MULTI-LEVEL RELATION AWARE NETWORK FOR PERSON RE-IDENTIFICATION

Jing Yang[1] *Canlong Zhang*[1] ⋆ *Zhixin Li*[1] *Yanping Tang*[2]

[1] Guangxi Key Lab of Multi-source Information Mining Security,
Guangxi Normal University, Guilin 541004,China
[2] School of Computer Science and Information Security,
Guilin University of Electronic Technology, Guilin 541004,China

ABSTRACT

Person attribute or pose information has improved person re-identification performance, however, inaccurate pose or attribute module will damage the final identification performance. Based on this, we propose a multi-scale relation aware network (MSRA) for person re-identification. Specifically, we design an attribute relation mining module to construct an attribute map through constraint loss to learn the correlation between different attributes. Besides, we construct a multi-level Pose Pyramid based on the physical structure of the human body, so as to model the internal relationship between pose points. Finally, we designed a cross-scale graph convolution to infer the cooperative structural relation between different layers of components and fused it with the attribute relation module to reinforce the feature. Many experiments on three large-scale datasets verify the effectiveness and state-of-the-art performance of the proposed method.

Index Terms— Person Re-identification, Attribute Relation, Pose Pyramid, Structural Relation

1. INTRODUCTION

The purpose of person re-identification (Re-ID) is to identify a specific person from non-overlapping cameras with different perspectives and positions. Some early person Re-ID works [1, 2, 3] mainly used low-level visual features(e.g. colour, texture) as feature descriptors. Such features are more sensitive to external interference. In contrast, high-level attributes (e.g. age, gender, hair length, clothing style) are relatively stable.

Recently, many works[4, 5, 6] have also separately extracted attribute features as important feature expressions. Tay et al.[7] combine body parts with attribute information to form an attribute attention module. Li et al.[8] propose that semantic attribute information should be used as auxiliary information for local matching. In fact, not all attributes are equally important, so the relationship mining and selection of different attributes is helpful for learning robust features. We use the graph convolution network(GCN) to mine the deep high-order relationships among different attribute features.

⋆Corresponding Author: zcltyp@163.com

Some researchers [9, 10, 11, 12] try to use the pose estimator to detect the key points of persons to align and match local pose regions. For example, in some work based on probabilistic graph models [13, 14] to learn the typical spatial relationship between key points, Miao et al.[15] propose eliminating occlusion regions by setting the response threshold of the key points. Yang et al.[16] propose a spatial-temporal graph convolution network (STGAN), which models the temporal relationship between adjacent frames and the spatial structure relationship within each frame and alleviates the occlusion problem in video matching by mining the complementary spatial-temporal information. However, these strategies rely too much on high-precision pose estimation.

Specifically, we propose a multi-scale relation aware network (MSRA) for person Re-ID. Considering the dependency between attributes, we take each learned attribute feature as a node in the graph and use a linear graph convolution to mine the attribute relation. The Pose Pyramid to learn cross-scale discriminant semantic features to alleviate the impact of inaccurate pose estimation. The attribute relation mining module and Pose Pyramid can learn complementary effective semantic features, and the final fusion features have stronger robustness. The main contributions of this paper are as follows:

(1) We propose a new multi-scale relation aware network(MSRA) end-to-end person Re-ID model, which combines the appearance attribute information and internal pose features of persons and eliminates the effect of similar appearances or similar pose on the identification.

(2) We design a new attribute relation mining module (ARM), which uses the relationship between the attribute features and the features learned by the constraint loss to construct the graph to obtain more robust relationship-aware attribute features to better re-identify.

(3) We design a Pose Pyramid, which dynamically convolves to mine the relationship among the nodes of the layers. The cross-layer graph computing unit uses multi-scale graphs in a coarse-to-fine manner to extract and fuse the local component features at multiple scales. Then, the fused attribute information is used as a guide to filter out useless information and optimize the representation of local features.

