Neurocomputing 432 (2021) 206–215

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning multi-granularity features from multi-granularity regions for person re-identification

Kaiwen Yang, Jiwei Yang, Xinmei Tian *

CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System, University of Science and Technology of China, Hefei, Anhui, China

ARTICLE INFO

Article history:
Received 6 July 2020
Revised 25 October 2020
Accepted 5 December 2020
Available online 16 December 2020
Communicated by Steven Hoi

Keywords:
Person re-ID
Multi-granularity feature fusion
Human region localization

ABSTRACT

Part-based methods for person re-identification have been widely studied. In existing part-based methods, although multiple parts are explored, only coarse-grained features of these parts are utilized. Thus, too much fine-grained information is discarded, which limits their ability to extract detailed discriminative features. To tackle this problem, we propose a novel person re-identification network to learn discriminative features across multiple granularities from body regions which are also multi-grained. Specifically, we detect multi-granularity body regions at different stages of a backbone network, and multi-granularity features are learned from body regions with corresponding granularities. To overcome the severe mismatching problem of fine-grained regions and to learn discriminative features, the detection of multi-granularity body regions and the learning of multi-granularity features are jointly optimized. This joint optimization pushes the learned features concentrating on body regions. Moreover, with the body regions well located, the multi-granularity features can be well aligned. Extensive experiments on four popular datasets show that our method is the state-of-the-art in recent years.

1. Introduction

Person re-identification (ReID) aims at identifying specific person from a set of surveillance cameras across time. It plays a significant role in many vision-related applications, e.g., video surveillance, content-based video retrieval, and identification from CCTV cameras. Compared to other computer vision tasks, ReID is of great challenge due to differences of background, deviations in shape, and occlusion of the subjects [1,2].

Image representation learning plays a crucial role in person ReID. As shown in Fig. 1(a), images are usually fed to deep convolutional neural networks (CNNs) to extract the final representation. However, the final features are often too coarse and lose too much detail information. To solve this problem, many part-based models have been proposed [2–5]. By learning discriminative local features as a complement to global features, they can extract additional rich features and thus achieve better ReID performance.

According to the local region generation way, part-based models can be divided into three categories: pose-based, attention-based, and stripe-based. In pose-based methods, prior knowledge, e.g., pose estimation or human segmentation, is used to locate local regions of a human body accurately [6,2,7–9]. These methods handle local regions of a human body by extra convolutional branches. The attention-based methods learn attention masks to select a focused foreground [10–12]. In stripe-based category, the feature maps are split into several predefined horizontal stripes [4,13–15]. However, they all perform ReID with the features from the last layer, which have coarse granularity and contain limited local information. Furthermore, these methods are based on the assumption that person images are well aligned, so the corresponding stripes can be matched. However, misalignment is very common in person ReID.

However, methods in all three categories have one common drawback: though multiple parts/regions are explored, only coarse-grained features of these parts are utilized, as shown in Fig. 1(b). The local regions are first cropped either at the input [6] or at different stages in backbone CNNs [2], and are then fed into convolutional branches afterwards, leading to that the final features of these regions are coarse-grained. This limits the diversity and discrimination of the final features.

To tackle this problem, we propose to learn multi-granularity features from multi-granularity body regions for person ReID. We detect local regions across multiple granularities at different stages of a backbone network. As shown in Fig. 1(c), we detect four fine-grained body parts in the first stage, two body parts in the second stage, the whole body region in the third stage, and the whole

* Corresponding author.
E-mail addresses: kwyang@mail.ustc.edu.cn (K. Yang), yjiwei@mail.ustc.edu.cn (J. Yang), xinmei@ustc.edu.cn (X. Tian).

https://doi.org/10.1016/j.neucom.2020.12.016
0925-2312/

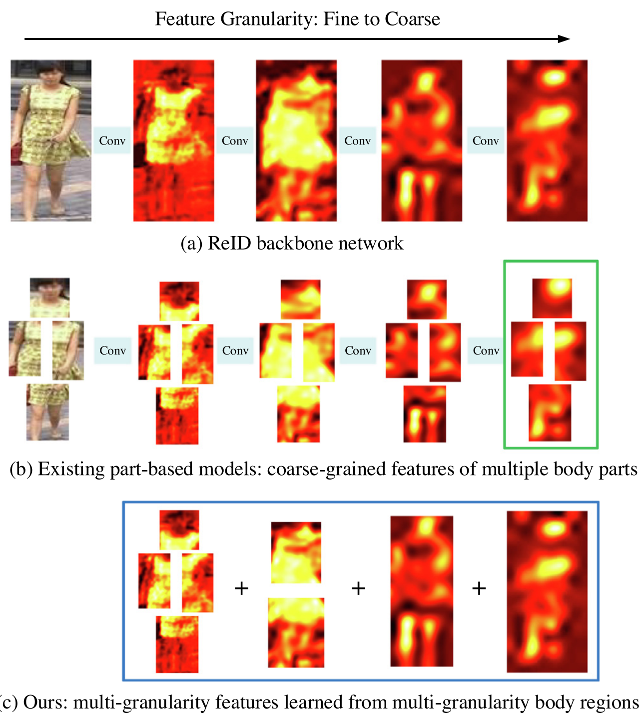


Fig. 1. (a): A backbone convolutional network for person ReID. (b): The green bounding box marks the coarse-grained features used in existing part-based models. (c): The blue bounding box marks the multi-granularity features learned from multi-granularity body regions in our model.

image in the forth stage. For regions in each granularity, instead of feeding them to extra local branches afterwards, we directly apply a feature extraction module to learn corresponding features. In this manner, the original granularities of features are kept, and more detail information are retained. Finally, multi-granularity features learned from multi-granularity body regions are fused for person ReID. Fig. 1 clearly illustrates the difference between our model and existing part-based models.

It is worth noting that we do not simply conduct the multi-granularity region detection and multi-granularity feature learning in a straightforward two-stage manner. There are two challenging problems in our task. First, when fine-grained features are used, the misalignment becomes a big problem, especially for fine-grained regions, since the fine-grained features are extracted from shallow layers in which the receptive field is small and thus are very sensitive to translation, pose variations, etc. This may be the reason why current works only use coarse-grained features for all parts. Second, the fine-grained features are very sensitive to noises or other image content which are not helpful for ReID. Thus, we face the problem of how to ensure the extracted fine-grained features are discriminative for person ReID. To tackle these problems, we design a model to jointly optimize the multi-granularity region detection and multi-granularity feature learning.

In summary, the contributions of this paper are threefold:

- We learn features across multiple granularities from the backbone network without feeding them to extra local branches. In this manner, the final features are diverse: both fine-grained features with rich details and coarse-grained abstract features are well reserved.
- Our multi-granularity features are learned from multi-granularity parts. The location of multi-granularity parts and the learning of multi-granularity features are jointly optimized. This joint optimization pushes the learned features focusing on human body regions. Moreover, with the multi-granularity parts accurately located, the multi-granularity features can be well aligned, and the background noise can be well reduced.
- The proposed method achieves the best performance on four person ReID datasets. Extensive experiments on these datasets verify the effectiveness of our approach. MGRe achieves 90.1%/96.2% mAP/Rank-1 in Market1501 and 82.0%/91.3% mAP/Rank-1 in DukeMTMC-reID.

