2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)

PRESERVING STRUCTURAL RELATIONSHIPS FOR PERSON RE-IDENTIFICATION

Liqiang Bao[1], *Bingpeng Ma*[1], *Hong Chang*[2], *Xilin Chen*[2,1]

[1]University of Chinese Academy of Sciences, Beijing 100049, China.
[2]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Bejing, 100190, China.
baoliqiang16@mails.ucas.ac.cn, bpma@ucas.ac.cn, {changhong, xlchen}@ict.ac.cn

ABSTRACT

Recent processes on many computer vision and multimedia researches heavily rely on Convolutional Neural Network (CNN) with pooling layer incorporated, where pooling operation reduces the amount of parameters and brings in translation invariance. However, we discover that pooling operation may destroy valuable structural relationship information, leading to defective feature learning in tasks such as person re-identification. In this paper, we propose a method called Structural Relationship Learning (SRL) to capture structural relationships by constructing a spatially structured graph based on the convolved features and propagate information over the edges. Coupled with pooling operation by metric fusion, SRL provides more comprehensive information for identity discrimination. Experiments are conducted on the iLDIS-VID, PRID2011 and MARS datasets and the results demonstrate the advantages of our proposed method.

***Index Terms*—** pooling layer, structural relationship learning coupled with pooling, spatially structured graph

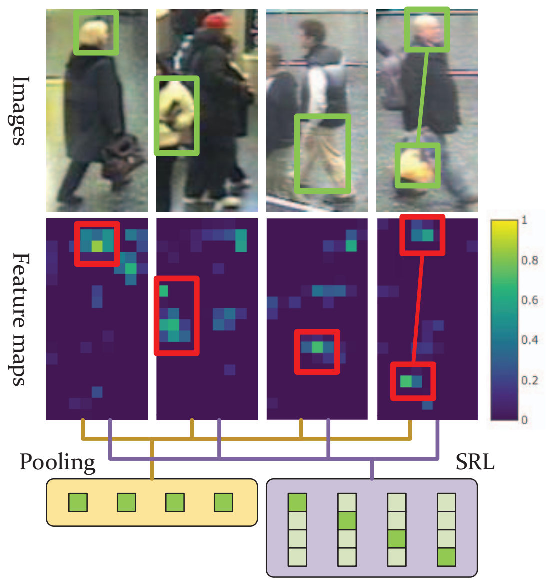

Fig. 1. A demonstration why structural relationship information is important. The feature maps for the given images manifest different structural relationships, which are particularly helpful to distinguish different pedestrians, yet the pooled features reflect no difference. Our proposed SRL is able to generate discriminative features by capturing these structural relationships.

1. INTRODUCTION

Identifying a person who has been previously captured by other non-overlapping cameras is commonly defined as *person re-identification* (ReID). This task has been paid more and more attention for increasing concern of public safety and social service such as criminal tracking and missing people search. Important though it is, this task is quite challenging due to changing environmental settings, illumination conditions, ineluctable occlusion and varying appearances.

Convolutional Neural Network (CNN) is a successful building block of many computer vision tasks including ReID [1, 2, 3, 4]. With the help of pooling operation, CNN not only reduces the the amount of parameters and computation in the networks, but also masters the magic of translation invariance. However, a common concern rises that pooling operation may harm the ability to learn representative features for some tasks. For example, Figure 1 is a visualization of the most relevant feature maps from the last layer of ResNet50 [5] for the yellow objects in the four images. Although the convolved feature maps contain useful regions, manifesting different structural relationships, the pooled features lose them. Therefore, proposing a feature generation method that can preserve the structural relationship information is of high importance to provide clearer clues for determining identities.

The study of preserving structural relationship information mainly focuses on learning spatial attention on feature maps. Jianlou Si et al. [6] propose to use a *dual attention mechanism* to learn context-aware feature sequences and perform dually attentive comparison for person ReID. Shuang Li et.al [7] introduce multiple spatiotemporal attention model

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00028

to automatically discover the diverse set of distinctive body to address the problem of occlusion. Besides, region-based methods are also based on the idea of preserving structural relationships. Dangwei Li et al. [8] stacks designed *multi-scale context-aware network* to learn powerful features over full body and body parts, and use *spatial transformer networks* to learn and localize deformable person parts.