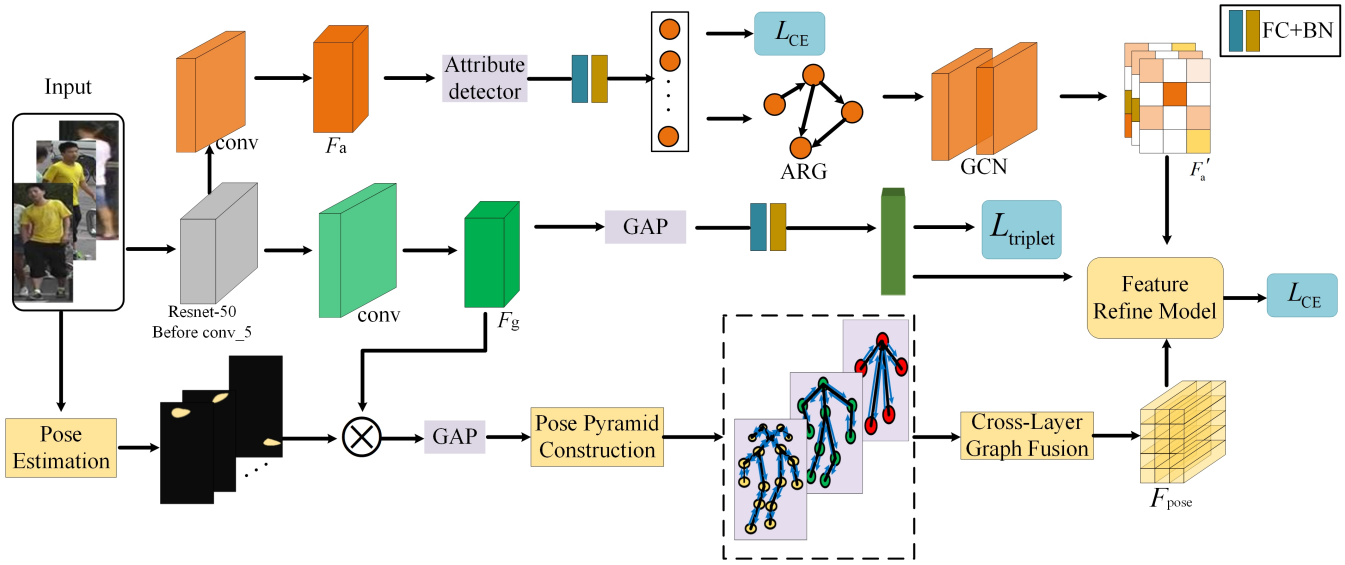


Fig. 1. The structure of proposed MSRA.

2. METHODOLOGY

2.1. Overview

We define the attribute labels of an input person image are $Z = \{z^1, z^2, ..., z^K\}$, and K is the number of predefined attribute labels. We divide the top branch of Fig.1 into two parts and modifie the baseline model ResNet-50 [17] to make it more suitable for their respective tasks. Specifically, the two parts of the top branch share the first four layers of the ResNet-50. In the gloal feature branch, the last 7×7 pooling is changed to global average pooling(GAP). The output features containing global features before pooling are expressed as $\boldsymbol{F}_g$. In the attribute relation mining branch, we remove the spatial downsampling layer,and the attribute features of the network output are expressed as $\boldsymbol{F}_a$.

2.2. Attribute relation mining module

In ARM, we use the attribute detector to process the global attribute feature $\boldsymbol{F}_a$ as shown in Formula 1, thus generating an attribute mask for each attribute:

$$\boldsymbol{M}_k = Sigmoid(\boldsymbol{W}_1\boldsymbol{F}_a + \boldsymbol{b}) \tag{1}$$

Each position of attribute feature $\boldsymbol{F}_a$ is averagely pooled. Then, we perform weighted average pooling with the M_k on feature map F_a, and get the attribute feature $\boldsymbol{F}_a{}^{(k)} \in \mathbb{R}^c$, $\boldsymbol{F}_a{}^{(k)} = \frac{1}{h\times w}\sum_{(x,y)}[F_a(x,y) \odot M_k(x,y)]$.

