed a listwise loss with adaptive margin. Chen *et al.* [27] designed a quadruplet loss, which led to model output with a larger inter-class variation and a smaller intra-class variation compared to the triplet loss. Lin *et al.* [28] exploited consistent-aware information under a deep learning framework to obtain the maximal correct matches for the whole camera network. Zhou *et al.* [29] used the point to set (P2S) metric to replace the point to point (P2P) distances, which jointly minimized the intra-class distance and maximized the inter-class distance.

Other researchers regarded the ReID problem as a classification problem, where a single network is applied to classify each image into a person category in training. In this case, the CNN features are extracted from the network in testing to calculate the similarity of the image pairs. Xiao *et al.* [3] developed a domain guided dropout algorithm based on the observation of CNN's training data from cross domains. It concluded that some neurons learned representations shared across several domains, while some others were effective for a specific domain. In addition to those facts, an inception layer, which is the nets-within-nets architecture of GoogLeNet [30] was adopted in that research. Chen *et al.* [4] combined two images horizontally to form an image that was used as input, and proposed a learning-to-rank algorithm to minimize the cost corresponding to the ranking disorders of the gallery. Franco and Oliveira [5] extracted features from intermediate and top layers. The former was wrapped in covariance matrices and integrated into the top layer features. Su *et al.* [31] proposed a semi-supervised framework to learn attributes that obtained a superior generalization ability across different datasets.

All of the aforementioned researches concentrated on the framework improvement, whereas other data distribution studies also produced important promotion for CNN. Up until now, dropout [32], [33] is a regularization scheme for the purpose of avoiding over-fitting in neural networks by preventing complex co-adaptations on training data. This creates major improvements over other regularization methods. Batch Normalization [34] accelerates training by reducing the internal covariate shift, where higher learning rates can be used.

The mid-level feature representation has also been a research focus in person ReID. Lin *et al.* [35] learned a correspondence structure to capture the patch-wise spatial pattern, and proposed a global matching constraint to exclude cross-view misalignments. Liu *et al.* [36] proposed a method to learn attentive deep features from an attention-based deep neural network to capture multiple attentions from low-level to semantic-level. Zhao *et al.* [37] proposed a scheme that learned different semantic features from different body regions, which were merged with a competitive scheme.

In this paper, we focus on the distribution and transformation of training data. Usually, non-linear transformation of outputs of convolutional layer is achieved by the activation layer, including Sigmoid, ReLU [38], PReLU [39], etc. Agostinelli *et al.* [40] designed a form of adaptive piecewise linear (APL) activation function which is learned independently by each neuron. The activation function is used to remove the redundancy but retains the features of data, which makes the classification easier. However, the activation function is based on a neuron, which ignores the relation of different neurons.

Our approach takes into account the overall distribution of training data and looks for the dispersion phenomenon of values. Based on the fact that different data has different discriminative power, the training data is divided into individual and general categories. An irregular space transformation model is proposed to improve the discrimination of the training data, which is based on the overall statistics of a feature map. After the transformation, the discrimination is increased and transmitted to the final Softmax classifier, layer by layer. Experimental results show that this application considerably increases the accuracy of identification.

III. MOTIVATION

As discussed in Section I, the high discriminative color is the minority color. We generalize this perspective to the feature map, i.e., the minority of the feature map has more discriminative information. Since the closer the two data are, then the smaller the differentiation is, we define the discriminative power as the absolute difference of two data, as shown in Eq. 1, where $z_k^{(l)}$ is the value of a feature map element $Z_k^{(l)}$, l is the layer index, and i, j denote the i^{th} and j^{th} component of data in layer l, respectively.

$$\triangle z_{ij}^{(l)} = |z_i^{(l)} - z_j^{(l)}|. \tag{1}$$

Multiplying a constant for each $z_i^{(l)}$ can achieve discriminative power enhancement, but the convolutional operation may pull the data back to its original value. Let $z_i^{(l+1)} = \lambda z_i^{(l)}$ where $\lambda > 1$, then $\triangle z_{ij}^{(l+1)} = \lambda \triangle z_{ij}^{(l)} > \triangle z_{ij}^{(l)}$. Suppose there is a convolutional operator of $z_i^{(l+2)} = f\left(\sum_{k=1}^{n} w_k^{(l+1)} z_k^{(l+1)} + b\right)$ that is applied on each kernels, where $w_k^{(l+1)}$, b and n are weight, bias and pixel number of kernel of layer $(l+1)$, respectively. The weights only need to be reduced to $\frac{1}{\lambda}$ of their original value, then the output is the same, i.e., $z_i^{(l+2)} = f\left(\sum_{k=1}^{n} \frac{1}{\lambda} w_k^{(l+1)} z_k^{(l+1)} + b\right) = f\left(\sum_{k=1}^{n} w_k^{(l)} z_k^{(l)} + b\right)$.

When we review the feature map of CNNs, we discover that most of the data are near the mean, as shown in Fig. 2(a). The training data are collected from a person ReID dataset named Market-1501 [41], where the vertical axis denotes the frequency of data, and μ and σ are the mean and variance, respectively. Fig. 2(b) shows an intuitive distribution of 1000 randomly selected data, where the horizontal axis denotes the randomly generated coordinate that is only used to avoid the overlap of points. It is apparent that the training data are concentrated in the interval of $[\mu - c\sigma, \mu + c\sigma]$, where c is a positive parameter, and $c = 1.2$ in Fig. 2. The data around the mean value only carry a small amount of information, which are called general data. The remaining

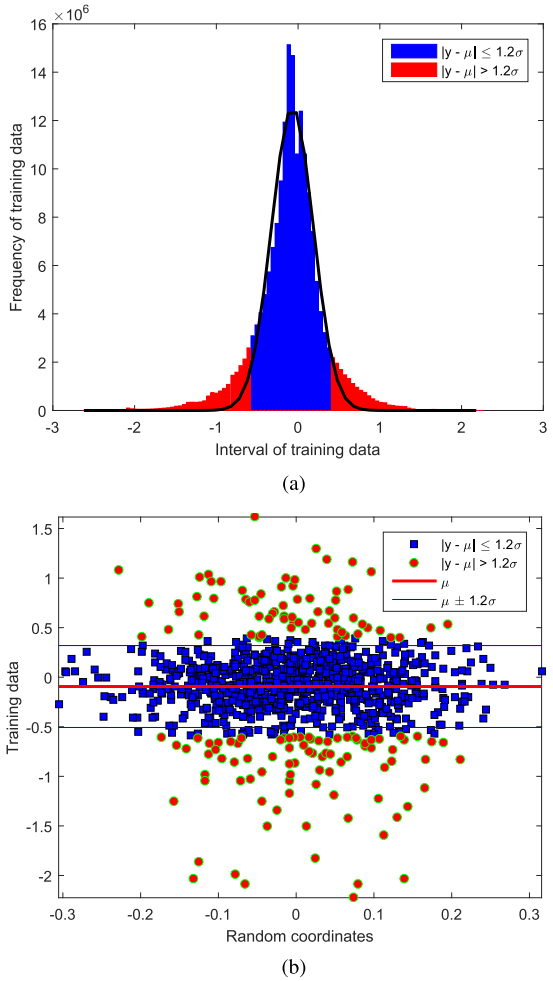


FIGURE 2. Distribution of training data on Market-1501 dataset. (a) Frequency histogram of 185,856,000 training data. (b) 1,000 randomly selected training data.

