This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Rich Feature Discovery with Class Activation Maps Augmentation for Person Re-Identification

Wenjie Yang [1,2], Houjing Huang [1,2], Zhang Zhang [3], Xiaotang Chen [1,2], Kaiqi Huang [1,2,5], Shu Zhang [4]
[1] CRISE, CASIA [2] University of Chinese Academy of Sciences
[3] CRIPAC & NLPR, CASIA [4] Deepwise AI Lab
[5] CAS Center for Excellence in Brain Science and Intelligence Technology
{wenjie.yang,houjing.huang,zzhang,xtchen,kaiqi.huang}@nlpr.ia.ac.cn, zhangshu@deepwise.com

Abstract

The fundamental challenge of small inter-person variation requires Person Re-Identification (Re-ID) models to capture sufficient fine-grained features. This paper proposes to discover diverse discriminative visual cues without extra assistance, e.g., pose estimation, human parsing. Specifically, a Class Activation Maps (CAM) augmentation model is proposed to expand the activation scope of baseline Re-ID model to explore rich visual cues, where the backbone network is extended by a series of ordered branches which share the same input but output complementary CAM. A novel Overlapped Activation Penalty is proposed to force the current branch to pay more attention to the image regions less activated by the previous ones, such that spatial diverse visual features can be discovered. The proposed model achieves state-of-the-art results on three Re-ID datasets. Moreover, a visualization approach termed ranking activation map (RAM) is proposed to explicitly interpret the ranking results in the test stage, which gives qualitative validations of the proposed method.

1. Introduction

Person Re-Identification (Re-ID) aims to identify a particular person across multiple non-overlapping cameras. It has important application prospects, e.g., large-scale person tracking and person search in video surveillance. Although significant progress has been achieved in the last decade, it is still confronted with many challenges, e.g., various background clutters, large variations of illuminations and camera views and articulated deformations of the human pose.

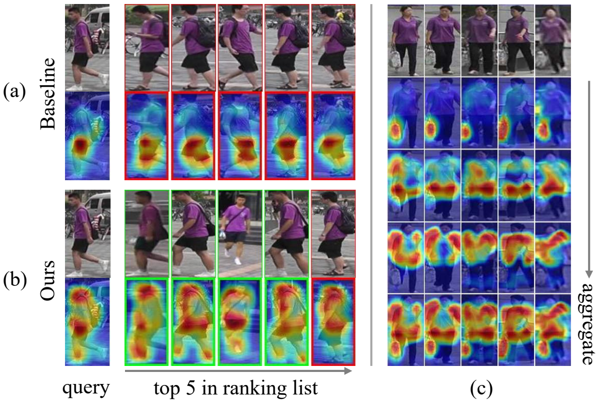


Figure 1: (a) and (b) show the proposed RAMs. The maps highlight the discriminative visual cues used by the baseline and the proposed model to rank the gallery images, respectively. Images with green and red boundary denote true positive and false positive. In (c), the 1st row shows images of the same ID and the CAM in the 2nd row highlights the image regions, i.e., *handbag*, used by the baseline model to identify this person. The CAM in 3rd-4th rows show the proposed method further discover more visual cues.

Moreover, the small inter-person variation makes it difficult for the Re-ID models to distinguish persons with similar appearance based on a small number of visual cues. As illustrated in Fig. 1(a), all the top 5 images are not the same identity with the queried one, as the baseline model mainly pays attention to the *black shorts*, but neglects the other key discriminative visual cues, e.g., *white shoes* and *bag strap*. To further illustrate the point, a diagnostic analysis is conducted on the baseline model. The analysis is based on a recently proposed visualization analytical tool, i.e., CAM [55]. The 2nd row in Fig. 1(c) presents the CAM of the person in Market1501 [49], which shows that the baseline model tends to identify this person based on a small number but discriminative cues, i.e., *handbag*. The small number of cues may be sufficient for distinguishing person IDs in the training set, however, it is essential for a Re-ID model to discover abundant discriminative visual cues so as to form a full-scale characteristic of each identity.

To discover abundant discriminative features from the limited training data, 1) some methods adopt particularly designed regularizations or constraints and proposed various metric learning losses beyond the classification loss, such as triplet loss [13], quadruplet loss [5], and group similarities learning [4]; 2) some methods devote to discover more fine-grained visual cues spread over the whole human body, where multi-branch networks are often proposed to learn fine-grained features from multiple body parts. These parts are obtained through either rigid spatial divisions [42, 26, 46, 7], latent parts localization [23, 27], pose estimation[39, 44, 33], human parsing [19], or attention map [48]; 3) some methods attempt to increase variations in the training data with data augmentation, e.g., random cropping (mirroring) [22], synthesized samples [52] by Generative Adversarial Network (GAN) [11] or adversarially occluded samples [14].

The proposed approach belongs to the second class in the above categorization which focuses on discovering dispersive fine-grained visual feature over whole human body. However, previous work needs an extra step for body parts localization with rigid spatial division, pose estimation or learning latent parts, which increases the complexities and uncertainties of algorithms. Inspired by the CAM in the 2nd row of Fig. 1(c), we propose to expand the activation scope of the baseline model, so that sufficient visual features can be learned over whole human body. Here, the visual discriminative regions are located by Class Activation Maps [55], thus the proposed model is named Class Activation Maps Augmentation (CAMA). In the CAMA, the backbone model is extended by a series of ordered branches, where a new loss function named Overlapped Activation Penalty (OAP) is introduced to force current branch to discover diverse visual cues from those regions less activated by the previous branches, so as to acquire diverse discriminative fine-grained features. To better interpret the ranking results, a visualization method i.e., Ranking Activation Maps, is proposed to explicitly visualize the associated visual features between the query and the gallery images in the ranking list. To the best of our knowledge, it is the first attempt to interpret the ranking results of person Re-ID.