2. Related work

Regarding discriminative feature learning for person ReID, many methods have been proposed to enhance certain regions in the feature maps. According to the region generation way, these methods can be divided into three categories: pose-based, stripe-based and attention-based.

2.1. Pose-based models

Pose-based models utilize prior pose information with the help of extra pose estimation models or human segmentation models. Zhao et al. first locate local body regions based on pose estimation models and then fuse these local regions hierarchically [2]. Li et al. extract deformable parts using spatial transform networks (STN) [16] based on defined spatial constraints of body parts. In addition, they proposed a gesture-based feature weighted sub-network to learn the weight of features and then selectively fuse features. In order to solve the problem of misalignment and posture shift of person data, Su et al. proposed a two-stream deep convolutional network: one for global features and the other for local features [6]. This method enhances the feature representation of body parts obviously. Saquib et al. utilized pose information in a rather straightforward way [17]. They added an additional input channel for each of the 14 main body keypoints, pushing the network to learn posture information by itself. Meanwhile, they added a branch to let the network to learn viewpoint information. Ustinova et al. noticed that the last stages of the original Bilinear-CNN architecture completely removes the geometric information from consideration by performing orderless pooling [18]. They achieved a better embedding by performing bilinear pooling in a more local way, where each pooling is confined to a predefined human body region. To address the issue of occlusion, Gao et al. [19] proposed a Pose-guided Visible Part Matching (PVPM) method that jointly learns the discriminative features with pose-guided attention and self-mines the part visibility in an end-to-end framework. Zhao et al. utilized composite models to extract specific salience features from different parts of the human body in an unsupervised way [20]. Wang et al. solved the issue of occlusion in a more explicit way[21]. They first used a CNN backbone and a key-points estimation model to extract semantic local features. Then, local features of an image were viewed as nodes of a graph and an adaptive direction graph convolutional (ADGC) layer was proposed to pass relation information between nodes. When aligning two groups of local features from two images, they viewed it as a graph matching problem.

2.2. Attention-based models

Attention-based models aims at eliminating the effects of background differences by learning attention masks to select a focused foreground. Li et al. propose a harmonious attention model to integrate soft pixel attention and hard regional attention [10]. Multiple attention masks are produced in multiple stages and merged with the final global features. An attribute attention network is proposed in [11]. This method generates attention masks according to attribute classification [11]. It learns attribute features with ReID features in a unified learning framework. Hou et al. proposed interaction-and-aggregation network (IANet) that learns the attention mask by modeling geometric variations [22]. In IANet, features are robust to body pose and scale variations. Chen et al. observed that previously learned salient features may hinder the network from learning other important information [23]. Thus, they introduced a cascaded suppression strategy, which enables the network to mine diverse potential useful features that be masked by the other salient features stage-by-stage. These previous approaches mainly learn attention using local convolutions, ignoring the mining of knowledge from global structure patterns. To address such issue, Chen et al. proposed an effective Relation-Aware Global Attention (RGA) module which captures the global structural information for better attention learning [24]. For the purpose of introducing additional contextual information, Yang et al. stacked numbers of convolutional layers in their encoder-decoder style attention module to achieve larger receptive fields [25]. However, most of the previous attention-based works concentrated only on coarse or first-order attention design, e.g. spatial and channels attention, while rarely exploring higher-order attention mechanism. To solve this problem, Chen et al. [26] and Xia et al. [27] utlized high-order attention information to produce the discriminative attention proposals. Chen et al. [26] proposed the High-Order Attention (HOA) module to capture the subtle differences among pedestrians. Xia et al. [27] proposed a novel attention mechanism to directly model long-range relationships via second-order feature statistics.

2.3. Stripe-based models

Stripe-based models split an image into predefined patches. Sun et al. [4] propose a simple but effective approach: part-based convolutional baseline (PCB). They cut the high-level feature map into six stripes evenly and the re-assign outliers in each part using a refined part pooling methods. Inspired by PCB, Wang et al. [14] proposed a multiple granularity network (MGN). MGN has three branches on top of the network: one branch is for global feature and the rest two branches are for local representations for person re-identification. However, MGN only splits the final coarse-grained feature maps into stripes. Based on PCB, Zheng et al. further integrate the gradual cues between local and global information through pyramidal branches [15].

3. Proposed method

In previous part-based methods, although multiple parts are utilized, only the most coarse-grained features of these parts are used to represent an image for ReID, i.e., the outputs of the last convolution layer of the feature extraction net. These coarse features are highly abstract and robust, but they discard too much detailed information. In this paper, we propose a novel method termed multiple granularities ReID (MGRe). As shown in Fig. 2, features and regions with fine-to-coarse granularities are jointly learned.

3.1. Network architecture

As shown in Fig. 2, our backbone is a ResNet-50 with four residual blocks. We locate multi-granularity body regions and extract multi-granularity features from the output feature maps of these four residual blocks in a fine-to-coarse manner via joint optimization.

For each of the first three residual blocks, its output feature maps are fed into a location module. The location module aims to locate body regions with different granularities. Four fine-grained body regions (head, left half of the torso, right half of the torso, and legs) are detected after residual block-1, two coarser body regions (upper and lower body) are detected after residual block-2, and the whole body region is detected after residual block-3. The whole image is utilized after residual block-4; thus, no location module is needed in this stage.

After the local regions are detected, a sampling module is applied to sample corresponding feature maps for each region. The sampled feature maps of all regions in each stage are concatenated via embedding operations and then used for person re-identification. The body region location and person re-identification are jointly trained. The final loss is the sum of three location loss terms and four ReID loss terms.

Since the fine-grained feature maps of the first two residual blocks lack semantic information and translation invariance, we enhance them by adding high-level coarse-grained feature maps. As shown in Fig. 2, we perform bilinear upsampling and convolution operations on the most coarse-grained feature maps output by residual block-4 to ensure their size and number of channels are the same as those of the two fine-grained feature maps. Then, they are added up in an element-wise manner. In this way, the two fine-grained feature maps are semantically guided.

Location Module. The location module automatically detects different body regions. The location of each body region is specified by 4 independent parameters: $\left(\Delta C_x, \Delta C_y, l_1, l_2\right)$. $(\Delta C_x, \Delta C_y)$ are the offsets between the predicted center and the predefined center of the body region. The predefined centers of body regions are set according to human geometry, as reported in Table 1. l_1 and l_2 are the width and height of the region respectively.

Given a feature map, we apply a block with convolution, ReLU, batch normalization, max pooling and fully connected layers to infer the following parameters:

$$\left(\tanh^{-1}(\Delta \boldsymbol{C_x}), \tanh^{-1}(\Delta \boldsymbol{C_y}), \log \boldsymbol{l_1}, \log \boldsymbol{l_2}\right) = \mathbf{L}(\mathscr{G}), \tag{1}$$

where $\mathbf{L}$ denotes the location module and $\mathscr{G}$ is the input feature map. $(\Delta \boldsymbol{C_x}, \Delta \boldsymbol{C_y})$ are scaled to $(-1, 1)$. The width and height are output on log scales to ensure positivity. All of these predicated parameters are vectors because multiple body regions might be located in one location module.