data, which have more discriminative information, are called individual data.

In order to directly increase the discrimination in training, our idea is to shrink the space of general data while expand the individual one's, i.e., apply a piecewise linear transformation on the training data, which is shown in Fig. 3. First, the data are split to three intervals $(-\infty, \mu - c\sigma)$, $[\mu - c\sigma, \mu + c\sigma]$ and $[\mu + c\sigma, +\infty]$ by a parameter $c(c > 0)$, mean μ and variance σ. The values in $[\mu - c\sigma, \mu + c\sigma]$ are general data, whereas the others are individual ones. Second, apply different linear transformations on each interval, which is shown in Eq. 2. We discuss how to determine the parameter in Section IV-C.

$$z_j^{(l+1)} = k_j z_j^{(l)} + b_j$$
$$s.t.\ k_1 > 1, \quad 0 < k_2 < 1,\ k_3 > 1,\ j = 1, 2, 3. \quad (2)$$

IV. LEARNING IRREGULAR SPACE TRANSFORMATION

A. FRAMEWORK

In order to protrude the individual data, we need to use the mean and the variance to separate them from the others.

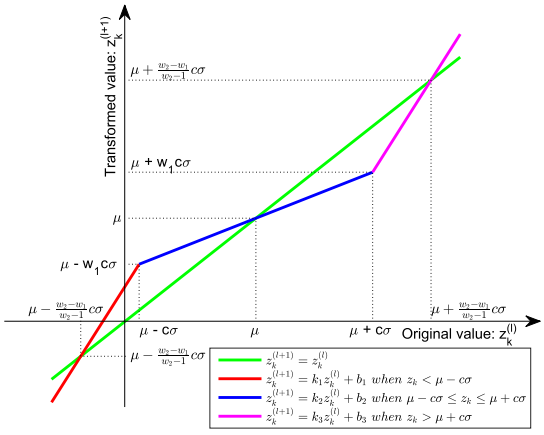


FIGURE 3. Irregular space transformation.

This means that we need a statistics collection procedure for each mini-batch. Afterwards a transformation is introduced to make the individual data more important than the general ones. Fig. 4 shows the framework of the Learning Irregular Space Transformation (LIST) model.

The forward propagation of training includes three stages:

1) Collecting statistics to obtain the local mean and variance of a mini-batch, which is the foundation of the training data classification. The statistics are accumulated and stored as global ones for the testing procedure. The processing details are introduced in sub-section IV-B.
2) Classifying data into individual and general ones is based on the crowding position (mean value) of data and the drifted distance from the center (variance). The training procedure alternatively uses local or global statistics for general and individual data classification, whereas the testing procedure uses the global ones.
3) An irregular space transformation is applied to different categories of data (sub-section IV-C).

The operation of back propagation is similar to other types of layers. We need to calculate the error terms and the partial derivatives of the cost function, which are transmitted from the next layer to the previous layer. This procedure is formulated and discussed in sub-section IV-D.

B. DATA CLASSIFICATION

The method to enhance discrimination is to increase the intra-class distance while downgrade the inner-class one. We need to distinguish which class a training value belongs to. As mentioned in section I, we consider the data that are concentrated around the mean value to be the general data, which have little information to classify pedestrian. The rest are the individual ones, which express the discriminative characteristics of a person.

Suppose there are m values $\{z_1^{(l)}, z_2^{(l)}, \ldots, z_m^{(l)}\}$ in a mini-batch of layer l, we collect statistics for the mean value and variance as shown in Eq. 3. The general data contain values in interval of $[\mu - c\sigma, \mu + c\sigma]$, whereas the individual

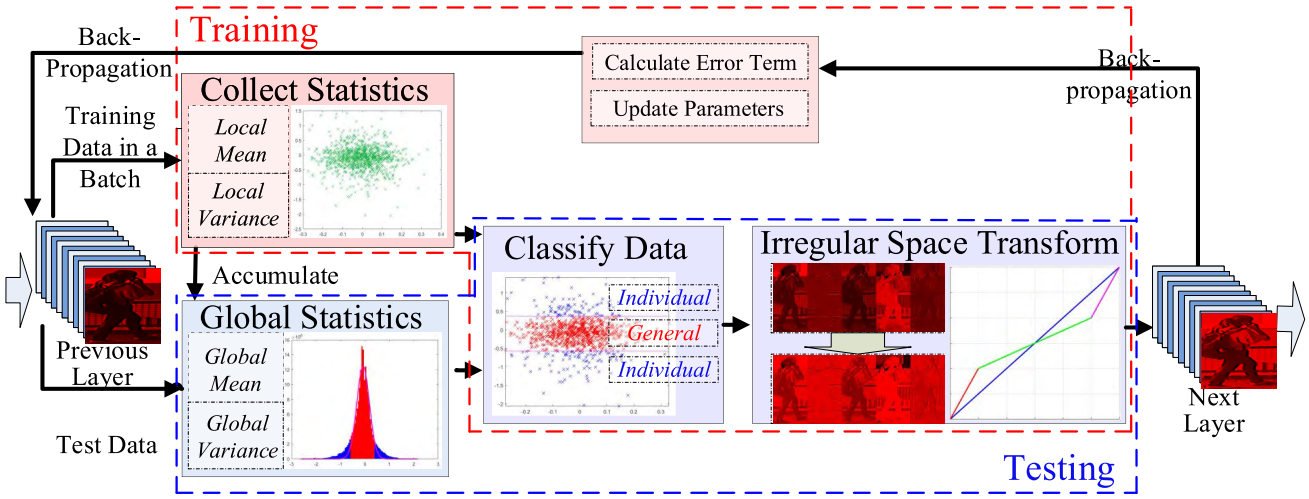


FIGURE 4. Framework of LIST model.

ones are in $(-\infty, \mu - c\sigma) \cup (\mu + c\sigma, +\infty)$.

$$\mu_{(n)} = \frac{1}{m}\sum_{k=1}^{m} z_k^{(l)}, \quad \sigma_{(n)}^2 = \frac{1}{m-1}\sum_{k=1}^{m}\left(z_k^{(l)} - \mu\right)^2, \quad (3)$$

where $\mu_{(n)}$ and $\sigma_{(n)}^2$ are the local mean value and variance of the n^{th} mini-batch, respectively.