In this paper, we present *Structural Relationship Learning* (SRL) to directly capture structural relationships in an efficient and elegant way. SRL attends over convolved features to model the relationships between useful regions with *Graph Convolutional Network* (GCN), considering that GCN is able to learn hidden layer representations that encode both local graph structure and features of nodes [9]. Intuitively, the pooling operation acts as a feature selection capturing the main information in the convolved feature maps, while SRL is supposed to answer how the information is spatially related.

Compared with the conventional pooling-based method, our method has two advantages: (a) SRL operates directly on a graph that explicitly preserves spatial structure, which learns structural relationship information that is difficult for pooling operation to preserve; (b) SRL works as a separate branch and can be easily embedded into any network to assist feature learning network in extracting more robust and discriminative features.

We evaluate the effectiveness of the proposed SRL in the application of ReID. Most datasets for ReID task contain multiple sets of images or videos captured by different cameras, and pedestrians have been well detected and cropped. Therefore, instead of focusing on translation invariance, we should pay more attention to structural relationships. Based on this consideration, we propose to use SRL coupled with pooling (SRL_P) to combine structural relationship features with the pooled features. The metric fusion of both will eventually lead to a more robust ReID system. Comprehensive ablation studies are conducted on iLIDS-VID, PRID2011 and MARS datasets. Our method achieves results on par with or better than the state-of-the-art approaches on all the datasets.

2. THE PROPOSED METHOD

In this section, we first describe the proposed SRL module in detail. And then we systematically introduce the application of SRL in ReID pipeline.

2.1. Structural Relationship Learning

Structural relationships not only provide the location information of important regions in the image, but also give the correlation or relationships between them. Thus is generally discussed by attention mechanism [6, 7] and region-based methods [10, 8]. Yet we discover that GCN is a more effective and explicit way to achieve this task in that it is able to learn hidden layer representations that encode both local graph structure and features of nodes.

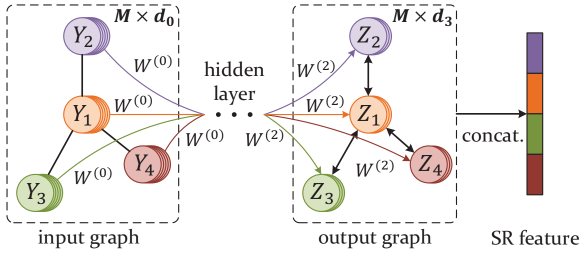

Fig. 2. Demonstration of Structural SRL. In every layer, features that are processed by weight matrix will be propagated over each other to learn the relationships. The graph structure (defined by adjacency matrix $\mathbf{A}$) is shared over layers.

The proposed SRL method is based on GCN, which directly operates on a graph that preserves the spatial structure of the convolved features. Let's denote the middle representation of a single input image as $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$ with C channels and each channel of size $H \times W$. Instead of using pooling operation to compress every channel into a single value and giving up the precious structural relationships, we permute the dimension order of $\mathbf{I}$ and denote it as $\mathbf{X} \in \mathbb{R}^{M \times d_0}$ with $M = H \times W$ features of dimension $d_0 = C$. In order to reveal the spatial structure, a symmetric adjacency matrix $\mathbf{A}$ is constructed with the value at position (i, j) is:

$$\mathbf{A}_{i,j} = \begin{cases} 1 & \mathbf{X}_i \text{ is adjacent to } \mathbf{X}_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\mathbf{X}_i$ is the i^{th} row of $\mathbf{X}$, indicating the i^{th} feature, and so is $\mathbf{X}_j$. $\mathbf{A}$ is a binary matrix which only enables the connection between spatially adjacent features in all four (three for edge features or two for corner features) directions.

For the learning of structural relationships, we showcase a three-layer GCN to operate on the graph defined by the adjacency matrix $\mathbf{A}$. The layer-wise propagation rule of the involved GCN has the following form:

$$\mathbf{Y}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{Y}^{(l)}\mathbf{W}^{(l)}\right). \tag{2}$$

In the above, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ is the adjacency matrix with added self-connection. $\mathbf{I}_N$ is the identity matrix, $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$ and $\mathbf{W}^{(l)}$ is a layer-specific trainable weight matrix. $\sigma(\cdot)$ denotes an activation function, such as the $ReLU(\cdot) = max(0, \cdot)$. $\mathbf{Y}^{(l)} \in \mathbb{R}^{M \times d}$ is the matrix of activations in the l^{th} layer and $\mathbf{Y}^{(0)} = \mathbf{X}$.