For location loss, we use pseudo labels predicted by a pose estimation model [28]. The pose estimation is not needed in testing. The joint optimization of location and ReID pushes learned features focusing on human body regions. Given an image, the coordinates of 17 human joint points are predicted. These 17 joint points are denoted as $\mathscr{I}$, and their horizontal and vertical coordinates are denoted as $\mathscr{X} = \{x_i, i \in \mathscr{I}\}$and $\mathscr{Y} = \{y_i, i \in \mathscr{I}\}$, respectively. Among these 17 joint points, points that belong to body region p constitute the set $\mathscr{I}_p$. Then, the horizontal and vertical coordinates of these joints are denoted as $\mathscr{X}_p = \{x_i, i \in \mathscr{I}_p\}$and $\mathscr{Y}_p = \{y_i, i \in \mathscr{I}_p\}$, respectively. As is shown in Eq. (1), the location module predicts the offsets of the center point and the length of the body regions. Then, the four corners of the predicted region can be determined:

$$\begin{cases} x_{\min} = (c_x + \Delta c_x) \times W - l_1/2 \\ x_{\max} = (c_x + \Delta c_x) \times W + l_1/2 \\ y_{\min} = (c_y + \Delta c_y) \times H - l_2/2 \\ y_{\max} = (c_y + \Delta c_y) \times H + l_2/2 \end{cases} \tag{2}$$

where W and H are the width and height of the input feature map respectively. Then the location loss can be written as the square of the difference between the locations predicted by our location module and HRNet:

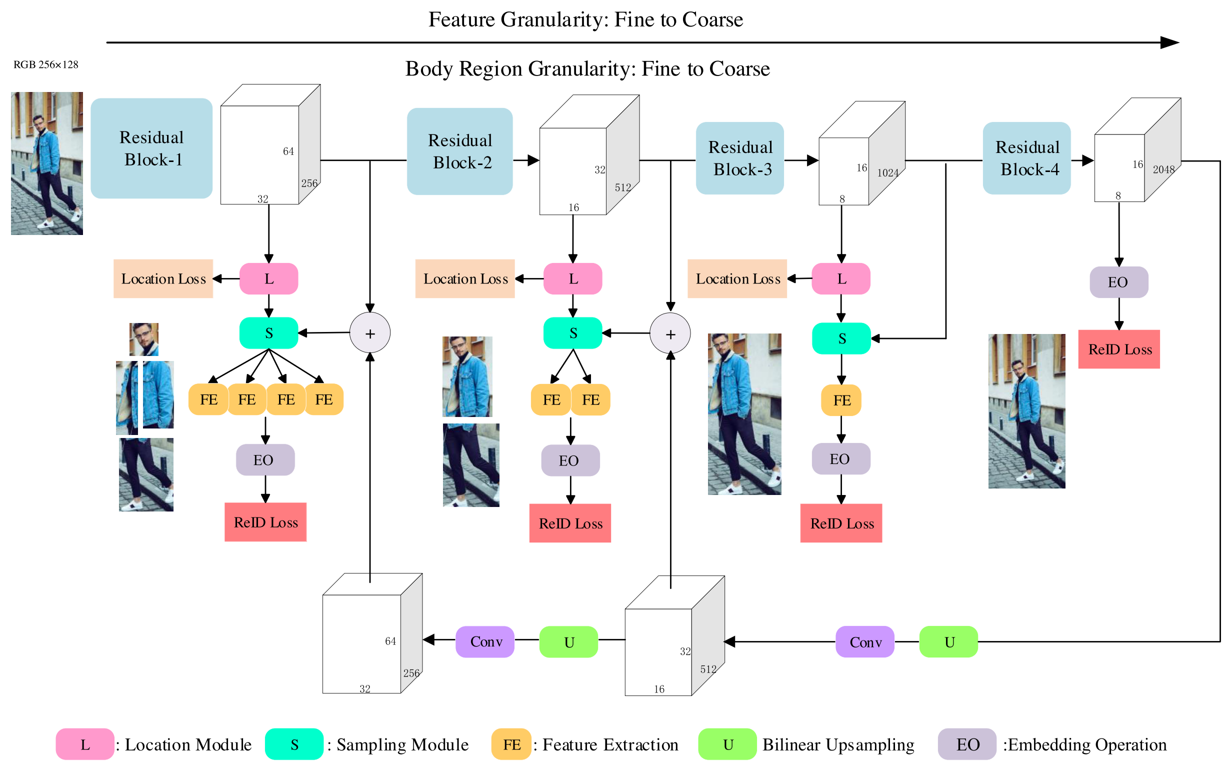


Fig. 2. The architecture of our proposed Multiple Granularities ReID (MGRe). The backbone network is a ResNet-50 that has four residual blocks. We extract features in four stages, i.e., the output feature maps of four residual blocks. In addition, two fine-grained feature maps in the backbone are fused with the upsampled feature maps. (Example image copyright Mihai Stefan (CC0 license)).

Table 1
Prior settings of body regions. C_x and C_y are the predefined centers, which are scaled to $(0, 1)$. N and M are the height and width of sampled feature maps, respectively.

Region	Stage	C_x	C_y	N	M
head	1	1/2	1/12	8	8
half left torso	1	1/2	1/3	32	16
half right torso	1	1/2	1/3	32	16
legs	1	1/2	3/4	24	32
upper body	2	1/2	1/3	16	16
lower body	2	1/2	3/4	16	16
whole body	3	1/2	1/2	16	8

$$\begin{aligned} L_{loc} = \; & \left[x_{min} - min(\mathcal{X}_p) + 0.05\right]^2 \\ & + \left[x_{max} - max(\mathcal{X}_p) - 0.05\right]^2 \\ & + \left[y_{min} - min(\mathcal{Y}_p) + 0.05\right]^2 \\ & + \left[y_{max} - max(\mathcal{Y}_p) - 0.05\right]^2, \end{aligned} \tag{3}$$

where 0.05 is the boundary margin.

Sampling Module. The sampling module samples corresponding feature maps for each region. It has two inputs: location parameters and a feature map. The sampling module samples the corresponding region in the given feature map to a partial feature map with a specific size, which are shown in Table 1.

Inspired by STN [29], the differentiable sampling module is designed. The transformation between input coordinates and output coordinates according to the four location parameters is calculated as

$$\begin{cases} x_i^s = (c_x + \Delta c_x) \times W + (x_i^t - M/2) \times l_1/M, \\ y_i^s = (c_y + \Delta c_y) \times H + (y_i^t - N/2) \times l_2/N, \end{cases} \tag{4}$$

where (x_i^t, y_i^t) are the coordinates of a point in the sampled output feature map and (x_i^s, y_i^s) are the corresponding source coordinates in the input feature map. W and H are the width and height of the input feature map respectively. M and N are the width and height of the sampled output feature map respectively. l_1 and l_2 are the width and height of the body region in the input feature map. Using bilinear interpolation, the value at (x_i^t, y_i^t) in the output feature map can be calculated as

$$V_i^c = \sum_{n=1}^{H}\sum_{m=1}^{W} U_{nm}^c \max\left(0, 1 - \left|x_i^s - m\right|\right) \max\left(0, 1 - \left|y_i^s - n\right|\right) \tag{5}$$

where U_{nm}^c is the value at (n, m) in the c-th channel of the input feature map.

Feature Extraction Module. The feature extraction module extracts features of the sampled partial feature maps. Each feature extraction module is designed as one residual block. In this manner, the extracted features have different granularities. Feature extraction modules do not share weights because they are responsible for extracting features with different granularities.