Once the mean and variance are calculated for each mini-batch, they are iteratively accumulated to the global mean and variance, respectively. In the testing procedure, the global statistics are used to replace the mini-batch computing. In the training procedure, the local or global statistics are alternatively used for computing. Suppose the global mean value from the 1^{st} to the n^{th} mini-batch is $\mu_n = \frac{1}{mn}\sum_{k=1}^{mn} z_k^{(l)}$, then the global mean value from the 1^{st} to the $(n+1)^{th}$ mini-batch is calculated by

$$\begin{aligned}\mu_{n+1} &= \frac{1}{m(n+1)}\sum_{k=1}^{m(n+1)} z_k^{(l)} \\ &= \frac{n}{n+1}\mu_n + \frac{1}{n+1}\mu_{(n+1)}, \end{aligned} \quad (4)$$

where $\mu_{(n+1)}$ is the local mean value of the $(n+1)^{th}$ mini-batch.

Similarly, the global variance is computed by

$$\begin{aligned}\sigma_{n+1}^2 &= \frac{1}{m(n+1)-1}\sum_{k=1}^{m(n+1)}(z_k^{(l)} - \mu_{n+1})^2 \\ &= \frac{(mn-1)\sigma_n^2 + (m-1)\sigma_{(n+1)}^2}{m(n+1)-1} \\ &\quad + \frac{mn(\mu_n - \mu_{(n+1)})^2}{[m(n+1)-1](n+1)}, \end{aligned} \quad (5)$$

where σ_{n+1}^2 and $\sigma_{(n+1)}^2$ are the global and the $(n+1)^{th}$ mini-batch local variance, respectively.

According to Eq. 4 and Eq. 5, the global mean and variance of the $(n+1)^{th}$ iteration are only dependent on the n^{th} iterative global and local mini-batch ones. Thus, the global statistics are computed during training, and no additional statistics collection procedure is required in testing.

C. IRREGULAR SPACE TRANSFORMATION

We use the piecewise linear transformation to achieve the shrinking and increasing operation for general and individual training data, as shown in Fig. 3. The transformation is subject to the following regulations, where w_1, w_2 and c are learning parameters, L_1, L_2 and L_3 are three lines in Eq. 2 when $j = 1, 2, 3$, respectively.

- The general interval $[\mu - c\sigma, \mu + c\sigma]$ is shrunk to $[\mu - w_1 c\sigma, \mu + w_1 c\sigma]$, where $0 < w_1 < 1$.
- L_1 and L_3 keep symmetry, i.e., $k_1 = k_3$.
- The individual intervals $(-\infty, \mu - c\sigma)$ and $(\mu + c\sigma, +\infty)$ are increased, i.e., $k_1 = k_3 = w_2 > 1$.
- L_1 and L_2 intersect at point $(\mu - c\sigma, \mu - w_1 c\sigma)$.
- L_2 and L_3 intersect at point $(\mu + c\sigma, \mu + w_1 c\sigma)$.

According to these constrains and Eq. 2, we can obtain the piecewise linear function, as shown in Eq. 6.

$$z_k^{(l+1)} = \begin{cases} w_2[z_k^{(l)} - (\mu - c\sigma)] + (\mu - w_1 c\sigma), & z_k^{(l)} < \mu - c\sigma \\ w_1[z_k^{(l)} - \mu] + \mu, & \mu - c\sigma \le z_k^{(l)} \le \mu + c\sigma \\ w_2[z_k^{(l)} - (\mu + c\sigma)] + (\mu + w_1 c\sigma), & z_k^{(l)} > \mu + c\sigma \end{cases}$$
$$s.t.\ c > 0, \quad 0 < w_1 < 1, \ w_2 > 1. \quad (6)$$

It is important to note that the shrunk interval is not $[\mu - c\sigma, \mu + c\sigma]$, but $[\mu - \frac{w_2 - w_1}{w_2 - 1}c\sigma, \mu + \frac{w_2 - w_1}{w_2 - 1}c\sigma]$, as shown in Fig. 3. Because c, w_1 and w_2 are learned parameters, the real interval size can be adjusted to a suitable one automatically.

Another thing to note is that the parameters of c, w_1 and w_2 are different in each channel of each LIST layer. Because the LIST layer is placed after every Batch-Normal layer, if there are n_L LIST layers in a network, and n_i channels in the i^{th} layer, the total number of groups of parameters c, w_1 and w_2 will be $\sum_{i=1}^{n_L} n_i$.

D. BACKPROPAGATION

In the error term calculation, the variance appears in the denominator. To avoid dividing a number by zero, σ is replaced by $\sqrt{\sigma^2+\varepsilon}$, where ε is a small positive number. After this substitution, the partial derivative of the statistics are computed as Eq. (7) and Eq. (8).

$$\frac{\partial \mu}{\partial z_k^{(l)}} = \frac{1}{m}, \quad \frac{\partial \sigma^2}{\partial z_k^{(l)}} = \frac{2}{m-1}(z_k^{(l)} - \mu). \tag{7}$$

$$\frac{\partial \sqrt{\sigma^2+\varepsilon}}{\partial z_k^{(l)}} = \frac{1}{2\sqrt{\sigma^2+\varepsilon}} \cdot \frac{\partial \sigma^2}{\partial z_k^{(l)}} = \frac{z_k^{(l)} - \mu}{(m-1)\sqrt{\sigma^2+\varepsilon}}. \tag{8}$$