Conditioning our SRL model $f(\mathbf{X}, \mathbf{A})$ on both the feature matrix $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$, we have the following learning process:

$$\mathbf{Z} = f(\mathbf{X}, \mathbf{A}) = \hat{\mathbf{A}}\ \left(\hat{\mathbf{A}}\ \text{ReLU}\left(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^{(0)}\right)\mathbf{W}^{(1)}\right)\mathbf{W}^{(2)}. \tag{3}$$

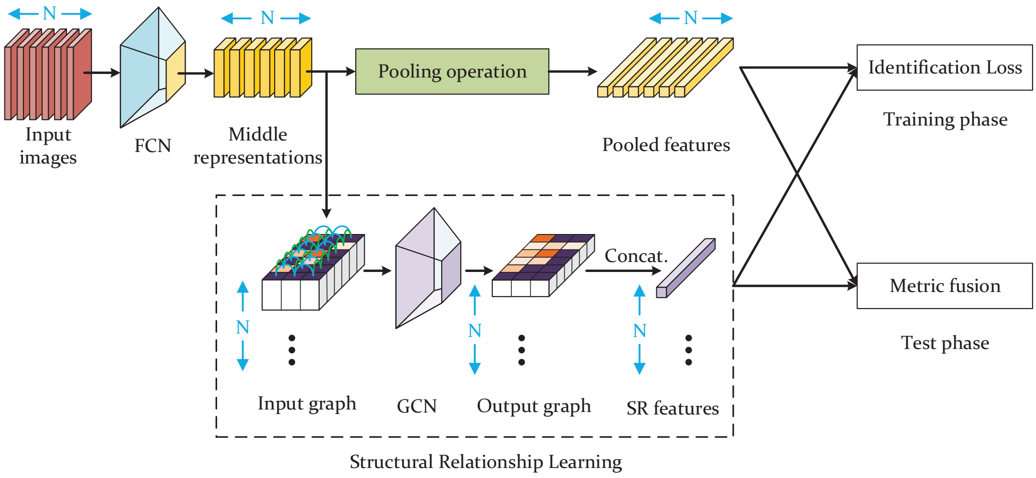

Fig. 3. The end-to-end ReID pipeline with SRL and pooling coupled. The input mini-batch will first be processed by the Fully Convolutional Network (FCN) to generate middle representations, which are further fed into two separate branches to extract the pooled features and structural relationship (SR) features. Identification loss and metric fusion are used in training and test phases respectively.

Where $\mathbf{Z} \in \mathbb{R}^{M \times d_3}$ is the output of GCN with the dimension of each feature vector reduced to d_3 and $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$. $\mathbf{W}^{(0)} \in \mathbb{R}^{d_0 \times d_1}$ is the input-to-hidden weight matrix that maps $\mathbf{X}$ to a latent feature space of dimension d_1. $\mathbf{W}^{(1)} \in \mathbb{R}^{d_1 \times d_2}$ is the hidden-to-hidden matrix and $\mathbf{W}^{(2)} \in \mathbb{R}^{d_2 \times d_3}$ is the hidden-to-output matrix.

A simple demonstration of SRL is in Fig. 2. The input graph is defined by adjacent matrix $\mathbf{A}$, which contains the structure information of the whole feature map. Every hidden layer learns to map the high dimensional node features into a lower feature space, and then propagate the information to each other. Thus relationships between features that are helpful to achieve learning objective will be enhanced and preserved. Eventually, all the output features are concatenated to form a spatial relationship feature f_{SRL} that captures the structural relationship information.

2.2. Application in ReID Pipeline

A diagram of our proposed SRL$_P$ ReID is shown in Fig. 3. The architecture is divided into two branches after the middle representation layer, where the top branch is guided by pooling operation to give the pooled features, and the bottom branch is composed of our proposed SRL to learn structural relationship information.

To be specific, given a mini-batch of N images, we first extract the middle representations $S = \{\mathbf{I}_1, \mathbf{I}_2, \cdots, \mathbf{I}_N\}$ by a Fully Convolutional Network (FCN). For the i^{th} image, as is discussed in Section 2.1, the middle representation is a matrix $\mathbf{I}_i \in \mathbb{R}^{C \times H \times W}$. The top branch utilizes pooling operation to compress each channel of $\mathbf{I}_i$ into a single value to give a pooled feature $f_{i,POOL} \in \mathbb{R}^{1 \times C}$. Meanwhile, the bottom branch applies SRL on $\mathbf{I}_i$ to generate a feature $f_{i,SRL}$ that preserves structural relationships. In training phase, like most existing approaches, we adopt the identification loss to supervise both branches. While in test phase, metric fusion is adopted to couple the contribution of both branches.

