Orientation-Guided Similarity Learning for Person Re-identification

Na Jiang*, Junqi Liu*, Chenxin Sun*, Zhong Zhou* and Wei Wu*
* State Key Lab of Virtual Reality Technology and Systems, Beihang University
Beijing, China
Email: {jiangna,zz}@buaa.edu.cn

***Abstract*—Person re-identification (re-id) is a promising topic in computer vision, which concentrates on similarity learning of individuals across different camera views. It remains challenging due to the unpredictable orientation variations, the partial occlusions, and the inaccurate detections. To solve these problems, we present an orientation-guided similarity learning architecture to learn discriminative feature representations and define similarity metric for person re-id. Our proposed architecture explicitly leverages pedestrian orientation and body part cues to enhance the generalization ability. In the architecture, an orientation-guided loss function that pulls the positive samples with the same orientations closer is designed to alleviate the orientation variations. Meanwhile, an aligned dense network with pose estimation is presented to extract robust global-local fusion representations, which effectively exploits local features to overcome partial occlusions. In the end, we introduce a two-stage Top-k re-ranking strategy to optimize initial re-id results by min-hash and weighted distance. Extensive experimental results demonstrate that our proposed approach significantly outperforms state-of-the-art re-id methods on the popular CUHK03, Market1501, and DukeMTMC-reID datasets.**

I. Introduction

Person re-identification (re-id) refers to the retrieval of specific probe images from large-scale gallery images or surveillance videos [1], [2]. It recently attracted increasing attentions since it has many possible applications in such areas as inter-camera tracking and anti-terrorism. However, the appearances displayed in detected images are apt to change with the pedestrian orientation, posture, occlusion, and monitoring environments. This significantly increases the difficulty of the person re-id problem. To mitigate the effects of these disturbing factors, various methods are proposed [3], [4], [5], [6], and recent deep learning based approaches exhibit promising performance and potentials [7], [8]. These deep learning based approaches treat person re-id as a classification problem and depend on the loss functions to train network parameters. Existing loss functions guide the network parameters learning the inter-class and intra-class similarity constraints for classification, while ignoring the influence of pedestrian orientation on these similarity constraints. It weakens the generalization capacity of the networks. Taking the original triplet loss function as an example [4], the trained inter-class similarity constraint with the function fails on the testing sets with unseen data distribution, especially on the gallery images with the same orientation as the probe images (see Fig. 1(a)).

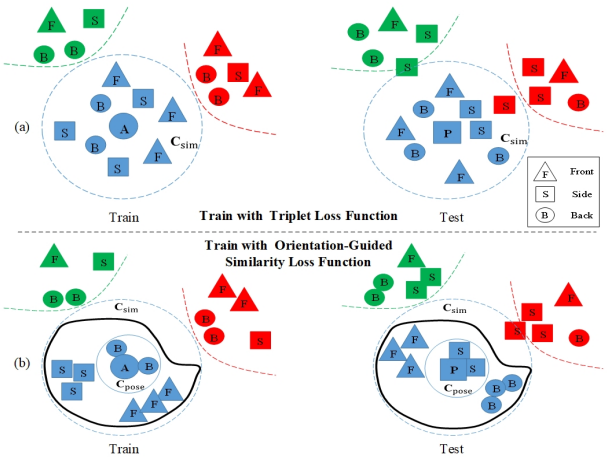

Fig. 1. A contrastive illustration between original triplet loss funtion and orientation-guided similarity loss function. Colors denote to identities, and shapes refer to orientations. A represents the anchor images, P denotes the probe images, the dotted lines represent the inter-class similarity constraint C_{sim}, and the blue solid lines represent the orientation-guided similarity constraint C_{pose}. What is more, the classifications guided by our proposed loss function are surrounded by the bold lines. (Best viewed in color)