Suppose the loss function and the error term of the $(l+1)^{th}$ layer are $J(\cdot)$ and $\delta_k^{(l+1)}$, respectively, we can obtain the error term of the l^{th} layer according to the chain rule of derivative as

$$\begin{aligned}\delta_k^{(l)} &= \frac{\partial J(\cdot)}{\partial z_k^{(l)}} = \frac{\partial J(\cdot)}{\partial z_k^{(l+1)}} \cdot \frac{\partial z_k^{(l+1)}}{\partial z_k^{(l)}} \\ &= \delta_k^{(l+1)} \begin{cases} (w_2 - w_1)c\widetilde{z}_k^{(l)} + \widetilde{w}_2, & z_k^{(l)} < \mu - c\sigma_\varepsilon \\ \widetilde{w}_1, & \mu - c\sigma_\varepsilon \le z_k^{(l)} \le \mu + c\sigma_\varepsilon \\ (w_1 - w_2)c\widetilde{z}_k^{(l)} + \widetilde{w}_2, & z_k^{(l)} > \mu + c\sigma_\varepsilon, \end{cases}\end{aligned} \tag{9}$$

where $\widetilde{z}_k^{(l)} = \frac{z_k^{(l)} - \mu}{(m-1)\sigma_\varepsilon}$, $\sigma_\varepsilon = \sqrt{\sigma^2+\varepsilon}$, and $\widetilde{w}_i = w_i\left(1 - \frac{1}{m}\right) + \frac{1}{m}$.

Let $I\{\cdot\}$ be the indicator function, which is shown in Eq. 10, the partial derivatives of the cost function can be calculated by Eq. 11, 12 and 13.

$$I\{x\} = \begin{cases} 0, & x = false \\ 1, & x = true. \end{cases} \tag{10}$$

$$\begin{aligned}\frac{\partial J(\cdot)}{\partial c} &= \sum_{k=1}^{m} \frac{\partial J(\cdot)}{\partial z_k^{(l+1)}} \cdot \frac{\partial z_k^{(l+1)}}{\partial c} \\ &= (w_2 - w_1)\sigma_\varepsilon \sum_{k=1}^{m} \delta_k^{(l+1)} I\{z_k^{(l)} < \mu - c\sigma_\varepsilon\} \\ &\quad - (w_2 - w_1)\sigma_\varepsilon \sum_{k=1}^{m} \delta_k^{(l+1)} I\{z_k^{(l)} > \mu + c\sigma_\varepsilon\}.\end{aligned} \tag{11}$$

$$\begin{aligned}\frac{\partial J(\cdot)}{\partial w_1} &= \sum_{k=1}^{m} \frac{\partial J(\cdot)}{\partial z_k^{(l+1)}} \cdot \frac{\partial z_k^{(l+1)}}{\partial w_1} \\ &= \sum_{k=1}^{m} \delta_k^{(l+1)} I\{\mu - c\sigma_\varepsilon \le z_k^{(l)} \le \mu + c\sigma_\varepsilon\}[z_k^{(l)} - \mu] \\ &\quad - c\,\sigma_\varepsilon \sum_{k=1}^{m} \delta_k^{(l+1)} I\{z_k^{(l)} < \mu - c\sigma_\varepsilon\} \\ &\quad + c\sigma_\varepsilon \sum_{k=1}^{m} \delta_k^{(l+1)} I\{z_k^{(l)} > \mu + c\sigma_\varepsilon\}.\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial J(\cdot)}{\partial w_2} &= \sum_{k=1}^{m} \frac{\partial J(\cdot)}{\partial z_k^{(l+1)}} \cdot \frac{\partial z_k^{(l+1)}}{\partial w_2} \\ &= -\sum_{k=1}^{m} \delta_k^{(l+1)} I\{z_k^{(l)} < \mu - c\sigma_\varepsilon\}[z_k^{(l)} - (\mu - c\sigma_\varepsilon)] \\ &\quad + \sum_{k=1}^{m} \delta_k^{(l+1)} I\{z_k^{(l)} > \mu + c\sigma_\varepsilon\}[z_k^{(l)} - (\mu + c\sigma_\varepsilon)].\end{aligned} \tag{13}$$

V. EXPERIMENTS

A. DATASETS

In this section, we evaluate the proposed method on six different datasets. Table 1 shows the camera, identity (ID) and image numbers.

CUHK01 [42], [43] includes 3,884 images of 971 pedestrians captured by two disjoint cameras, with each person having two images for each camera. Large inter-camera variations in this dataset make the person ReID experiment challenging.

CUHK03 [6] is one of the largest person ReID datasets, which has 1,467 IDs from five different pairs of cameras on campus. The detected and manually labeled bounding boxes are all used for training in our experiments, and have an average of 4.8 detected and manually labeled bounding boxes in each view.

Market-1501 [41] is another largest dataset, which contains 32,688 bounding boxes of 1,501 identities produced by Deformable Part Model (DPM). Each person is captured by $2 \sim 6$ non-overlap cameras. This dataset contains 2,798 distractors (produced by DPM false detection) and 3,819 junk images (has zero influence to the ReID accuracy) in the test set.

PRID2011 [44] records 385 persons from one view and 749 from the other one. The first 200 persons appear in both camera views. We use the multi-shot images for training and the single-shot ones for testing.

i-LIDS [45] person ReID dataset comes from the 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) dataset, which includes 476 images of 119 individuals using surveillance cameras in an airport.

VIPeR [46] contains 1,264 images of 632 people from two cameras. This dataset is very challenging due to the low image quality, coupled with large variations in illumination, poses and viewpoints.

Table 1 also shows the settings for the training, validation and testing groups. In the CUHK01, i-LIDS and VIPeR datasets, 485, 60 and 100 IDs are randomly chosen as the testing sets and the remainders are the training sets. In the CUHK03 dataset, the 20th pre-split group is used as the testing set. In the PRID2011 dataset, the first 100 IDs are used as the testing set. The grouping of Market-1501 is followed by the publisher's groups.

B. SETTINGS AND EVALUATION PROTOCOLS

We employ the popular networks, GoogLeNet-v3(Inception-v3) [47] and ResNet50 [48], as the baselines. We first classify identities using the networks with the softmax with loss

TABLE 1. Dataset grouping for training, validation and testing.

Dataset	#cameras	#identities				#images		
		total	train & val	test probe	test gallery	train & val	test probe	test gallery
CUHK01	2	971	486	485	485	2,044	970	970
CUHK03	10	1,467	1,367	100	100	26,253	952	988
Market-1501	6	1,501+2	751	750	750+2	12,936	3,368	19,732
PRID2011	2	749	649	100	649	75,417	100	649
i-LIDS	2	119	59	60	60	238	107	131
VIPeR	2	632	532	100	100	1,064	100	100
Total	–	5,441	3,844	1,595	2,146	43,469	5,597	22,670

layer, and then extract features to compute the distance to rank the candidates, but there are some differences from the original frameworks. First, we add a fully connected layer (with 384 outputs) before the classifier layer (with 3,844 outputs), which is used to extract features for the classification network. Second, we use the Cosine Distance to compute the rank list after the features are extracted. Third, our LIST layer is placed after every batch normalization layer of Inception-v3 and ResNet50, which are called Inc-LIST and Res-LIST for short, respectively.