Embedding Operation. As illustrated in Fig. 3, we perform embedding operations on the output feature maps of feature extraction modules. Identification loss L^{ID} and triplet loss $L^{triplet}$ [30,31] are used. During inference, four feature vectors in the four

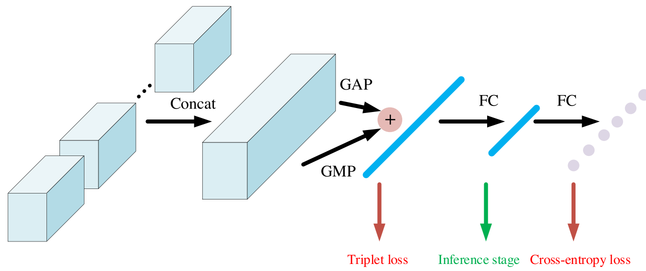


Fig. 3. Illustration of embedding operations. In residual block-3 and residual block-4, concatenation is omitted because there is only one feature map. GAP: global average pooling. GMP: global max pooling.

embedding operations are weighted and concatenated after normalization to form the final feature:

$$\boldsymbol{F}^{final} = [\lambda_1^F \hat{\boldsymbol{F}}_1, \lambda_2^F \hat{\boldsymbol{F}}_2, \lambda_3^F \hat{\boldsymbol{F}}_3, \lambda_4^F \hat{\boldsymbol{F}}_4] \tag{6}$$

where λ^F is the weight factor for each feature vector.

3.2. Loss function

For the first three residual blocks, we have L^{loc} for location tasks and L^{ID} and $L^{triplet}$ [30,31] for person ReID. Thus, the loss function for the i-th output feature map of the first three residual blocks is

$$L_i = L_i^{loc} + L_i^{ID} + L_i^{triplet} \tag{7}$$

After the fourth residual block, there is no location module, and the loss function is

$$L_4 = L_4^{ID} + L_4^{triplet} \tag{8}$$

The final loss function can be written as:

$$L = \lambda \sum_{i=1}^{3} (L_i^{loc} + L_i^{ID} + L_i^{triplet}) + (L_4^{ID} + L_4^{triplet}) \tag{9}$$

where λ is the weight factor and i denotes the ith residual block.

4. Experiments

4.1. Datasets

We present our experiments on the following four widely used person ReID datasets.

Market1501. This dataset [42] contains 32,668 images of 1501 identities. Bounding boxes are given by a pedestrian detector of a deformable part model. The dataset is divided into a training set with 12,936 images of 751 persons and a testing set of 750 persons containing 3,368 query images and 19,732 gallery images.

DukeMTMC-reID. In this dataset [43,44], there are 1,404 identities appearing in more than two cameras and 408 identities appears in only one camera. It is divided into a training set of 702 identities with 16,522 images and a testing set of 702 identities with 19,889 images. Only single query is supported for testing.

CUHK03. Following the new protocol similar to that of Market1501, the CUHK03 dataset [45] is split into training set of 767 identities and testing set of 700 identities appears in only one camera. This dataset has two methods of annotating bounding box, including labeled by a human or detected by a detector.

MSMT17. This is a large-scale dataset [46] that contains 126,441 images taken by 15 cameras. This dataset is very challenging because it has both outdoor and indoor scenes. It is divided into a training set of 32,621 images and a testing set of 93,820 images.

4.2. Implementation details

All images are resized to 256 × 128. Random flipping and erasing are applied for data augmentation. Soft labels are used. We use ResNet-50 [47] pretrained on ImageNet with instance-batch normalization [48] as the base model. The Adam optimizer and warming up with a slope of 0.01 are used for training. MGRe is trained with an initial learning rate of 0.00035. The learning rate is decayed by 0.3 every 30 epoches from 120 to 210 epochs. We train the model for 240 epochs in total. During inference, four 512-dim feature vectors forms the final 2048-dim feature vector, as specified in Eq. (6). The loss is formulated as in Eq. (9). The weight factors λ^F and λ are automatically selected on the validation set.

We follow the standard evaluation protocol in ReID. Specifically, we adopt the cumulative matching characteristics (CMC) at Rank-1 (R-1) and mean average precision (mAP) for performance evaluation.

4.3. Comparison with state-of-the-art approaches

MGRe is compared with 26 state-of-the-art methods proposed in recent years to show our considerable performance advantage over all the existing competitors. The experimental results are summarized in Table 2. The compared methods are divided into three categories. Methods category "Attention" learn attention masks to enhance feature representation. Methods belong to category "Stripe" divide the feature map of an input image into several horizontal stripes to exploit features from multiple parts. Methods in category "Pose" leverage the coarse pose/part semantic information to assist ReID.

Market1501 and DukeMTMC. The results on Market1501 and DukeMTMC are summarized in Table 2. MGRe achieves the best mAP on Market1501 and outperforms all pose-based methods.

Our MGRe belongs to the "Pose" category and exhibits superiority to all other models in this category, including Spindle, Pose-driven, AACN, PIE, SPReID, P^2-Net and DSA-reID. Spindle Net [2] also crops human body regions in different stages, which is somewhat similar to our proposed approach. However, Spindle crops small human body regions in coarse-grained feature maps and crops large human body regions in fine-grained feature maps. Moreover, only the coarse granularity features of these parts are used in Spindle. In contrast, our MGRe crops proper-grained body regions from corresponding-grained feature maps, and the final features in MGRe have multiple granularities. DSA-reID achieves the best results in this category. Our MGRe surpasses it by a large margin, e.g., 7.7% mAP and 5.1% Rank-1 on DukeMTMC. MGRe jointly optimizes the location of body regions and person re-identification while DSA-reID separates these two processes. Our joint optimization strategy on the one hand brings performance gain and on the other hand saves inference time.

MGRe is also the state-of-the-art comparing with other two categories. The latest work SCAL [35] in category "Attention" achieves the best performance. Our MGRe surpasses SCAL and achieves an increase of 2.4% mAP and 2.3% Rank-1 on DukeMTMC. It worth noting that st-ReID [37], the best one in category "Stripe" utilizes temporal information, which is also very useful. MGRe can achieve comparable results with st-ReID without using temporal information. Moreover, MGRe also significantly outperforms Pyramid, the second best one in category "Stripe", and achieves an increase of 3.0% mAP on Market1501. It is because Pyramid only splits the final coarse-grained feature maps into stripes, while MGRe extracts multi-granularity body regions from multi-granularity feature maps.

CUHK03. The results on CUHK03 are also summarized in Table 2. Here, both labeled and detected settings are used MGRe outper-

Table 2
Performance (%) comparison with state-of-the-art methods. Superscript * indicates models that rely on other datasets, e.g., pose estimation, human segmentation or person attributes during training and inference. Superscript † denotes using temporal information. Our method does not need other datasets during inference. Bold numbers denote the best performance, while numbers with underlines denote the second best.