Consider the pooled features of both *query* set and *gallery* set as $\mathbf{F}^t_{POOL} \in \mathbb{R}^{N \times C}$, where $t \in \{q, g\}$. In our implementation, consine distance is used to measure the similarity between two features. Thus the distance matrix between the *query* set and the *gallery* set is calculated by dot product:

$$\mathbf{DM}_{POOL} = \mathbf{F}^q_{POOL} \cdot \mathbf{F}^{g\top}_{POOL}. \tag{4}$$

Here, the i^{th} row of $\mathbf{DM}_{POOL}$ indicates the cosine distances between the i^{th} pooled feature of *query* set and all the pooled features of *gallery* set. Similarly, the distance matrix of the SRL features is calculated as $\mathbf{DM}_{SRL}$.

The *metric fusion* strategy is used to couple two branches, which is realized by the weighted sum of $\mathbf{DM}_{POOL}$ and $\mathbf{DM}_{SRL}$:

$$\mathbf{DM} = w_{POOL}\mathbf{DM}_{POOL} + (1 - w_{POOL})\mathbf{DM}_{SRL}. \tag{5}$$

Here, w_{POOL} is the weight of $\mathbf{DM}_{POOL}$, representing the importance of the top pooling branch, while $1 - w_{POOL}$ represents the importance of the bottom SRL branch.

3. EXPERIMENTAL RESULTS

3.1. Implementation Details

We use ResNet50 [5] as the FCN to extract middle representations, which is pretrained on the ImageNet [11] dataset and

Table 1. The influence of GCN layers on the performance of SRL branch.

Dataset	GCNs	top-1	top-5	top-10	mAP
iLIDS-VID	1	**82.7**	**98.0**	**100.0**	**89.5**
	2	82.0	96.0	98.0	87.7
	3	73.3	88.7	93.3	80.4
PRID2011	1	**93.3**	**100.0**	**100.0**	**96.2**
	2	88.8	97.8	98.9	93.1
	3	80.9	93.3	96.6	86.3
MARS	1	**84.4**	**93.5**	**95.5**	**75.8**
	2	81.8	92.4	94.1	72.8
	3	80.2	92.3	93.8	70.0

Table 2. SRL verses Pooling.

Dataset	Branch	top-1	top-5	top-10	mAP
iLIDS-VID	Pooling	78.0	96.0	99.3	85.0
	SRL	82.7	98.0	**100.0**	89.5
	SRL_P	**86.7**	**98.7**	**100.0**	**91.5**
PRID2011	Pooling	91.0	97.8	98.9	94.4
	SRL	93.3	**100.0**	**100.0**	96.2
	SRL_P	**95.5**	**100.0**	**100.0**	**97.0**
MARS	Pooling	78.6	90.6	93.0	70.2
	SRL	**84.4**	**93.5**	**95.5**	**75.8**
	SRL_P	**84.4**	**93.5**	**95.5**	**75.8**

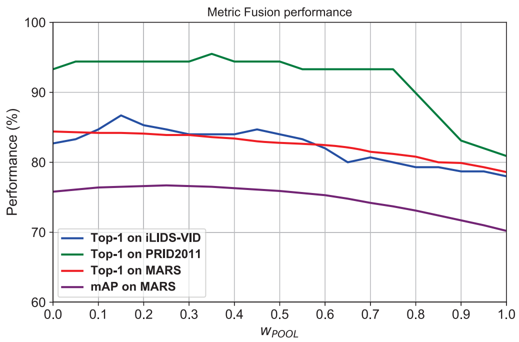

Fig. 4. Main performance plots over w_{POOL}.

then used to initialize the modules except for the SRL branch. All the input images are resized to 288×144. For data augmentation, random cropping and random horizontal flipping are adopted.

We select *Stochastic Gradient Descent* (SGD) as the optimization method. In training phase, we set 1e-3 as the initial learning rate for the pretrained modules and 1e-2 for the non-pretrained modules. All the learning rates will be reduced by 10 times every 3 epoches, in a totol of 9 epoches. A mini-batch for training consists of 8 randomly selected person images belonging to 8 different persons. In test phase, we use all the images of a person to get a single robust feature by averaging learned features.