The implementation framework of our LIST model is based on CAFFE [49]. We train the networks 80,000 iterations by jointly crossing the six datasets, and then fine-tune 40,000 iterations on each dataset. Because there are totally 3,844 identities in the six training and validation sets, we classify the images into 3,844 categories in the jointly training. In the fine-tuning process, the number of categories is the same as in the jointly training, although some identity values are never used in a single dataset. For example, there are 751 identities in Market-1501, then there are $3,844 - 751 = 3,093$ identity values that are never taken in the fine-tuning on this dataset. In testing, although no testing identity appears in the 3,844 identities, we still use the same number of categories to classify persons and extract features from the fully connected layer, and then use the Cosine Distance to rank the gallery identities.

In the experiments, we employ the commonly used Rank-1 accuracy, Cumulative Match Characteristic (CMC) curve, and the mean average precision (mAP) to evaluate the methods. Rank-1 accuracy refers to the traditional notion of classification accuracy, which is the percentage of probe images that are perfectly matched to their corresponding gallery image. The CMC curve summarizes the chance of the correct match appearing in the top $1, 2, \ldots, n$ of the ranked list. The first point of the CMC curve is the Rank-1 accuracy.

If multiple gallery ground truths exist, the CMC curve is biased because "recall" is not considered [41]. In this case, the Precision-Recall curve for each query is calculated, which is known as average precision (AP). Then, the mean value of the APs of all queries, i.e., the mAP, is calculated, which considers both precision and recall of an algorithm, thus providing a more comprehensive evaluation.

In the experimental datasets, each ID has multiple instances, except for VIPeR. Following the popular used evaluation protocols, we only apply one query image for search. There are two methods [50] to calculate the CMC curve:

- single-shot versus single-shot (SvsS), if each image in a set represents a different individual;
- single-shot versus multiple-shot (SvsM), if each ID has several images in gallery.

In the case of the former, we randomly choose an image for each ID to calculate the CMC curve, and repeat it 100 times to compute the mean as the final result. For the later, only the first match is counted regardless of how many ground truth matches are in the gallery – this is usually called Single Query (SQ) [41], [51]–[54] (from the query viewpoint, we do not know that the two images belong to one person) or Multi-shot [52] (from the candidate viewpoint, all gallery images are used) for short.

There are also two methods to calculate the AP. Suppose there are a total of n images and m matched images to a query image in the gallery, and the ranks of the matched images are $s_1, s_2, \ldots, s_m$. Let $P(k)$ be the precision at a cutoff of k images, and $\triangle r(k)$ be the change in recall that happens between cutoff $k-1$ and cutoff k. Eq. 14 shows the calculation of the traditional Average Precision (TAP), whereas Eq. 15 shows the Interpolated Average Precision (IAP) – usually, both are all called AP. Because the TAP is not a monotonic function, the IAP is more popular in image retrieval, so we use that to calculate the mAP in our experiments.

$$AP = \sum_{k=1}^{n} P(k)\triangle r(k) = \frac{1}{m}\sum_{k=1}^{m}\frac{k}{s_k}. \tag{14}$$

$$AP = \sum_{k=1}^{n} \max_{\widetilde{k}\geq k}\left\{P(\widetilde{k})\right\}\triangle r(k) = \frac{1}{m}\sum_{k=1}^{m}\max_{\widetilde{k}\geq k}\left\{\frac{\widetilde{k}}{s_{\widetilde{k}}}\right\}. \tag{15}$$

C. EFFECTIVENESS ANALYSIS

Fig. 5 shows some matching examples of baseline and LIST on the market-1501 dataset, which is based on the experimentation of single query measurement. The left in a dashed rectangle is the probe image, and the two rows on the right are the top 1–20 matched results of baseline and LIST. The label under each image is the ID, in which the first character of "P" or "G" means "Probe" or "Gallery" respectively, and the following number is the ID number. The blue decimal under each gallery label is the Vector Cosine Angle (CVA) of that gallery image and the probe image. The images in the red and blue rectangles are the correctly matched and mismatched ones, respectively.