	Method	Market1501		DukeMTMC		CUHK03			
						Detected		Labeled	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Attention	HA-CNN [10](CVPR2018)	75.7	91.2	63.8	80.5	38.6	41.7	41.0	44.4
	DuATM [32](CVPR2018)	76.6	91.4	64.6	81.8	–	–	–	–
	Mancs [33](ECCV2018)	82.3	93.1	71.8	84.9	60.5	65.5	63.9	69.0
	AANet* [11](CVPR2019)	82.5	93.9	72.6	86.4	–	–	–	–
	CASN [12](CVPR2019)	82.8	94.4	73.7	88.7	68.0	73.7	64.4	71.5
	IANet [22](CVPR2019)	83.1	94.4	73.4	87.1	–	–	–	–
	RGA-SC [24](CVPR2020)	88.4	96.1	–	–	74.5	79.6	77.4	81.1
	ABD-Net[34](ICCV2019)	88.3	95.6	78.6	89.0	–	–	–	–
	SCAL[35](ICCV2019)	89.3	95.8	79.6	89.0	68.6	71.1	72.3	74.8
Stripe	Deep-Person [36](PR2020)	79.6	92.3	64.8	80.9	–	–	–	–
	PCB+RPP [4](ECCV2018)	81.6	93.8	69.2	83.3	57.5	63.7	–	–
	HPM [13] (AAAI2019)	82.7	94.2	74.3	86.6	–	–	–	–
	MGN [14](MM2018)	86.9	95.7	78.4	88.7	66.0	68.0	67.4	68.0
	Pyramid [15](CVPR2019)	88.2	95.7	79.0	89.0	74.8	78.9	76.9	78.9
	st-ReID† [37](AAAI2019)	87.6	**98.1**	**83.9**	**94.4**	–	–	–	–
Pose	Spindle* [2] (CVPR2017)	–	76.9	–	–	–	–	–	–
	Pose-driven* [38](CVPR2017)	63.4	84.1	–	–	–	–	–	–
	AACN*[39](ICCV2018)	66.9	85.9	59.2	76.8	–	–	–	–
	PIE* [5](TIP2019)	69.3	85.2	63.3	79.2	–	–	–	–
	SPReID* [7](CVPR2018)	81.3	92.5	71.0	84.4	–	–	–	–
	P^2-Net [40](ICCV2019)	85.6	95.2	73.1	86.5	68.9	74.9	73.6	78.3
	DSA-reID* [41](CVPR2019)	87.6	95.7	74.3	86.2	73.1	78.2	75.8	78.2
	MGRe (ours)	**90.1**	96.2	82.0	91.3	**79.3**	**81.6**	**82.4**	**84.9**

forms all other methods in three categories by a large margin. RGA-SC [24] achieves the second best result, and our method surpasses it by at least 4.8% and 2.0% in mAP and Rank-1 under the detected setting, 5.0% mAP and 3.8% Rank-1 under the labeled setting. Moreover, it also significantly outperforms MGN [14] by over 10% mAP whose motivation is somewhat similar to MGRe. The reason lies in that MGN only utilizes multi-grained feature stripes, while MGRe extracts multi-granularity body regions from multi-granularity feature maps.

MSMT17. We further evaluate our method on a large-scale dataset MSMT17. This dataset is released in 2018, therefore only a few latest works report their results on this dataset. The results are summarized in Table 3. Our method again outperforms all existing methods. Although JG-Net [51] utilizes generative models to generate more training samples, MGRe still outperforms it and achieves an increase of 10.0% mAP and 5.9% Rank-1. ABD-Net from category "Attention" achieves the second best result on this dataset. It applies strong orthogonal constraints to features and convolution kernels of the network, leading to high computational complexity. Similar to the results on Market1501 and DukeMTMC-reID, MGRe again outperforms ABD-Net on this dataset.

Table 3
Performance (%) comparison with state-of-the-art methods on MSMT17. Bold numbers denote the best performance, while numbers with underlines denote the second best. MGRe achieves the best performance on all datasets.

Method	mAP	R-1	R-5
Google [49](CVPR2015)	23.0	47.6	75.0
Pose-driven [6](CVPR2017)	29.7	58.0	73.6
GLAD [50](MM2017)	34.0	61.4	76.8
IANet [22](CVPR2019)	46.8	75.5	85.5
JG-Net [51](CVPR2019)	52.3	77.2	87.4
RGA-SC [24](CVPR2020)	57.5	80.3	–
ABD-Net [34] (ICCV2019)	60.8	82.3	90.6
MGRe (ours)	**62.3**	**83.1**	**91.3**

4.4. Ablation study

To verify the effectiveness of each component and setting in MGRe, we conduct several ablation studies on Market1501 and DukeMTMC.

Component analysis. To investigate the effectiveness of modules in different residual blocks, we do corresponding experiments. Results are presented in Table 4. *Baseline+i* denotes modules are inserted into the *i*-th block. *Baseline+1+2+3* is the final MGRe. It verifies that MGRe outperforms the baseline by a large margin. Compared with *Baseline+2+3*, MGRe has certain performance improvement in both datasets. There are two reasons for that the improvement is slight. First, there may exist some redundant features between Residual Block-1 and Residual Block-2/3. But comparing Baseline and *Baseline+1*, we can see performance booming. *Baseline+1* outperforms *Baseline* by over 3% mAP on both Market1501 and DukeMTMC, which clearly reveals the effectiveness of Residual Block-1. Residual Block-1 provides fine-grained features which are helpful for differentiating some difficult pairs. Moreover, it provides deep supervision to low-level features, thus leading to a faster and better convergence for the network. Second, the result of *Baseline+2+3* is already very impressive, so it is difficult to achieve significant increase based on *Baseline+2+3*. In addition, *Baseline+2* and *Baseline+3* also achieve good performance. We can conclude that modules in each block are effective. By combining features from all three blocks, *MGRe (Baseline+1+2+3)* achieves the best results.

Architecture. Previous analysis show that MGRe reaches the best performance by extracting multi-granularity features from corresponding multi-granularity body regions. In order to verify the effectiveness of the multi-granularity architecture, we change the architecture to single-grained features and single-grained regions.

As reported in Table 5, *Single-grained Features* means that all of the proposed modules, including location, sampling and FE modules are performed on the output feature maps of residual block-4, i.e., the features are all coarse-grained. *Single-grained Regions*

Table 4
Evaluation (%) of components of MGRe architecture. *Baseline+i* denotes inserting modules to the *i*-th block. *Baseline+1+2+3* is the final MGRe. Bold numbers denote the best performance, while numbers with underlines denote the second best.

Method	Market1501		DukeMTMC	
	mAP	Rank-1	mAP	Rank-1
Baseline	85.8	94.3	76.6	86.5
Baseline+1	89.0	95.4	79.7	90.1
Baseline+2	89.0	95.2	80.0	90.6
Baseline+3	88.8	95.5	80.0	90.1
Baseline+2+3	90.0	95.5	81.7	90.7
MGRe(Baseline+1+2+3)	**90.1**	**96.2**	**82.0**	**91.3**

Table 5
Evaluation (%) of effectiveness of MGRe architecture on Market1501. *Single-grained Features* means all the proposed modules including location, sampling and FE modules are performed on the 4-th block. Bold numbers denote the best performance, while numbers with underlines denote the second best. *Single-grained Regions* means only global features are extracted in all blocks.