3.2. Evaluation

Datasets. To evaluate the effectiveness of our proposed approach for ReID, we conduct extensive experiments and ablation study on three popular ReID datasets, iLIDS-VID [12], PRID2011 [13] and MARS [14]. iLIDS-VID dataset contains 600 trajectories for 300 person identities. Due to heavy occlusion, iLIDS-VID is very challenging for ReID task. PRID2011 has 385 trajectories from camera A and 749 trajectories from camera B. Among them, only the first 200 people appear in both cameras. MARS is an extension of the Market1501 [15] dataset. It is the first large scale video based ReID dataset. Since all bounding boxes and tracklets are generated automatically, it contains distractors and each identity may have more than one tracklets.

Metrics. We adopt the *Cumulative Matching Characteristics* (CMC) top-1, top-5, top-10 and top-20 accuracies and *Mean Average Precision* (mAP) as evaluation metrics and strictly follow the original evaluation protocol provided by each dataset.

3.3. Ablation Study

In this section, we investigate the effectiveness of our proposed SRL by conducting a series of comprehensive ablation studies on iLIDS-VID and PRID2011 datasets.

Number of GCN layers. A single-layer GCN can propagate information in the first-order domain of nodes, while multi-layer GCNs can take a larger domain into account, meaning the receptive field of a single node is highly related to the number of GCNs. In this part, we investigate how the number of GCN layers affect the perception of nodes, by analyzing the performance SRL solely.

Table 1 demonstrates the influence of GCN layers on the performance of SRL branch. We find that performance tends to decline with the increase of layers. On the one hand, we owe it to that more layers may increase the risk of underfitting, on the other hand, we think it also indicate that larger node receptive field will weaken the learning of local structural relationships.

Metric fusion. Figure 4 depicts the main performance plots over w_{POOL}, including top-1 accuracies on all the three datasets and mAP for MARS. mAP is specially plotted for MARS because it has more than 2 cameras. Since the metric fusion is the weighted sum of DM_{POOL} and DM_{SRL}, $w_{POOL} = 0$ means only the SRL branch is used, while

Table 3. Comparison on the iLIDS-VID dataset.

Method	top-1	top-5	top-10	top-20
TDL [16]	56.3	87.6	95.6	98.3
CNN+XQDA [14]	53.0	81.4	-	95.1
CNN+RNN [2]	58	84	91	96
QAN [4]	68.0	86.8	95.4	97.4
SDM [17]	60.2	84.7	91.7	95.2
JSTRNN [18]	55.2	86.5	-	97.0
ASTPN [19]	62	86	94	98
RQEN [10]	77.1	93.2	97.7	99.4
DRSA [7]	80.2	-	-	-
SRL_P	**86.7**	**98.7**	**100.0**	**100.0**

Table 4. Comparison on the PRID2011 dataset.

Method	top-1	top-5	top-10	top-20
TDL [16]	56.7	80.0	87.6	93.6
CNN+XQDA [14]	77.3	93.5	-	99.3
CNN+RNN [2]	70	90	95	97
QAN [4]	90.3	98.2	99.3	100.0
SDM [17]	85.2	97.1	98.9	99.6
JSTRNN [18]	79.4	94.4	-	99.3
ASTPN [19]	77	95	99	99
RQEN [10]	91.8	98.4	99.3	99.8
DRSA [7]	93.2	-	-	-
SRL_P	**95.5**	**100.0**	**100.0**	**100.0**

$w_{POOL} = 1$ means the opposite. Obviously, from left to right, there is a slight upward trend and then a downward trend for most curves, which proves that the SRL branch is relatively more important than the Pooling branch. In addition, we find it interesting that the curves for MARS dataset are smoother than others. This is for the reason that lager dataset comes with more stable evaluation results.

SRL verses Pooling. Since each branch only captures part of the information, we now investigate how these two branches work separately and collaboratively. Table 2 is the comparison results, where the SRL uses a single-layer GCN. We can find that SRL digs more comprehensive information contained in features, manifesting unparalleled potential. Note that for the MARS dataset, coupling pooling branch to SRL shows no gain at all, this proves that fully trained SRL is capable of learning complete information from the feature maps, thus is more suitable for feature induction.

3.4. Comprison with State-of-the-art methods

Results on iLIDS-VID dataset. The results of our proposed method and other state-of-the-art methods on the iLIDS-VID dataset are listed in Table 3. Our proposed method outperforms all the compared methods.