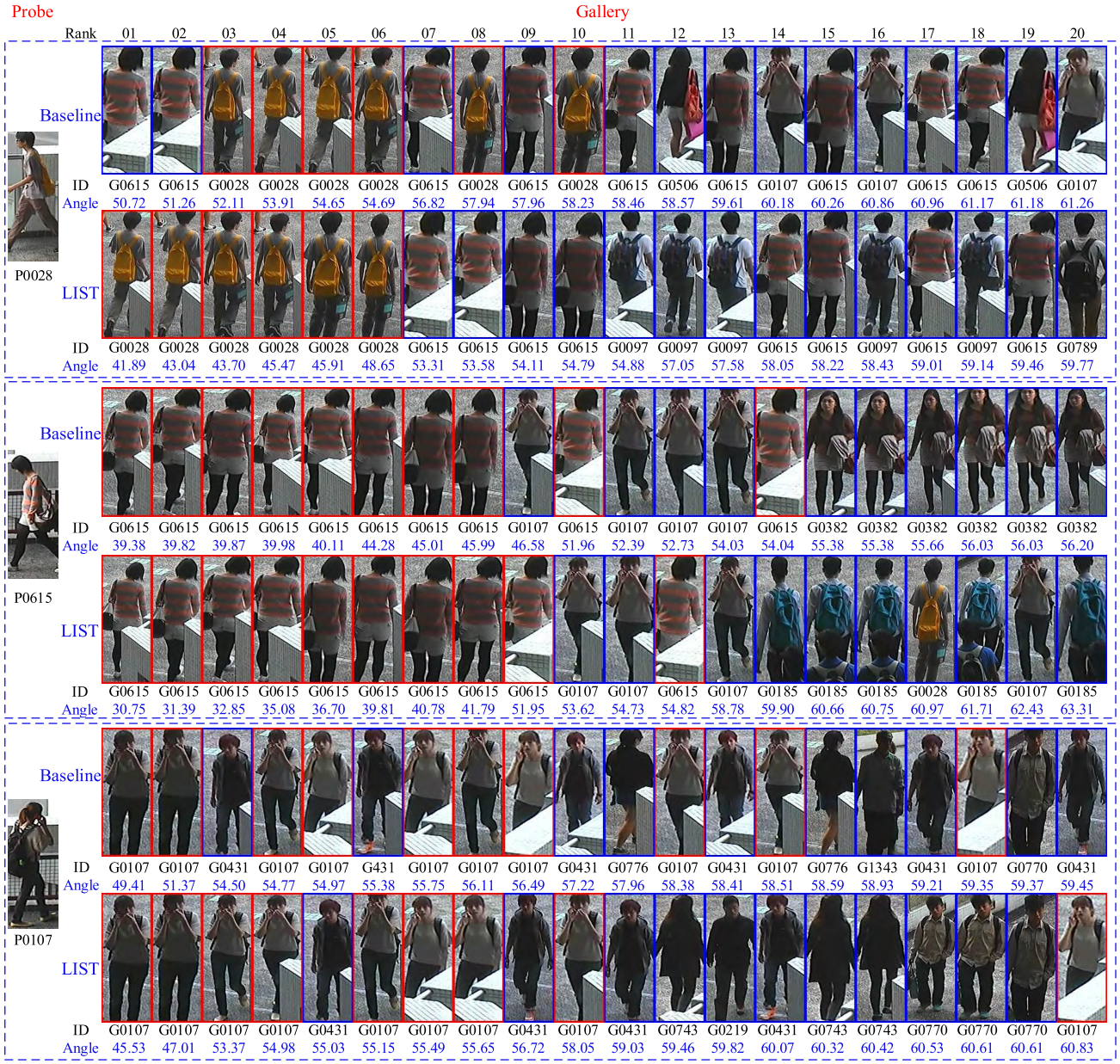


FIGURE 5. Examples comparison of baseline and LIST on the market-1501 dataset.

The first query of P0028 shows that it has a similar background with G0615 and G0107. The LIST method correctly matches G0028 in the top 6 rank whereas the baseline does not, which shows that the LIST has higher discriminative ability. The LIST also has a wider CVA range (from 41.89° to 59.77°) than the baseline (from 50.72° to 61.26°). The queries of P0615 and P0107 also depict that the LIST method is less affected by the similar backgrounds than the baseline.

D. COMPARISON WITH THE BASELINE

Table 2 and Table 3 compare the Inc-LIST and Res-LIST with their baselines, respectively, where the SvsS, SQ rank-1 accuracy and mAP are given. For the two frameworks, it is apparent that the rank-1 SvsS accuracy of our LISTs are higher than that of the corresponding baselines on all six datasets. In addition to that, the rank-1 SQ accuracy and mAP of the LISTs are higher than the baselines on five datasets, except for the i-LIDS dataset.

There are two reasons why the SQ and mAP of the LISTs are lower than the baselines on i-LIDS. (1) The i-LIDS has only 238 training images, which is an insufficient number to collect the discriminative statistics. (2) The occlusion by other humans and accessories (the bags, draw-bar boxes, etc.) in the airport scene results in very different backgrounds,

TABLE 2. Comparison with the Inception-v3 baseline.

Dataset	Method	SvsS	SQ	mAP
CUHK01	Inception-v3	74.05	78.87	79.87
	Inc-LIST	77.63	82.47	82.75
	↑	3.58	3.60	2.88
CUHK03	Inception-v3	81.56	86.76	84.57
	Inc-LIST	87.05	93.07	89.30
	↑	5.49	6.31	4.73
Market-1501	Inception-v3	72.59	84.80	66.54
	Inc-LIST	74.95	86.43	70.06
	↑	2.36	1.63	3.52
PRID 2011	Inception-v3	50.00	36.00	47.02
	Inc-LIST	55.00	39.00	50.55
	↑	5.00	3.00	3.53
iLIDS	Inception-v3	59.68	73.83	66.65
	Inc-LIST	62.48	70.09	65.98
	↑	2.80	-3.74	-0.67
VIPeR	Inception-v3	61.00	61.00	72.99
	Inc-LIST	72.00	72.00	80.92
	↑	11.00	11.00	7.93

TABLE 3. Comparison with the Resnet50 baseline.

Dataset	Method	SvsS	SQ	mAP
CUHK01	Resnet50	62.14	69.38	69.27
	Res-LIST	65.06	70.31	72.31
	↑	2.92	0.93	3.04
CUHK03	Resnet50	76.60	84.45	81.40
	Res-LIST	78.24	86.13	82.74
	↑	1.64	1.68	1.34
Market-1501	Resnet50	65.77	79.33	59.04
	Res-LIST	68.36	81.41	61.96
	↑	2.59	2.08	2.92
PRID 2011	Resnet50	48.00	27.00	39.88
	Res-LIST	51.00	30.00	42.05
	↑	3.00	3.00	2.17
iLIDS	Resnet50	49.70	66.36	57.02
	Res-LIST	50.17	64.49	56.82
	↑	0.47	-1.87	-0.20
VIPeR	Resnet50	62.00	62.00	73.92
	Res-LIST	68.00	68.00	77.34
	↑	6.00	6.00	3.42

which leads to the results that the background pixels being in the minority count and the representation space is expanded. As such, the inaccurate and false discriminative information leads to a negative influence for ranking.

The performance gain on the VIPeR dataset is much higher than the other ones, because the background of VIPeR is more similar than that of other datasets. Thus, the most background pixels are in the majority count, and their representation space is shrunk. This protrudes the foreground pixels and achieves better performance.

E. COMPARISON WITH STATE-OF-THE-ART METHODS

We compare the results of the proposed Inc-LIST and Res-LIST approaches against other state-of-the-art methods using the SvsS CMC curve with the top 1–20 ranks, the single query top 1, 5 and 10 ranks, and the mAP evaluation. All the compared results come from their published papers. Some papers only report discrete ranks (e.g. top 1, 5, 10 and 20 ranks), so we connect them in the CMC curve graph and ignore the missing ranks. For those reports ranking less than 20 (e.g. top 1 to 10 ranks), we only draw the reported segments and neglect the remainders.