The individual attribute losses of K attributes are calculated first, and K cross entropy loss functions are used here. After sending attribute $\boldsymbol{F}_a{}^{(k)}$ to the fully connected(FC) layer, it outputs $\boldsymbol{l}^k$, we use $\boldsymbol{l}^k$ as the graph nodes to construct an attribute relationship graph (ARG). The GCN takes K predicted attribute nodes $\tilde{z} = \{l^k\}_{k=1}^K$ and the initialized attribute relationship adjacency matrix $\boldsymbol{S}_a$ as the input to construct a linear graph convolution layer:

$$\boldsymbol{A} = ReLU(\boldsymbol{D}^{-1/2}\boldsymbol{S}_a\boldsymbol{D}^{-1/2}\tilde{\boldsymbol{z}}\boldsymbol{W}_2) \tag{2}$$

where $\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$ is standardized by an adjacency matrix, $\boldsymbol{D} = diag(d_1, d_2, ..., d_k)$. The elements of each row in $\boldsymbol{D}$ are equal to the sum of the elements in the corresponding rows of $\boldsymbol{S}_a$. The relation adjacency matrix between pairs of attribute nodes is $\boldsymbol{S}_a\left(\boldsymbol{l}^k, \boldsymbol{l}^i\right) = \phi(\boldsymbol{l}^k)^T\phi'(\boldsymbol{l}^i)$. Where ϕ and ϕ' are embedding functions. Finally, the final output of the attribute relationship mining module is $\boldsymbol{F}_a{}' = FC(\boldsymbol{A})$.

2.3. Pose Pyramid module

The purpose of Pose Pyramid is to divide the human body regions into different granularities, and use GCN to mine the semantic information of different components. Firstly, we use HRNet[18] to extract the key points. The upper, middle and lower three layers of the Pose Pyramid contain 17, 10 and 5 nodes n_j, of which $j = \{1, 2, 3\}$. We take n_j node features $R_j = \{\boldsymbol{r}_j^t\}_{t=1}^{n_j}$ of the j-th layer as the input. The node features of each layer are obtained by pooling the product of global features and heat map according to the number of predefined nodes. We design a GCN within a layer to transfer information between nodes. Taking the first layer GCN layer as an example, the structure information dissemination and update process of graph nodes in the layer is expressed as follows:

$$\boldsymbol{r}_p^1 = \sigma(\sum_{q\in n_1}\boldsymbol{S}_{p,q}^1\boldsymbol{W}_5\boldsymbol{r}_q^1) \tag{3}$$

where $\boldsymbol{r}_p^1$ and $\boldsymbol{r}_q^1$ are the graph nodes of the first layer, and $p, q \in [1, n_1]$. $\boldsymbol{S}_{p,q}^1$ is the structural relation matrix constructed

by graph node p and other nodes in the first layer.

$$S^1_{p,q} = \frac{\exp(LeakyReLU(W_7{}^T[W_6 r^1_p]\|[W_6 r^1_q]))}{\sum_{t\in n_1}\exp(LeakyReLU(W_7^T[W_6 r^1_p]\|[W_6 r^1_t]))} \quad (4)$$

where $\|$ is the connection feature of r^1_p and r^1_q, and LeakyReLU is a nonlinear activation function. If viewing original semantics node features $R_i \in \mathbb{R}^{n_i \times c}$ as one-order features, then the output node features $\tilde{R}_i \in \mathbb{R}^{n_i \times c}$ after the above node features updating can be regarded as high-order features with spatial relation information.

To realize the interlayer diffusion of structural relations, we design a cross-layer graph convolution fusion strategy(CLG). First, we derive the cross-layer adjacency matrix $\tilde{S}_{v,v+1} \in \mathbb{R}^{n_v \times n_{v+1}}$, $v \in \{1, 2\}$. For each layer of nodes, feature aggregation and enhancement are performed first. Let's take the first layer of graph nodes r^1_p and r^1_q as an example:

$$B^1_p = \sum_{q=1}^{n_1} f_1([\omega(r^1_p), \sigma(r^1_q - r^1_p)]) \quad (5)$$

$$V^1_p = \delta([r^1_p, B^1_p]) \quad (6)$$

where $[\cdot,\cdot]$ is the concat operation. f_1, ω, σ and δ are all learnable embedding functions, and all of them include a fully connected layer, a normalization layer, and a ReLU activation function layer. Then, the dependency matrix between adjacent layers is calculated by $\tilde{S}_{v,v+1} = soft\max[(V^v)^T(V^{v+1})]$.