To solve the problem, we propose the orientation-guided similarity learning architecture, which consists of a similarity learning module and a feature extraction module. In the similarity learning module, we introduce the pedestrian orientations to design an orientation-guided similarity loss function. It can pull the positive samples with the same orientations closer and impose the orientation-guided similarity constraint to alleviate the orientation variations and improve the generalization of the proposed architecture. As shown in Fig. 1(b), the model trained by the orientation-guided similarity loss function makes the positive samples with the same orientations closer and minimize misclassifications caused by the lack of orientation-guided similarity constraint. In the feature extraction module, we modify and design an aligned dense framework to extract discriminative representations. The discriminative representations are generated with the combination of global features and three major local features, which are conducive to improving the feature resistance to partial occlusions and inaccurate detections. Compared with existing methods exploiting fixed ratio horizontal stripes or body part models to infer partial regions of interest (ROIs) for local feature extraction [4],

[9], our aligned dense framework can achieve more accurate body part ROIs since we introduce 2D joint points from pose estimation [10]. Different from other methods of using pose estimation to infer body parts [1], [11], our feature extraction framework adopt the dense connection used in DenseNet. It can strengthen feature propagation and support feature reuse. These differences and improvements enable our proposed framework to take full advantage of local features and realize feature alignment. Since the appearance features extracted from the single image is not comprehensive enough to achieve the best similarity metric, we develop a two-stage Top-k re-ranking strategy to further optimize feature matching by min-hash and weighted distance. Extensive experimental results on three popular public datasets show that our proposed approach significantly outperforms state-of-the-art re-id methods.

II. RELATED WORKS

Most existing methods can be classified into two important components of feature extraction and metric learning. Therefore, we elaborate on reviewing existing works concerning feature extraction [5], [6], [7], [8], [11], [12], [13], [14], [15] and metric learning [4], [9], [10], [16], [17], [18], [19], [20] in this section.

A. Feautre Extraction

In the traditional person re-id approaches without deep learning, the color features and hand-crafted features are often employed as feature descriptors [9]. Extracting these features is simple and efficient, whereas the discrimination of these features will be weakened when pedestrian orientations or monitoring environments change among different cameras. To enhance the robustness of appearance features, deep learning structures [21], [22] are introduced in person re-id. Xiao et al. propose a CNN framework with domain guided dropout to improve the feature representation [5]. It achieves significant improvement compared to the traditional approaches. Since then, other deep learning based re-id methods such as Deepreid [6], Gated Network [23], SVDNet [7], and Spindle Net [11] have been put forward to further improve the performance of person re-id. To increase data volume and prevent the overfitting, numerous large-scale datasets like CUHK03 [6], Market1501 [24], and DukeMTMC-reID [7] have been released successively. Zheng et al. also exploit generative adversarial networks (GANs) to generate unlabeled data for data augmentation [8]. The method enhances the generalization ability of the trained model by expanding the training sets. Although the above-mentioned approaches have proposed multifarious contributions, most of them neglect the crucial orientation factor. In this paper, we shed a new light on the exploiting of the orientation, and propose an orientation-guided similarity learning architecture by taking account of the orientation factor.

B. Metric Learning

Many metric learning algorithms have been proposed to optimize the distance metric, for instance, cross-view quadratic discriminant analysis (XQDA) [9] and Discriminative Null Space [25]. Most of them depend on complex mathematical formulas and are independent of the feature extraction. With the advent of end-to-end person re-id architectures based on deep learning, there are some subtle changes in the metric learning algorithms. In addition to calculating the similarity distances between images or sequences in the test phase, training the network model has also become their responsibility. Hence, the metric learning begins to pay attention to the design of loss functions. Zheng et al. propose the joint training strategy of double loss functions [18], which gives inspiration to our training methods. Compared with the network models trained only with one loss function, the joint training strategy of multiple loss functions can significantly improve the performance of deep learning architectures. Moreover, there are two methods which focus on designing loss functions [4], [19] worth studying. They aim to train a larger inter-class similarity constraint and a smaller intra-class similarity constraint compared to the original triplet loss function. Different from them, we not only introduce pedestrian orientations to add orientation-guided similarity constraint, but also introduce body part cues to achieve more accurate body part ROIs for local feature extraction. These improvements help our proposed architecture to make better use of local features and loss functions. To our best knowledge, this is the first reported effort to take orientation cues into consideration and design the similarity loss function.