This research compares LIST with 19 person ReID approaches using the SvsS CMC curve, including AMC-SWM [60], CIND [8], CSL [15], DGD [3], DVDL [61], GOG [62], JLSCR [12], JUDEA [9], LDCF [23], LOMO [59], LRME [63], LSSCDL [64], MLAPG [65], MTL-LORAE [66], Quadruplet [27], RDC [11], SCSP [67], SSM [57], and TCP [10], which are shown in Fig. 6. Experimental results show that our Inc-LIST method has increased over the rank-1 performance by 10.42%, 19.52%, 5.83%, 14.4%, 2.08% and 18.27% on the datasets of CUHK03 (87.05%), Market-1501 (74.95%), CUHK01 (77.63%), PRID2011 (55.00%), i-LIDS(62.48%) and VIPeR(72.00%), respectively. In addition to those facts, the TCP method outperformed the LIST from the top 4 rank on the i-LIDS dataset, because there are no sufficient images to collect the discriminative statistics.

TABLE 4. Comparison with state-of-the-art.

Market-1501	Rank-1	Rank-5	Rank-10	mAP
Embedding [55]	79.51	–	–	59.87
APR [53]	84.29	93.20	95.19	64.67
TriNet [54]	84.92	94.21	–	69.14
MSCAN [23]	80.31	–	–	57.53
Re-ranking [56]	77.11	–	–	63.63
SSM [57]	82.21	–	–	68.80
Our Res-LIST	**81.41**	**87.41**	**89.88**	**61.96**
Our Inc-LIST	**86.43**	**94.77**	**96.79**	**70.06**
CUHK03	Rank-1	Rank-5	Rank-10	mAP
Embedding [55]	73.10	87.00	92.50	68.20
Re-ranking [56]	61.60	–	–	67.60
LOMO+XQDA [58]	46.30	78.90	88.60	–
SI-CI [12]	52.20	84.30	94.80	–
DNS [59]	54.70	80.10	88.30	–
Our Res-LIST	**86.13**	**88.03**	**89.81**	**82.74**
Our Inc-LIST	**93.07**	**96.53**	**98.11**	**89.30**

When we compared Inc-LIST with ReID methods using the SQ and mAP evaluation, we only provided the results on CUHK03 and Market-1501 because of the lack of reported performance on the other datasets. Table 4 shows the comparison results with 9 methods, including APR [53], DNS [59], Embedding [55], LOMO+XQDA [58], MSCAN [23], Re-ranking [56], SI-CI [12], SSM [57] and TriNet [54]. It is evident that our Inc-LIST method outperforms all the compared state-of-the-art ones.

It must be noted that, to overcome the scale issue of small datasets, we firstly train the networks by jointly crossing the six datasets, and then fine-tune on each dataset. This makes the features more robust on different datasets but less discriminative on the larger ones themselves [3]. In the compared state-of-the-art methods, DGD, CIND, Quadruplet and TriNet trained in the same or a similar way, but the others did not.

Another issue to be aware of is that ResNet50 performs much worse than the Inception-v3 when we regard the ReID as a classification problem, and this leads to the similar result for Res-LIST and Inc-LIST. Using softmax with loss in

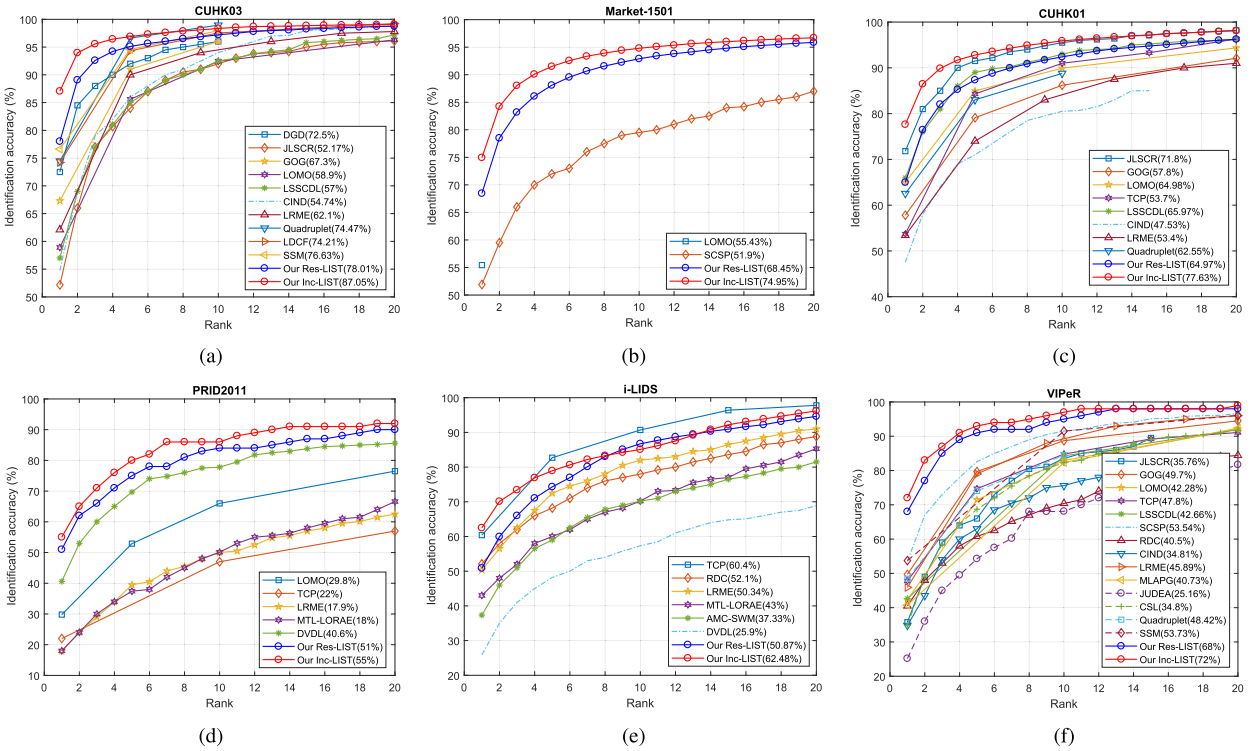


FIGURE 6. CMC curves of different methods on six datasets. (a) CUHK03. (b) Market-1501. (c) CUHK01. (d) PRID2011. (e) i-LIDS. (f) VIPeR.

ResNet50 is also worse than the triplet loss (TriNet), which leads to the worse performance of Res-LIST than the TriNet in Table 4.

VI. CONCLUSION

This paper focused on the feature map distribution and found a rule to quantitatively measure the discriminative power of the feature map. The study proposed a method to distinguish individual data from general ones in order to separate the space into several segments. It also presented an irregular space transformation that enhances the classification accuracy. The proposed approach achieved high discriminative and generalization power. We experimented on six person ReID datasets to validate the effectiveness of our method. By applying on the Inception-v3 and ResNet50 networks of our LIST layer, we proved that the proposed approach can be used in most of the existing classification CNNs. Moreover, our results exceeded state-of-the-art methods with sufficient training data, and we demonstrated the effectiveness of the proposed method.