Method	mAP	R-1	R-5
Single-grained Features	87.5	95.1	98.2
Single-grained Regions	89.6	95.3	**98.5**
MGRe	**90.1**	**96.2**	**98.5**

means that we only extract global features in all the output feature maps of residual blocks, i.e., the regions are all coarse-grained. The results clearly demonstrate that MGRe outperforms other designs. Spindle [2] and PIE [5] extract multiple body regions and then feed them to convolution layers with the same depth, which is somewhat similar to *Single-grained Features*. This experiment further verifies that our MGRe outperforms other pose-based models. Comparing the results of MGRe to that of *Single-grained Regions*, we can see that MGRe outperforms *Single-grained Regions* in both mAP and Rank-1. It clearly shows the sampling modules improves the results. The reason lies in that there exist noises in feature maps, especially in low-level fine-grained feature maps. Moreover, there are misalignment problems, especially for fine-grained regions, since the fine-grained features are extracted from shallow layers in which the receptive field is small and thus are very sensitive to translation, pose variations, etc. Sampling modules can filter out the useless noise and make the features well aligned. In addition, through experiments we find that extracting whole-image features in the output feature map of residual block-3 can achieve slightly better results than extracting whole-body features. The reason may be that each pixel in high-level feature maps has large receptive fields; thus, cropping feature maps may lose much information.

Performance with Occlusion. When certain body regions are occluded, pose-based methods may fail to extract features from those regions. In Table 6, we compare the performance of MGRe in occluded situation to the best methods in other two categories. Because SCAL and Pyramid have not released official codes, we choose the second best methods, e.g., ABD-Net and MGN. We randomly erase the testing set with a probability of 0.5 to simulate the occlusion. We can see that ABD-Net and MGN all decline sharply in performance. The results clearly illustrate that our MGRe is the most robust to occlusion. The reason lies in that with pose guidance, MGRe can be aware of those occlusion. And with the joint optimization of location and feature learning, MGRe tends to focus on other interested areas. In the contrast, attention-based may pay attention to those occluded areas and stripe-based methods cannot handle body region misalignment without pose guidance.

Table 6
Performance (%) with occlusion. *Occlusion* denotes the performance of corresponding methods when facing with occlusion on Market1501. Superscript * represents the results are reproduced by us for fair comparison. Bold numbers denote the best performance, while numbers with underlines denote the second best.

	Method	mAP	R-1
Attention	ABD-Net[34]*	88.0	95.0
	ABD-Net [34](occlusion)*	80.7	91.6
	Decline	−7.3	−4.3
Striple	MGN[14]*	86.6	94.8
	MGN[14](occlusion)*	77.8	90.8
	Decline	−8.8	−4.0
Pose	MGRe	90.1	96.2
	MGRe (occlusion)	85.9	94.0
	Decline	−**4.2**	−**2.2**

Effectiveness of joint optimization. To evaluate the effectiveness of the joint optimization (JO) of the multi-granularity region detection and multi-granularity ReID feature learning, we conduct corresponding experiments. The results are summarized in Table 7. 1) *Baseline+JO* denotes inserting location modules and the location loss to the baseline without sampling and FE modules. In other words, *Baseline+JO* learns the location of human region and the pseudo labels are given by pose-estimation method. It worth mentioning that *Baseline+JO* only utilizes the final global features, the same as *Baseline*. Thus comparing *Baseline+JO* with *Baseline*, we can clearly see pose-estimation brings improvement in the training phase. Through learning to predict the location of human regions, the network will pay more attention to these regions and will provide more discriminative features. 2) *MGRe w/o JO* denotes directly using an off-the-shelf pose-estimation model to locate body regions. It separates the location task and ReID feature learning task, leading to ReID task unable to achieve performance boost through joint optimization. We can see MGRe outperforms *MGRe w/o JO* by 0.6% mAP and 0.6% Rank-1. Moreover, *MGRe w/o JO* relies on pose-estimation model during test phase which can bring additional GFLOPs.

Table 7
Effectiveness (%) of joint optimization of location and ReID on Market1501. *Baseline +JO* denotes that the location modules are inserted into the baseline and then is jointly optimized. *MGRe w/o JO* denotes using an off-the-shelf pose-estimation model to locate body regions. Bold numbers denote the best performance, while numbers with underlines denote the second best.

Method	mAP	R-1	R-5
Baseline	85.8	94.3	98.1
Baseline +JO	86.6	95.0	98.3
MGRe w/o JO	89.5	95.6	**98.5**
MGRe	**90.1**	**96.2**	**98.5**

Visualization of Focused Regions in MGRe. Furthermore, we perform class activation map (CAM) [52] to visualize which regions the networks focus on during the training stage. The results are shown in Fig. 4. This figure clearly demonstrates that MGRe pays more attention to more areas of the human body, while the base model mainly focuses on some specific areas. In fine-grained feature maps (the 2nd and 3nd columns in Fig. 4), MGRe focuses on some important local areas of the human body, while in the coarse-grained feature maps (the 4th and 5th columns in Fig. 4), MGRe pays attention to global areas. By synthesizing the features with the four granularities, the final activation map (the 6th column in Fig. 4) not only focuses on global human body regions but also notices these scattered local areas.

Feature dimension. In this subsection, we investigate the influence of the feature dimension. Experimental results with different final feature dimension C on Market1501 dataset are shown in Fig. 5. It is worth mentioning that our final feature is composed of four parts, each part feature dimension accounts for a quarter of the final feature dimension. As the figure illustrates that both the curves of mAP and Rank-1 initially shows an upward trend and start to decline when C exceeds 2048. When the final feature dimension C is set to 2048, a relatively high ReID performance can be obtained. This indicates that simply adding the feature dimension cannot bring much performance gain and we must ensure the diversity of features when adding the feature dimension.

Parameter analysis. In this subsection, we investigate the influence of the weight parameter λ_1^F to λ_4^F. All of them are important for our MGRe. Firstly, grid search method was used to search the most appropriate value on the validation set of Market1501. The validation set is composed of 100 person IDs, split from the training set of Market1501. Because these four parameters always interfere with each other, we only change one of the four parameters and fix the others to observe the variation. The results are shown in Fig. 6. As shown in the figure, all of the curves of mAP increase first and then trend downward. Meanwhile, λ_2^F is the most sensitive to MGRe because its results vary the most. When $\lambda_1^F = 0.35, \lambda_2^F = 0.5, \lambda_3^F = 0.7$ and $\lambda_4^F = 0.8$, a relatively high ReID performance can be obtained. This verifies that both fine-grained and coarse-grained features are significant for our model.

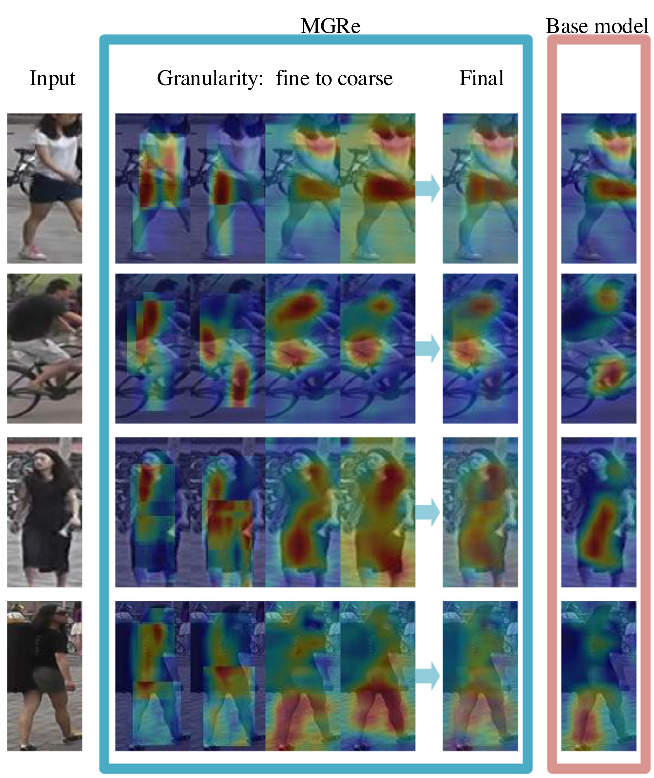


Fig. 4. Comparison of class activation maps between MGRe and base model during training stage.