In order to adaptively focus on the key related features in pose components at different spatial levels and enhance the global feature learning, we calculate the feature $\tilde{R}_{v,v+1}$ after cross-layer graph convolution fusion according to the dependency matrix.

$$\tilde{R}_{v,v+1} \leftarrow \alpha(\tilde{S}_{v,v+1}(r^v)W_{v,v+1}) \quad (7)$$

where $W_{v,v+1} \in \mathbb{R}^{c\times c}$ and α are learnable parameters, the fusion pose feature $\alpha(\tilde{S}_{v,v+1}[r_v]W_{v,v+1})$ of the two layers can adaptively absorb information clues from the corresponding parts of the body. Finally, the multi-layer graph feature expression F_{pose} of Pose Pyramid is obtained:

$$F_{pose} = \tilde{R}_1 + \lambda(\sum_{v=1}^{2}\tilde{R}_{v,v+1}) \quad (8)$$

2.4. Feature Enhancement Module

Attributes can be used as supplementary information of pose features and enhance the robustness of features. We design a feature enhancement module, and the details are shown in Fig.2. The output features after fusion are as follows:

$$F_{fuse} = Sigmoid[W_{10}\tanh(W_8 F_{pose} + W_9 F_a{}' + b')] \odot F_{pose} \quad (9)$$

After strengthening the feature module, F_{fuse} and global features are linked together. $F_{\text{out}} = F_{\text{fuse}} \oplus F_{\text{g}}$, where F_{out} contains three different types of information: global identity level features, refined posture features and attribute features.

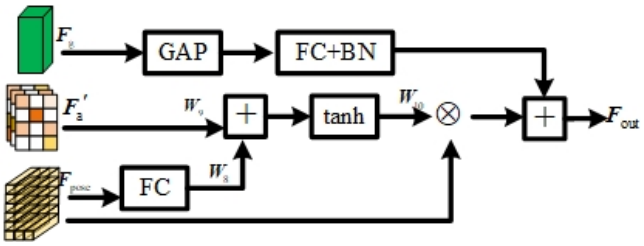


Fig. 2. Structure diagram of the feature enhancement module." + " is element by element addition.

3. EXPERIMENT

3.1. Experimental details

We have performed experiments on two datasets: Market-1501[19] and DukeMTMC-ReID[20]. All attribute labels in the training process are from Lin et al.[5]. For evaluation, we use the cumulative match characteristic (CMC) and mean average precision (mAP) as the evaluation metrics.

In the training process, all input images are resized to 256×128. All training batches are 32, and the number of training iterations of the network model is 240. Adam is used to optimize the model of the entire network in small batches. The initial learning rate of the training process is set to 0.01, and in the last 60 batches, the model learning rate is adjusted to 0.001. The balance coefficient λ is set to 0.6. For global features F_{g}, we use triple loss function to optimize. For fusion features F_{out}, we use cross entropy loss function to optimize. For the total loss of K attributes, global feature loss and fusion feature loss, their weight coefficients are set as 0.3, 1, 1.2, respectively.

3.2. Comparison with the state-of-the-art

On the Market-1501 and DukeMTMC-ReID datasets, we set up three groups of experiments: only using person attribute features[7, 21, 4], only using person pose features[14, 22, 10] and using global features[23, 24, 21] (such as attention mechanisms and GCN). On the Market-1501 dataset, among the three comparative experimental models, the models with the best mAP accuracy are RAN[4], DSA[10] and SCSN[23], and their mAP accuracies are 89.2%, 87.6% and 88.5%, respectively. Although these three models use different features, the accuracy is not much different, which also explains pedestrians from the side. In all comparative experiments, our MSRA model is superior to the existing models. On the DukeMTMC-ReID dataset, the accuracy of our model Rank-1 reaches 92.2%, which is 1.2% higher than the existing optimal model. The results show that MSRA in Rank-1 has significantly better accuracy and mAP than all other methods, and these comparisons further prove the effectiveness of this model.