III. THE PROPOSED PERSON RE-IDENTIFICATION APPROACH

In this section, we describe the overall outline of the proposed person re-id approach (see Fig. 2) , where we mainly introduce the similarity learning based on orientation-guided similarity loss function, aligned dense framework, and two-stage Top-k re-ranking strategy.

A. Outline

Fig. 2 shows the outline of our proposed person re-id method, which consists of quintuple inputs, orientation-guided similarity learning architecture, and re-ranking strategy. For any given image, we first employ 2D pose estimation to detect the joint points, and then infer the pedestrian orientations and calculate the body part ROIs. The orientations are the major basis for dividing the images as quintuple inputs for training, and the body part ROIs will be transformed into ROI Pooling layer for local feature extraction. Secondly, we train the orientation-guided similarity learning architecture to simultaneously extract discriminative global features and robust local features. In the training phase, we exploit the softmax loss function and the proposed orientation-guided similarity loss function to jointly train the network parameters. The fusion of features and the joint training strategy are beneficial to alleviate the influences of orientation variations and occlusions. At last, we introduce a two-stage Top-k re-ranking strategy which automatically adjusts the order of re-id

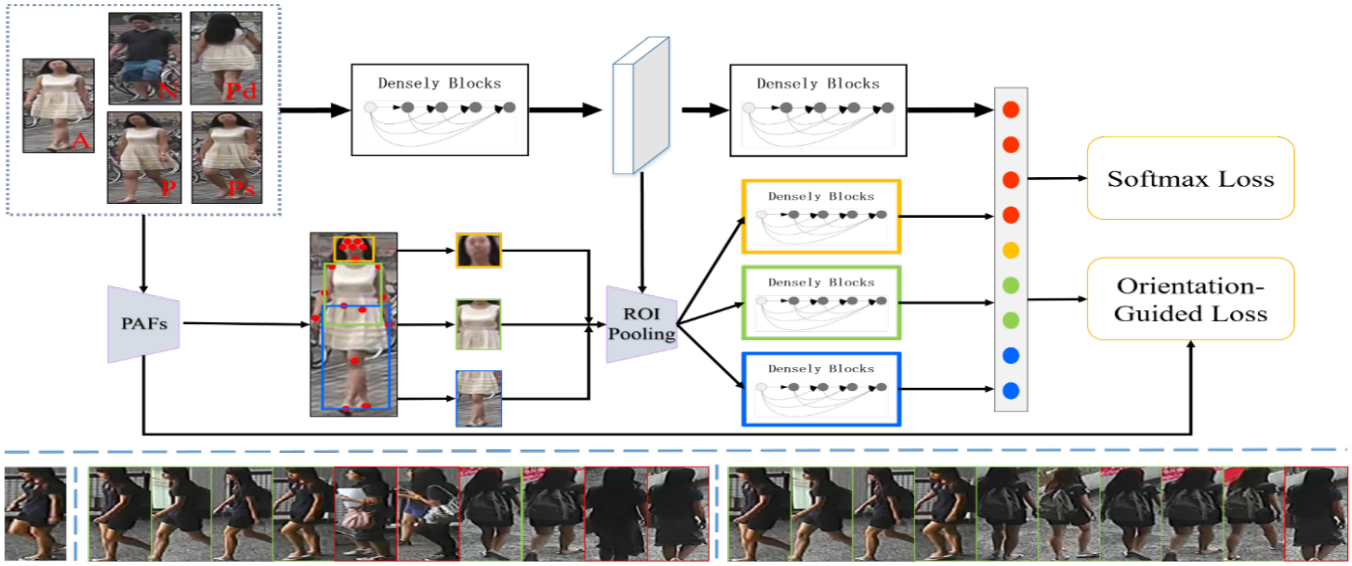

Fig. 2. Outline of our proposed method. A denotes an anchor image, P represents the positive sample, N indicates the negative sample, P_s refers to the positive samples that have the same orientation with the anchor images and P_d shows the different orientation ones. In Fig.2 (b), the thickness of lines means the weights among feature distances (Best viewed in color).

results to further improve the generalization and accuracy of person re-id. Specific methods and details are described below.