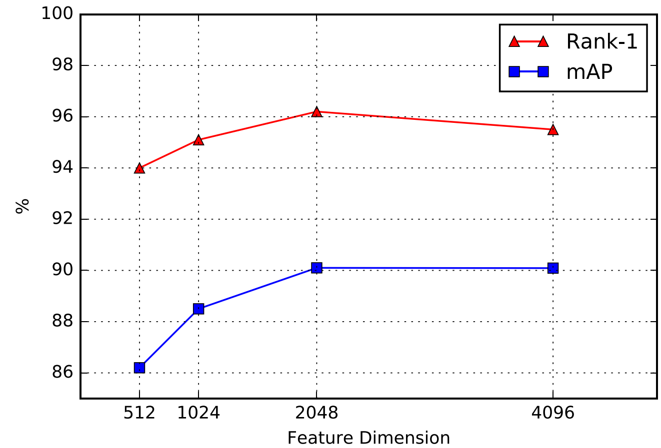


Fig. 5. The ReID performance (mAP and Rank-1 accuracy) on the Market1501 dataset with different final feature dimension.

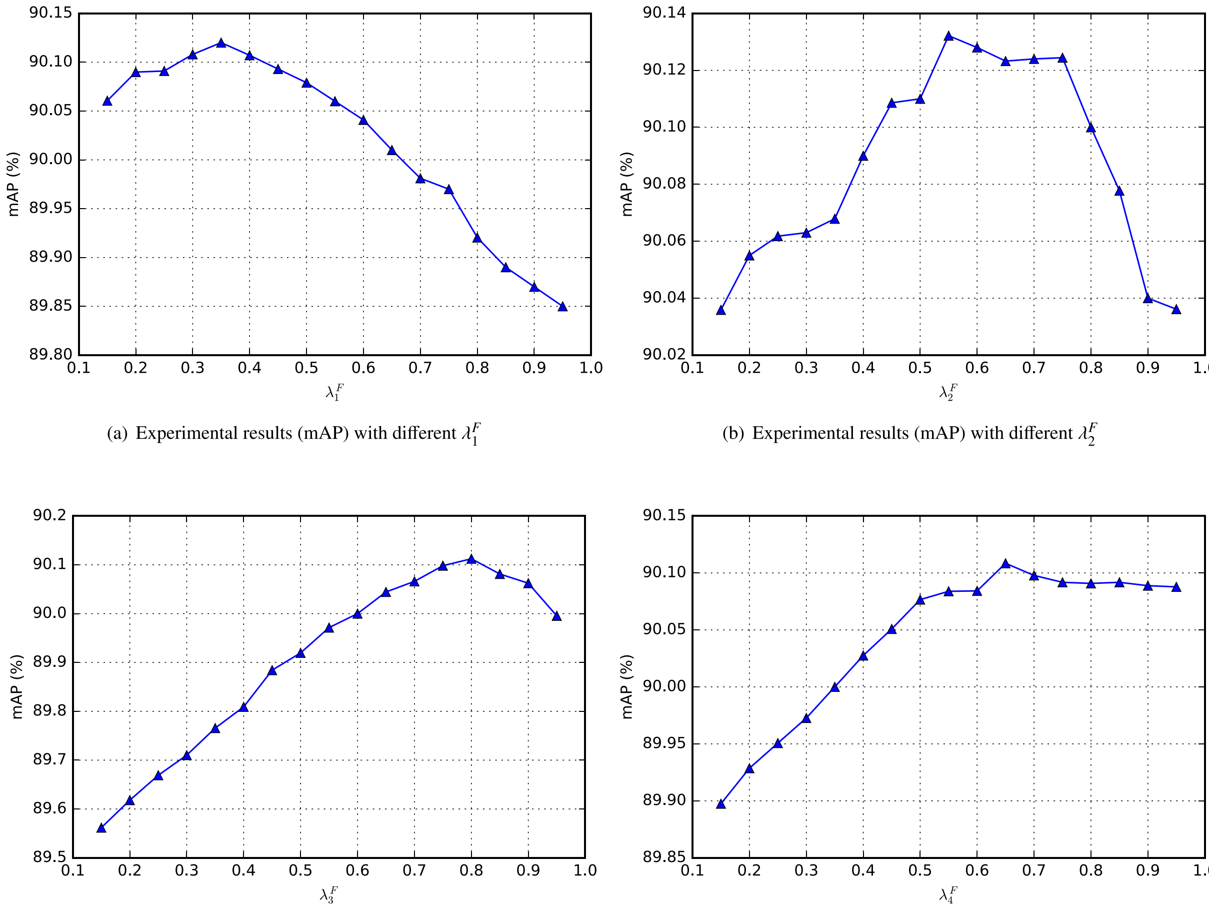
Comparison with large-scale networks. In this subsection, we compare MGRe with some large-scale networks including *Baseline+ResNet* and *ResNest*. The results are summarized in Table 8. We can see that the MGRe has a similar computational cost with *Baseline+ResNest101* but achieves better accuracy in both datasets. MGRe outperforms *Baseline+ResNest101* by 1.1% and 1.9% mAP on Market1501 and DukeMTMC respectively. It is worth noting that the feature dimension of MGRe and *Baseline+ResNest101* are both 2048 (MGRe concatenates four 512-d features from 4 stages). What is more, MGRe significantly outperforms *Baseline+ResNest200* and *Baseline+ResNest269* which have larger computational costs. These experiments clearly reveal the superiority of MGRe.

5. Conclusion

In this paper, we propose a novel multiple granularities ReID approach for learning discriminative local and global features. In MGRe, features with fine-to-coarse granularities are learned from corresponding fine-to-coarse grained body regions in different stages of the backbone network. Thus, we can obtain discriminative features where both fine-grained details and coarse-grained abstract information are learned. In addition, the location of body region and the learning of ReID features are optimized jointly. This joint optimization strategy on the one hand push network to focus on human body areas and on the other hand saves inference time. Extensive ablation studies and comparisons verify the effectiveness of the proposed method.

CRediT authorship contribution statement

Kaiwen Yang: Conceptualization, Methodology, Visualization, Software, Writing - original draft. **Jiwei Yang:** Data curation, Software, Investigation, Validation. **Xinmei Tian:** Supervision, Investigation, Funding acquisition, Writing - review & editing.



(a) Experimental results (mAP) with different λ_1^F

(b) Experimental results (mAP) with different λ_2^F

(c) Experimental results (mAP) with different λ_3^F

(d) Experimental results (mAP) with different λ_4^F

Fig. 6. Experimental results (mAP) on Market1501 dataset when parameter λ^F varies.

Table 8
Comparison between MGRe and large-scale networks including ResNet and ResNeSt with different layers on Market1501 and DukeMTMC. Both accuracy and computational cost are considered. Bold numbers denote the best performance, while numbers with underlines denote the second best.