Table 1. Performance(%) comparison with the state-to-the-arts model on market-1501 and DuekmtMTMC-reID.

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
AANet[7]	93.9	82.4	86.4	72.6
AAB[6]	96.1	88.6	89.9	80.4
RAN[4]	95.2	89.2	88.9	79.1
HOReID[14]	94.2	84.9	86.9	75.6
PSE[22]	90.4	80.5	82.8	74.5
DAS[10]	95.7	87.6	86.2	74.3
SCSN[23]	95.7	88.5	91.0	79.0
Pyramid[24]	95.7	88.2	89.0	79.0
PCB+RPP[21]	93.8	81.6	83.3	69.2
MSRA	**95.9**	**90.6**	**92.2**	**81.6**

Table 2. The impact of different CLG Settings on performance(%).

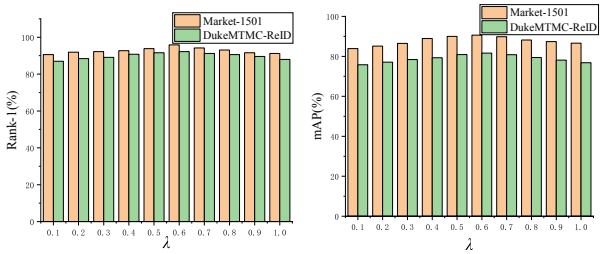
Methods		Market-1501		DukeMTMC-ReID	
		Rank-1	mAP	Rank-1	mAP
$\theta = 1$	$CLG-j_1$	94.3	89.6	91.3	80.7
	$CLG-j_1j_2$	95.0	90.2	91.8	81.2
	$CLG-j_1j_2j_3$	**95.9**	**90.6**	**92.2**	**81.6**
$\theta = 2$	$CLG-j_1$	93.6	88.7	90.7	79.3
	$CLG-j_1j_2$	94.2	89.2	91.2	79.9
	$CLG-j_1j_2j_3$	94.8	89.8	91.5	80.4

3.3. Ablation experiments

In Table 2, we analyse the influence of number α of cross-layer convolution CLGs on the experimental performance and analyse the performance of fusing pose features between different layers. The experimental results are shown in Table 2. When $\theta = 2$, two layers of CLG are used when cross-layer features are fused. Compared with $\theta = 1$, the performance of both datasets significantly decreases, which reflects that although double-layer CLG can fuse more fine-grained information, it contains considerable redundant information, which weakens the final feature expression. Moreover, $CLG-j_1j_2j_3$ contains more local features of the granularity level. Hence, its effect is better than $CLG-j_1j_2$ and $CLG-j_1$, which proves that CLG can effectively mine features of different scales and fuse their complementary information.

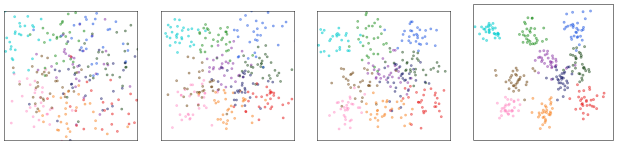
Super parameter λ in Formula8 controls the final fusion of the three-layer features of the Pose Pyramid. Fig.3 shows the influence of λ on identification accuracy on different datasets. When λ is small, the pose key points fuse less than other local features with different granularities. The performance is poor, which proves the necessity of blocking the features extracted by a convolutional neural network. When λ gradually increases, the performance is improved. When $\lambda = 0.6$, the performance of both datasets is the best. When $\lambda > 0.6$, the fused features contain too much redundant information, and the performance declines.

As shown in Fig.4, we use t-SNE[25] to intuitively visualize their distribution. In Fig.4(a), the samples of different



(a) (b)

Fig. 3. The Influence of balance coefficient λ of pose feature fusion in different layers of Pose Pyramid on performance.