The upper half of the table shows methods in which no structural relationships are considered. In the implementation, all the methods use CNNs with different architectures to learn representative deep features. As a contrast, the lower half are methods that use structural relationships to varying degrees. JSTRNN [18] uses spatial pyramid pooling to enable the perception for structural relationships. However, it still ignores the structural relationship of single scale feature map. ASTPN [19] uses spatial recurrent model to capture structural relationships in different directions of consecutive images. The limitation is that regions in spatially convolved features are not very sequential for RNN to learn useful information. RQEN [10] is an extension of QAN [4] that tries to learn structural relationships by simple partitioning. And finally DRSA [7] introduces multiple spatiotemporal attention model to automatically discover that diverse set of distinctive body to address the problem of occlusion, which tries to capture structural relationships by attention mechanism. Nevertheless, it still relies on pooling to extract features. Compared with the above methods, our SRL explicitly uses graph structure to retain the spatial structure of convoluted features, and learns relationships by information propagation over edges, resulting in a superior method.

Table 5. Comparison on the MARS dataset.

Method	mAP	top-1	top-5	top-20
CNN+XQDA [14]	47.6	65.3	82.0	89.0
QAN [4]	51.7	73.7	84.9	91.6
SDM [17]	-	71.2	85.7	91.8
TriNet [20]	67.7	79.8	91.4	-
JSTRNN [18]	50.7	70.6	90.0	**97.6**
DCFBLP [8]	56.1	71.8	86.6	93.0
DRSA [7]	65.8	82.3	-	-
SRL_P	**75.8**	**84.4**	**93.5**	96.8

Results on PRID2011 dataset. Table 4 illustrates the performance of our proposed method and state-of-the-art methods on the PRID2011 dataset. Our proposed method outperforms all the methods introduced above, which verify the effectiveness of our method.

Results on MARS dataset. As the first large scale video-based ReID dataset, MARS is a more objective and fair evaluation criteria for multi-shot methods. The results of our proposed method and other state-of-the-art methods on the MARS dataset are shown in Table 5. Note that all the experiments are conducted in the single query mode. Our method outperforms all the compared methods by a large margin.

4. CONCLUSION

In this paper, we discover that pooling operation may destroy valuable structural relationship information, leading to defective feature learning in tasks such as person re-identification. We propose a method called *Structural Relationship Learning* (SRL) to capture structural relationships by constructing a spatially structured graph based on the convolved features and propagate information over the edges. Coupled with pooling operation by metric fusion, SRL provides more comprehensive information for identity discrimination.

Acknowledgement This work is partially supported by National Key R&D Program of China (No.2017YFA0700800), Natural Science Foundation of China (NSFC): 61876171 and 61572465, and Beijing Municipal Science and Technology Program: Z181100003918012.

5. REFERENCES

[1] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159. 1

[2] Niall Mclaughlin, Jesus Martinez Del Rincon, and Paul Miller, "Recurrent convolutional network for video-based person re-identification," in *ICCV*, 2016, pp. 1325–1334. 1, 5

[3] Dahjung Chung, Khalid Tahboub, and Edward J Delp, "A two stream siamese convolutional neural network for person re-identification," in *ICCV*, 2017. 1

[4] Yu Liu, Junjie Yan, and Wanli Ouyang, "Quality aware network for set to set recognition," in *CVPR*, 2017, vol. 2, p. 8. 1, 5

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 1, 3

[6] Jianlou Si Si, Honggang Zhang, Chunguang Li Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," *CoRR*, vol. abs/1803.09937, 2018. 1, 2

[7] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *CVPR*, 2018, pp. 369–378. 1, 2, 5

[8] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, 2017, pp. 384–393. 2, 5

[9] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016. 2

[10] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai, "Region-based quality estimation network for large-scale person re-identification," *CoRR*, vol. abs/1711.08766, 2017. 2, 5

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255. 3

[12] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, "Person re-identification by video ranking," in *ECCV*. Springer, 2014, pp. 688–703. 4

[13] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof, "Person Re-Identification by Descriptive and Discriminative Classification," in *SCIA*, 2011. 4

[14] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016. 4, 5

[15] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015. 4

[16] Jinjie You, Ancong Wu, Xiang Li, and Wei Shi Zheng, "Top-push video-based person re-identification," in *ICCV*, 2016, pp. 1345–1353. 5

[17] Jianfu Zhang, Naiyan Wang, and Liqing Zhang, "Multi-shot pedestrian re-identification via sequential decision making," *arXiv preprint arXiv:1712.07257*, 2017. 5

[18] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *CVPR*, 2017, pp. 6776–6785. 5

[19] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," *arXiv preprint arXiv:1708.02286*, 2017. 5

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017. 5