B. Orientation-Guided Similarity Learning Architecture

1) Orientation Estimation and Body Parts Localization: *Orientation Estimation*. In our proposed architecture, the orientation-guided similarity loss function requires datasets to be divided into three subsets of the front, back and side to constitute quintuple inputs. To meet this demand, we first calculate the clockwise angles between the pedestrian shoulder vector V_i and the vertical vector $V_{vertical}$ (from top to bottom) to estimate the orientation and classify the training images (see Fig. 3). The clockwise angle is defined as:

$$V_i = p_{rsho} - p_{lsho} = \{x_{rsho} - x_{lsho}, y_{rsho} - x_{lsho}\} = \{x_v, y_v\} \quad (1)$$

$$\theta_i = \begin{cases} 0^\circ & ,where\ x_v = 0, y_v < 0 \\ 180^\circ & ,where\ x_v = 0, y_v > 0 \\ arccos\dfrac{V_i \cdot V_{vetical}}{\|V_i\| \times \|V_{vetical}\|} & ,where\ x_v > 0 \\ 360 - arccos\dfrac{V_i \cdot V_{vetical}}{\|V_i\| \times \|V_{vetical}\|} & ,where\ x_v < 0 \end{cases} \quad (2)$$

where the shoulder vector V_i is started from left shoulder and ended at right shoulder, the shoulder joint points are detected from Part Affinity Fields (PAFs) [10], $\|\ \|$ represents L2-norm.

According to the angles calculated by Eq.2, all images in the datasets can be divided into three subsets by Eq.3. Each subset covers120°, and then the front images are labeled to 1, the side group is labeled to 2, and the backs are labeled to 3.

$$c_i = \begin{cases} 1, where\ \theta_i \in [210^\circ, 330^\circ] \\ 2, where\ \theta_i \in (150^\circ, 210^\circ) \cup [0^\circ, 30^\circ) \cup (330^\circ, 360^\circ] \\ 3, where\ \theta_i \in [30^\circ, 150^\circ] \end{cases} \quad (3)$$

where c_i indicates the orientation label of the i-th image. If the left or right shoulder joint point is lost, the c_i will be marked as 2 directly. During the training, the samples that fail to achieve joint points are discarded. It will effectively eliminate the low-quality training samples, which helps to improve network convergence and robustness.

(a) (b) (c)

Fig. 3. Examples of orientation estimation. The red arrow indicates the pedestrian shoulder vector V_i and the blue arrow represents the vertical vector $V_{vertical}$.

Body Parts Localization. To deal with the occlusion, the idea of combining local features with global features has been studied in many person re-id studies [4], [9]. They generally use fixed ratio horizontal stripes or body part models to infer the locations of ROIs for local feature extraction (see Fig. 4(b)). While handling the images with misalignment or orientation variations, mandatory ROIs will bring noises and weaken the effect of local features.

The phenomenon stimulates us to consider accurate ROIs so as to give full play to local features. Therefore, we introduce pose estimation to detect joint points and then calculate the locations of body part ROIs based on these joint points. Considering the fact that arms are easily obscured and have inferior discrimination, we select head, torso and legs as major ROIs. The head S_{head} is determined by the joint points index Set_a

(a) (b)

Fig. 4. The comparison of ROIs generated by different methods. Fig.4 (a) is the division generated by joint points, Fig.4(b) shows the results from horizontal stripes. Obviously, the accurate ROIs generated by the joint points can be used to extract the correct local features from the non-aligned persons, which is very beneficial for feature alignment in the phase of similarity metric.