(a) baseline (b) ARM (c) Pose Pyramid (d) MSRA

Fig. 4. Visualization results of 10 randomly selected samples in the Market-1501 test set.

people are mixed and difficult to distinguish. After adding attribute features and pose features, respectively, the cluster of the same kind is closer than the baseline type, which verifies the experimental results of this paper. In our model, different clusters maintain a long distance, while clusters of the same kind are more compact. The results are different, which fully verifies that the model in this paper can effectively improve the identification performance.

4. CONCLUSION

To improve pedestrian recognition,we propose a multi-scale relation aware network for person Re-ID. As complementary features, attributes and pose can eliminate the noise caused by detection deviation and increase the granularity of identification. At present, the attribute relationship mining module in this paper also relies on the manual annotation of attributes. In future work, we will consider combining weak supervised learning with attributes or other characteristics of pedestrians to further improve the recognition performance.

5. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China(No.61866004,61966004,61962007), the Guangxi Natural Science Foundation (No.2018GXNSFDA-281009, 2019GXNSFDA245018, 2018GXNSFDA294001), Research Fund of Guangxi Key Lab of Multi-source Information Mining and Security (No.20-A-03-01), and Guangxi "Bagui Scholar" Teams for Innovation and Research Project.

6. REFERENCES

[1] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *CVPR*, 2018, pp. 2275–2284.

[2] Jianwen Wu, Ye Zhao, and Xueliang Liu, "Enhancing person retrieval with joint person detection, attribute learning, and identification," in *PCM*. Springer, 2018, pp. 113–124.

[3] Yu-Tong Cao, Jingya Wang, and Dacheng Tao, "Symbiotic adversarial learning for attribute-based person search," in *ECCV*. Springer, 2020, pp. 230–247.

[4] Yuxuan Shi, Hefei Ling, Lei Wu, Jialie Shen, and Ping Li, "Learning refined attribute-aligned network with attribute selection for person re-identification," *Neurocomputing*, vol. 402, pp. 124–133, 2020.

[5] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[6] Jianfu Zhang, Li Niu, and Liqing Zhang, "Person re-identification with reinforced attribute attention selection," *IEEE Transactions on Image Processing*, vol. 30, pp. 603–616, 2020.

[7] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap, "Aanet: Attribute attention network for person re-identifications," in *CVPR*, 2019, pp. 7134–7143.

[8] Shuzhao Li, Huimin Yu, and Roland Hu, "Attributes-aided part detection and refinement for person re-identification," *Pattern Recognition*, vol. 97, pp. 107016, 2020.

[9] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *CVPR*, 2018.

[10] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen, "Densely semantically aligned person re-identification," in *CVPR*, 2019, pp. 667–676.

[11] Kai Wang, Shichao Dong, Nian Liu, Junhui Yang, Tao Li, and Qinghua Hu, "Pa-net: Learning local features using by pose attention for short-term person re-identification," *Information Sciences*, vol. 565, pp. 196–209, 2021.

[12] J. Yin, A. Wu, and W. S. Zheng, "Fine-grained person re-identification," *International Journal of Computer Vision*, vol. 128, no. 12, 2020.

[13] "Prgcn: Probability prediction with graph convolutional network for person re-identification," *Neurocomputing*, vol. 423, no. 12, pp. 57–70, 2021.

[14] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun, "High-order information matters: Learning relation and topology for occluded person re-identification," in *CVPR*, 2020, pp. 6448–6457.

[15] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *ICCV*, 2019.

[16] "Spatial-temporal graph convolutional network for video-based person re-identification," in *CVPR*, 2020.

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

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019, pp. 5693–5703.

[19] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[20] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*. Springer, 2016, pp. 17–35.

[21] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018, pp. 480–496.

[22] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *CVPR*, 2018, pp. 420–429.

[23] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang, "Salience-guided cascaded suppression network for person re-identification," in *CVPR*, 2020, pp. 3300–3310.

[24] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *CVPR*, 2019, pp. 8514–8522.

[25] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.