Method	Market1501		DukeMTMC		GFLOPs	Latency(s)
	mAP(%)	R-1(%)	mAP(%)	R-1(%)		
Baseline+ResNet101	86.9	94.3	77.4	87.5	**6.50**	**1.06**
Baseline+ResNet152	87.3	94.6	78.0	87.6	8.95	1.85
Baseline+ResNeSt101	89.0	95.6	80.1	89.5	7.87	2.19
Baseline+ResNeSt200	88.8	95.7	79.9	89.4	12.61	2.37
Baseline+ResNeSt269	86.1	95.2	77.3	88.3	18.06	4.12
MGRe	**90.1**	**96.2**	**82.0**	**91.3**	7.89	1.79

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

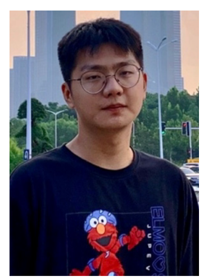
Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 61872329.

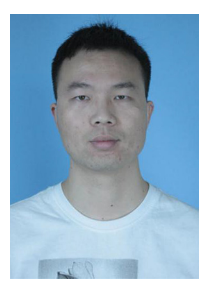
References

[1] F. Zheng, L. Shao, Learning cross-view binary identities for fast person re-identification, in: IJCAI, 2016, pp. 2399–2406.
[2] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: CVPR, 2017, pp. 1077–1085.
[3] J. Yang, X. Shen, X. Tian, H. Li, J. Huang, X.-S. Hua, Local convolutional neural networks for person re-identification, in: ACM MM, ACM, 2018, pp. 1074–1082.
[4] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: ECCV, 2018, pp. 480–496.
[5] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose invariant embedding for deep person re-identification, IEEE Transactions on Image Processing 28 (9) (2019) 4500–4509.
[6] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: CVPR, 2017, pp. 3960–3969.
[7] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: CVPR, 2018, pp. 1062–1071.
[8] J. Miao, Y. Wu, P. Liu, Y. Ding, Y. Yang, Pose-guided feature alignment for occluded person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 542–551.

[9] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, J. Feng, Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8450–8459.
[10] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: CVPR, 2018, pp. 2285–2294.
[11] C.-P. Tay, S. Roy, K.-H. Yap, Aanet, Attribute attention network for person re-identifications, in: CVPR, 2019, pp. 7134–7143.
[12] M. Zheng, S. Karanam, Z. Wu, R.J. Radke, Re-identification with consistent attentive siamese networks, in: CVPR, 2019, pp. 5735–5744.
[13] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, T. Huang, Horizontal pyramid matching for person re-identification, in: AAAI, vol. 33, 2019, pp. 8295–8302..
[14] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: ACM MM, ACM, 2018, pp. 274–282.
[15] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, R. Ji, Pyramidal person re-identification via multi-loss dynamic training, in: CVPR, 2019, pp. 8514–8522.
[16] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: IJCAI, AAAI Press, 2017, pp. 2194–2200.
[17] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 420–429.
[18] E. Ustinova, Y. Ganin, V. Lempitsky, Multi-region bilinear convolutional neural networks for person re-identification, in: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2017, pp. 1–6.
[19] S. Gao, J. Wang, H. Lu, Z. Liu, Pose-guided visible part matching for occluded person reid, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11744–11752.
[20] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.
[21] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, J. Sun, High-order information matters: learning relation and topology for occluded person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6449–6458.
[22] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-and-aggregation network for person re-identification, in: CVPR, 2019, pp. 9317–9326.
[23] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, Y. Yang, Salience-guided cascaded suppression network for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3300–3310.
[24] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3186–3195.
[25] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification, Pattern Recognition 86 (2019) 143–155.
[26] B. Chen, W. Deng, J. Hu, Mixed high-order attention network for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 371–381.
[27] B.N. Xia, Y. Gong, Y. Zhang, C. Poellabauer, Second-order non-local attention networks for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3760–3769.
[28] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: CVPR, 2019, pp. 5693–5703.
[29] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: NeurIPS, 2015, pp. 2017–2025.
[30] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognition 48 (10) (2015) 2993–3003.
[31] G. Wang, J. Lai, X. Xie, P2snet: Can an image match a video for person re-identification in an end-to-end way?, IEEE Transactions on Circuits and Systems for Video Technology 28 (10) (2017) 2777–2787
[32] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A.C. Kot, G. Wang, Dual attention matching network for context-aware feature sequence based person re-identification, in: CVPR, 2018, pp. 5363–5372.
[33] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs, A multi-task attentional network with curriculum sampling for person re-identification, in: ECCV, 2018, pp. 365–381.
[34] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, Z. Wang, Abd-net, Attentive but diverse person re-identification, in: ICCV, 2019, pp. 8351–8361.
[35] G. Chen, C. Lin, L. Ren, J. Lu, J. Zhou, Self-critical attention learning for person re-identification, in: ICCV, 2019, pp. 9637–9646.
[36] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-person: Learning discriminative deep features for person re-identification, Pattern Recognition 98 (2020) 107036.
[37] G. Wang, J. Lai, P. Huang, X. Xie, Spatial-temporal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8933–8940..
[38] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: CVPR, 2017, pp. 384–393.
[39] J. Xu, R. Zhao, F. Zhu, H. Wang, W. Ouyang, Attention-aware compositional network for person re-identification, in: ICCV, 2018, pp. 2119–2128.
[40] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, K. Han, Beyond human parts: Dual part-aligned representations for person re-identification, in: ICCV, 2019, pp. 3642–3651.
[41] Z. Zhang, C. Lan, W. Zeng, Z. Chen, Densely semantically aligned person re-identification, in: CVPR, 2019, pp. 667–676.
[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: CVPR, 2015, pp. 1116–1124.
[43] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: ICCV, 2017, pp. 3754–3762..
[44] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking, 2016.
[45] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid, Deep filter pairing neural network for person re-identification, in: CVPR, 2014, pp. 152–159.
[46] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: CVPR, 2018, pp. 79–88.
[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[48] X. Pan, P. Luo, J. Shi, X. Tang, Two at once: Enhancing learning and generalization capacities via ibn-net, in: ECCV, 2018, pp. 464–479.
[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: CVPR, 2015, pp. 1–9.
[50] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, Glad, Global-local-alignment descriptor for pedestrian retrieval, in: ACM MM, ACM, 2017, pp. 420–428.
[51] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, Joint discriminative and generative learning for person re-identification, in: CVPR, 2019, pp. 2138–2147.
[52] A. Kwaśniewska, J. Rumiński, P. Rad, Deep features class activation map for thermal face detection and tracking, in: HSI, IEEE, 2017, pp. 41–47.



Kaiwen Yang received the B.E. degree from the Xidian University, Xi'an, China, in 2019. He is currently working towards the master degree in the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei, China. His research interests lie primarily in person re-identification, representation learning and machine learning.



Jiwei Yang received the PhD degree in information and communication engineering from University of Science and Technology of China, Hefei, China, in 2019. He is currently an algorithm engineer at Huawei, China. His current research interests include machine learning and its applications to computer vision.

Xinmei Tian received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. She is an Associate Professor in the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei, China. Her current research interests include multimedia information retrieval and machine learning. She received the Excellent Doctoral Dissertation of Chinese Academy of Sciences award in 2012 and the Nomination of National Excellent Doctoral Dissertation award in 2013.