=[1,2,17,18], the torso S_{torso} and the legs S_{leg} are depended on Set_b =[3,4,5,6,7,8,9,12], Set_c =[9,10,11,12,13,14], respectively. The eighteen joint points detected by PAFs are sequentially represented as follows: nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, left ankle, left eye, right eye, left ear, right ear. For each ROI index set, we calculate the tightest horizontal bounding box that can cover all joint points in the as the corresponding ROI. As shown in Fig. 4(a), the body part ROIs inferred by joint points are not only more accurate, but also can effectively cope with the misalignment of pedestrian images.

2) Orientation-Guided Similarity Loss Function: To introduce the orientation cues, we design an orientation-guided similarity loss function for the large-scale person re-id. It stems from two important rules that summarized through a large number of experiments: 1) the feature distances between positive sample pairs are smaller than the feature distances between negative sample pairs; 2) the feature distances between positive sample pairs with same orientation are also smaller than the feature distances between positive sample pairs with different orientation. The first rule can be formulated as inter-class similarity constraint and can be included in the original triplet loss function. It is defined as follow:

$$D_{id}(I_i^a, I_i^p, I_i^n) = [d(f(I_i^a), f(I_i^p)) - d(f(I_i^a), f(I_i^n)) + \alpha]_+ \quad (4)$$

$$d(x, y) = \|x - y\|_2^2 \quad (5)$$

$$[x]_+ = max(x, 0) \quad (6)$$

where $D_{id}(I_i^a, I_i^p, I_i^n)$ represents inter-class similarity constraint that the feature distances between all positive samples are smaller than the feature distances between negative samples. I_i^a represents the anchor image in a triplet input, I_i^p denotes the positive sample of anchor image, I_i^n expresses the negative sample of the anchor image. $d(x, y)$ represents the L2-norm distance between x and y. α is a limit margin between positive and negative samples, N is the number of triples. To meet the second rule, we introduce the orientation-guided similarity constraint $D_{pose}(I_i^a, I_i^{ps}, I_i^{pd})$ based on the inter-class similarity constraint and define it as follow:

$$D_{pose}(I_i^a, I_i^{ps}, I_i^{pd}) = [d(f(I_i^a), f(I_i^{ps})) - d(f(I_i^a), f(I_i^{pd})) + \beta]_+ \quad (7)$$

where I_i^{ps} indicates positive sample of I_i^a with the same posture, I_i^{pd} refer to the positive sample of I_i^a with different postures. β is a threshold of orientation-guided similarity constraint which is used to pull the positive samples with the same orientations. N is the number of inputs. The constraints defined by Eq.4 and Eq.7 are then be employed to derive the orientation-guided similarity loss function. It is summarized as Eq.8:

$$L_{quin}(I, w) = \frac{1}{N}\sum_{i=1}^{N}(D_{id}(I_i^a, I_i^p, I_i^n) + \lambda D_{pose}(I_i^a, I_i^{ps}, I_i^{pd})) \quad (8)$$

where λ is a weight of balancing the two similarity constraints, and w represents the current network parameters.

C. Aligned Dense Framework.

Comparing with the most outstanding the ResNet [21] with the DenseNet [22], we find that the DenseNet which reuses low-level semantic features at the latter layers by the skip-connection is more suitable for the fine-grained person re-id than the ResNet which focuses on solving the gradient vanishing problem. We thus modify and design an aligned dense framework based on DenseNet, as shown in Table I.

TABLE I
DETAILED STRUCTURE OF THE ALIGNED DENSELY FRAMEWORK

Layers	Backbone $(k \times k/s/p) \times (num)$	Head $(k \times k/s/p) \times (num)$	Torso/Legs $(k \times k/s/p) \times (num)$
Conv1	$(7 \times 7/2/3) \times 1$	–	–
Max Pool	$(3 \times 3/2/1) \times 1$	–	–
Den1	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 6$	–	–
Den2	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 12$	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 4$	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 6$
Den3	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 24$	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 8$	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 12$
Den4	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 16$	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 4$	$\begin{pmatrix}1\times1/-/-\\3\times3/-/1\end{pmatrix} \times 8$
Ave Pool	$(8 \times 8/-/-) \times 1$	$(8 \times 8/-/-) \times 1$	$(8 \times 8/-/-) \times 1$
FC	512	128	256