ACKNOWLEDGMENT

Thank you for the support from HAWKEYE Group.

REFERENCES

[1] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 1–20.

[2] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke. (2016). "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets." [Online]. Available: https://arxiv.org/abs/1605.09653, doi: 10.1109/TPAMI.2018.2807450.

[3] 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.

[4] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[5] A. Franco and L. Oliveira, "Convolutional covariance features: Conception, integration and performance in person re-identification," *Pattern Recognit.*, vol. 61, pp. 593–609, Jan. 2017.

[6] 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.

[7] 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.

[8] 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.

[9] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3765–3773.

[10] 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.

[11] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.

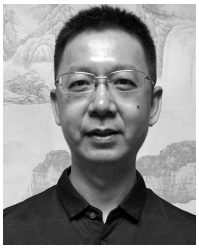
[12] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1288–1296.

[13] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng. (2017). "Deep hybrid similarity learning for person re-identification." [Online]. Available: https://arxiv.org/abs/1702.04858

[14] 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

[15] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3200–3208.
[16] H. Sheng, Y. Huang, Y. Zheng, J. Chen, and Z. Xiong, "Person re-identification via learning visual similarity on corresponding patch pairs," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Cham, Switzerland: Springer, 2015, pp. 787–798.
[17] Y. Huang, H. Sheng, and Z. Xiong, "Person re-identification based on hierarchical bipartite graph matching," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4255–4259.
[18] 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.
[19] 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.
[20] Y. Huang, H. Sheng, Y. Zheng, and Z. Xiong, "DeepDiff: Learning deep difference features on human body parts for person re-identification," *Neurocomputing*, vol. 241, pp. 191–203, Jun. 2017.
[21] 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
[22] 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.
[23] 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.
[24] X. Wang, Q. Wu, X. Lin, Z. Zhuo, and L. Huang, "Pedestrian identification based on fusion of multiple features and multiple classifiers," *Neurocomputing*, vol. 188, pp. 151–159, May 2016.
[25] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. (2016). "End-to-end comparative attention networks for person re-identification." [Online]. Available: https://arxiv.org/abs/1606.04404
[26] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang, "DeepList: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 513–524, Mar. 2017.
[27] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 403–412.
[28] 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.
[29] 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. 3741–3750.
[30] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[31] 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.
[32] G. Hinton, N. Srivastava, A. Krizhevsky, R. R. Salakhutdinov, and I. Sutskever, "Improving neural networks by preventing co-adaptation of feature detectors," *Comput. Sci.*, vol. 3, no. 4, pp. 212–223, Jul. 2012.
[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (PMLR)*, in Proceedings of Machine Learning Research, vol. 37, Jul. 2015, pp. 448–456.
[35] W. Lin *et al.*, "Learning correspondence structures for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2438–2453, May 2017.
[36] X. Liu *et al.*, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 350–359.
[37] 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. 907–915.
[38] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *J. Mach. Learn. Res.*, vol. 15, no. 4, pp. 315–323, 2011.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.
[40] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi. (2014). "Learning activation functions to improve deep neural networks." [Online]. Available: https://arxiv.org/abs/1412.6830
[41] 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.
[42] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 31–44.
[43] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3594–3601.
[44] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scand. Conf. Image Anal.* Berlin, Germany: Springer, 2011, pp. 91–102.
[45] W. S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 91–110.
[46] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, vol. 3, 2007, pp. 1–7.
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[48] 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.
[49] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093
[50] L. Bazzani, M. Cristani, and V. Murino, "SDALF: Modeling human appearance with symmetry-driven accumulation of local features," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 43–69.
[51] L. Zheng, Y. Yang, and A. G. Hauptmann. (2016). "Person re-identification: Past, present and future." [Online]. Available: https://arxiv.org/abs/1610.02984
[52] Z. Zheng, L. Zheng, and Y. Yang. (2016). "A discriminatively learned CNN embedding for person re-identification." [Online]. Available: https://arxiv.org/abs/1611.05666
[53] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. (2017). "Improving person re-identification by attribute and identity learning." [Online]. Available: https://arxiv.org/abs/1703.07220
[54] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737
[55] Z. Zheng, L. Zheng, and Y. Yang. (2016). "A discriminatively learned CNN embedding for person re-identification." [Online]. Available: https://arxiv.org/abs/1611.05666
[56] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1318–1327.
[57] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3356–3365.
[58] 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.
[59] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.
[60] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4678–4686.
[61] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4516–4524.
[62] 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.
[63] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1846–1855.

[64] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1278–1287.
[65] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3685–3693.
[66] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3739–3747.
[67] 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.



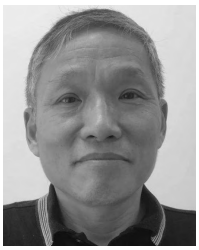
YANWEI ZHENG received the B.S. degree from Shandong Jianzhu University, Jinan, China, in 1999, and the M.S. degree from Shandong University, Jinan, in 2004. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, Beijing, China. He joined the University of Jinan, Jinan, and became a lecturer from 2004 to 2013. His research interests include machine learning, computer vision, and especially person re-identification.

HAO SHENG received the B.S. and Ph.D. degrees from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2003 and 2009, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, pattern recognition, and machine learning.

YANG LIU received the B.S. degree from the School of Advanced Engineering, Beihang University, Beijing, China, in 2009, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interests include deep learning, computer vision, and especially person re-identification.

KAI LV received the B.S. degree from the School of Computer Science and Technology, Tianjin University of Science and Technology, Tianjin, China, in 2012. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include machine learning, computer vision, and penson and vehicle re-identification.

WEI KE received the Ph.D. degree from the School of Computer Science and Engineering, Beihang University. He is currently an Associate Professor of the Computing Program, Macau Polytechnic Institute. His research interests include programming languages, image processing, computer graphics, and tool support for object-oriented and component-based engineering and systems. His recent research focuses on the design and implementation of open platforms for the applications of computer graphics and pattern recognition, including programming tools, environments, and frameworks.



ZHANG XIONG received the B.S. degree from Harbin Engineering University in 1982 and the M.S. degree from Beihang University, Beijing, China, in 1985. He is currently a Professor with the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, information security, and data vitalization.