As illustrated in Table I, the feature extraction framework includes a backbone network (growth rate k = 32) and three branch networks (growth rate k = 64). They share the weights from Conv1 to Den1. Due to the sizes of the inferred ROIs are not fixed, we add a ROI Pooling layer behind the Den1 to connect the shared feature maps and branch networks. The branch structure of torso is the same with the legs, but they are different from the branch of head. In the three branch networks, we also set different output size for each fully connected layer to adjust the proportion of global features and local features. To accelerate the convergence and alleviate the impact of unseen data, BN layers and ReLUs are also inserted behind each CNN layer in the aligned dense framework.

After extracting features, we first use the Euclidean distance to achieve the initial re-id results. Considering the feature distances between single image pairs are not comprehensive, we present a two-stage re-ranking method to further optimize the results. We define the initial Top-k results of probe image p as $N(p,k) = r_1, r_2, \ldots, r_k$. The similar images of $i-$th result r_i are defined as $N(r_i,k)$. It means the re-id results that the result r_i is regarded as a probe image. In the first phase of re-ranking, we replace the Euclidean distances between the probe image and the current Top-k results with the minhash values between $N(p,k)$ and $N(r_i,k)$, and then adjust the order of results to improve the accuracy. In the second phase, we first assume that the first m results achieved from the first phase are completely correct. And then we require that subsequent results not only are similar to the probe images but also similar to the first m results. The feature distances between the probe image and the current results are updated again with the weighted distance, which is calculated as follows:

$$d_{re} = \sum_{i=1}^{m} \rho_i \cdot d(r_i, r_j) + \rho_p \cdot d(p, r_j) \quad m < j < k \tag{9}$$

where $d_{re}(*,*)$ denotes the weighted distance, ρ_p and ρ_i are the weights of feature distances, whose specific values can be determined empirically. In this paper, ρ_p and ρ_i are set to 0.6, 0.1, respectively. m is empirically set to 4.

IV. EXPERIMENTAL RESULTS

In this section, we conduct the following experiments to analyze the effectiveness of contributions and evaluate the proposed person re-id algorithm by comparing with the state-of-the-art methods. These experiments are run on the server with GTX TITAN XP and Xeon E5 CPU.

A. Analysis of Contribution Effectiveness

To verify that the effectiveness of contributions described in this paper, we design the following analysis experiments which are implemented on the Caffe platform. The experimental results from different optimization are shown in Table II.

TABLE II
ANALYSIS RESULTS OF CONTRIBUTIONS ON DUKEMTMC-REID

Setting	DukeMTMC-reID	
	Rank-1	mAP
baseline	67.95	47.49
Global+Local	69.88	48.96
Softmax+Triplet	71.23	50.00
Softmax+Orientation	73.56	53.46
Our Method (O)	76.12	58.05
Our Method (R)	**76.35**	**63.69**

In Table II, the baseline results are generated from the backbone network, and the double loss results don't consider of local features. O represents original results from orientation-guided similarity learning architecture and R represents the results of our whole architecture with re-ranking strategy. As illustrated in Table II, the local feature and the proposed loss function improve the rank-1 accuracy by 1.93% and 5.61% compared with the baseline, respectively. Meanwhile, Table II also shows that our proposed re-ranking strategy further improves the algorithm performance, especially the mAP. The effectiveness of contributions can also be seen in retrieval results of person re-id (see Fig. 5).

Fig. 5. Retrieval Results of our method on different datasets. The first column represents probe images, and each row demonstrates corresponding retrieval results. The green bounding boxes indicate the results which identities are same with the probe images, the red images are different ones.

As shown in Fig. 5(a), some images with partial occlusion can be retrieved correctly, which further proves that the aligned dense framework can alleviate the occlusion and misalignment. And Fig. 5(b) displays the adaptability of our method to orientation variations, owing to the proposed loss function and the introduction of orientation cues.

B. Performance Comparison on Public Datasets

The proposed method is compared with some recent state-of-the-art algorithms on three large-scale public datasets. These experiments are repeated for 10 times and the average performances are demonstrated in Table III and IV.

TABLE III
COMPARISON WITH STATE-OF-THE-ART APPROACHES ON MARKET1501 AND DUKEMTMC-REID

Market1501	Rank-1	mAP	Duke-reID	Rank-1	mAP
Gated[23]	65.88	39.55	ResN50[17]	65.22	22.99
ResN50[17]	73.90	47.78	GAN[7]	67.68	47.13
Re-rank[16]	77.11	63.63	OIM[26]	68.1	–
Siamese[18]	79.51	59.87	Siamese[18]	68.9	49.3
ACRN[12]	83.61	62.60	APR[13]	70.69	51.88
PDC[1]	84.14	63.41	PAN[14]	71.59	51.51
APR[13]	84.29	64.67	ACRN[12]	72.58	51.96
Our Method	**87.11**	**70.23**	Our Method	**76.35**	**63.69**

TABLE IV
COMPARISON WITH STATE-OF-THE-ART APPROACHES ON CUHK03

CUHK03	Labeled		Detected	
	Rank-1	Rank-5	Rank-1	Rank-5
LOMO+XQDA[9]	52.20	82.23	46.25	78.90
NSFT[25]	62.55	90.05	54.70	84.75
GOG[15]	67.30	91.00	65.50	88.40
EDM[20]	61.32	88.90	52.09	82.87
Context-aware[2]	74.21	94.33	67.99	91.04
PDC[1]	88.70	**98.61**	78.29	94.83
Our Method	**89.29**	97.50	**83.31**	**96.93**

In the Table III and IV, the bold fonts represent the best results. Experimental results demonstrate that our proposed method has achieved excellent performance on both Rank-1 and mAP. Especially, PDC and our method which both introduce pose estimation into person re-id achieve better performance than others, which further demonstrates the validity of orientation-guided similarity loss function.

V. CONCLUSION

In this paper, we present an orientation-guided similarity learning architecture and a two-stage Top-k re-ranking strategy for person re-id. The proposed architecture specifically leverages the pedestrian orientations and body part ROIs to learn discriminative feature representations. The re-ranking strategy effectively employs the feature distances between similar image groups to optimize the similarity metric. Extensive experimental results on three popular datasets demonstrate that our proposed approach is superior to many state-of-the-art methods. In our future work, we will extend this work to the video-based person re-id and explore the multiple query strategy.

ACKNOWLEDGE

This work is supported by the Natural Science Foundation of China under Grant No.61572061, 61472020, 61502020, the National 863 Program of China under Grant No.2015AA016403 and the China Postdoctoral Science Foundation under Grant No. 2013M540039.

REFERENCES

[1] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *2017 IEEE International Conference on Computer Vision*, 2017, pp. 3980–3989.

[2] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.

[3] T. Matsukawa and E. Suzuki, "Person re-identification using cnn features learned from combination of attributes," in *Pattern Recognition, 2016 23rd International Conference on*, 2016, pp. 2428–2433.

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[5] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1249–1258.

[6] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[7] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, 2017.

[8] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," *arXiv preprint arXiv:1703.05693*, 2017.

[9] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2017, p. 7.

[11] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1077–1085.

[12] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Computer Vision and Pattern Recognition Workshops, 2017 IEEE Conference on*, 2017, pp. 1435–1443.

[13] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.

[14] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *arXiv preprint arXiv:1707.00408*, 2017.

[15] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1363–1372.

[16] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3652–3661.

[17] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[18] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, p. 13, 2017.

[19] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2017.

[20] 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 *European Conference on Computer Vision*, 2016, pp. 732–748.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.

[23] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*, 2016, pp. 791–808.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[25] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1239–1248.

[26] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3376–3385.