The main contributions of this paper can be summarized as threefold. (1) An end-to-end multi-branch model is proposed to discover sufficient and diverse discriminative fine-grained features flexibly, without the need of rigid spatial division or extra part localization modules. (2) A novel loss function, i.e., OAP, is proposed to force different branches in the CAMA to learn complementary visual feature from different body regions effectively. (3) Extensive experimental results show that a superior performance can be achieved over other state-of-the-art methods on three large datasets, where a new visualization method, i.e., RAM, is proposed to interpret the ranking results of Re-ID for the first time.

2. Related Work

Person Re-ID. Most person Re-ID methods focus on learning an effective feature extractor [42, 40], or metric that pulls the same identities closer while pushes different ones away [9, 13, 5, 25, 36, 4]. Being first introduced in [46, 25], the deep learning based methods have been dominating the person Re-ID community.

The metric learning based methods adopt some regularizations or constraints to guide the Re-ID model to obtain a set of diverse features, such as triplet loss[13], group similarities learning [4] or quadruplet loss [5], where the essential factor lies in the quality of hard sample mining.

To learn an effective feature extractor to capture abundant discriminative features, some methods aggregate global and local representation and show promising performances. They leverage explicit pose estimation [2], human parsing [10], or Spatial Transform Networks (STN) [16] to locate body parts [39, 44, 33, 19, 27, 23], or directly use the predefined rigid parts (horizontal stripes or grids) for fine-grained feature extraction [42, 26, 46, 7, 1]. Compared to the above methods, the proposed approach does not depend on any external parts localization models. However, the global representation extracted from the top layer of the person Re-ID classification network does not adequately retain visual clues that are crucial for person Re-ID, e.g., fine-grained attributes (sunglasses, shoes) and some texture/edge features at lower semantic level [23, 3, 14]. Therefore, some researchers [3, 43] propose to fuse discriminative visual features at multiple semantic levels.

Some other methods, e.g., Huang *et al.* [14] augment the variation of training data by generating occluded samples; Song *et al.* [38] introduce segmentation masks as guidance to extract features invariant to background clutters; Shen *et al.* [34] aims to improve the post-processing (i.e., reranking [53]) to make it possible to learn in an end-to-end manner. Moreover, instead of measuring similarity with Euclidean distance, a Kronecker Product Matching module is employed to match feature maps of different persons [36]. Similar idea can alse be found in [37].

Network Visualization. Convolutional neural networks (CNN) are usually treated as a black-box function that maps a given input to a task-specific output. There has been much work devoted to explore how the CNN works, e.g., DeconvNet [47] visualizes what patterns activate a specific neurons, Network Inversion [30] sheds light on the information represented at each layer by inverting them to synthesize an input image, and Class Activation Map (CAM) is proposed to visualize the input image regions used when CNN making decisions [55]. In this work, we enhance the capability of Re-ID model for capturing diverse fine-grained knowledge through inspecting the intrinsic working mechanisms of a trained network, which is closely related with current researches on network interpretability and visualization.

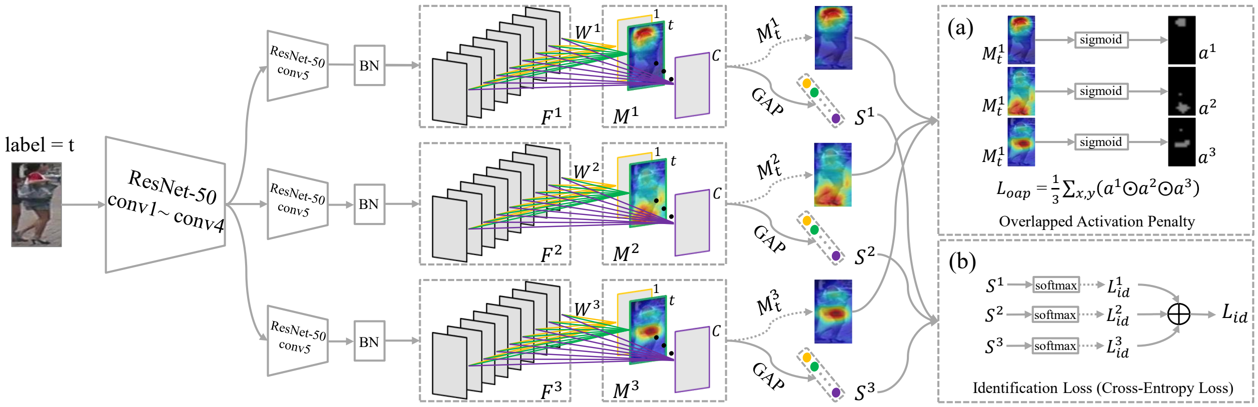


Figure 2: The CAMA model with 3 branches. The image with label t passes through ResNet-50 and Batch Normalization (BN) layer to form feature maps $F^i \in \mathbb{R}^{h\times w\times d}$, which are weighted summed (Eq. (4)) by $W^i \in \mathbb{R}^{d\times C}$ to obtain the CAM, i.e., $M^i \in \mathbb{R}^{h\times w\times C}$. Then Global Average Pooling (GAP) is applied on M^i to obtain the class scores $S^i \in \mathbb{R}^{1\times C}$, where C is the number of training classes. The t-th channel of M^i, i.e., $M_t^i \in \mathbb{R}^{h\times w}$, highlights the image regions used by branch i to identify the input image. In (a), the M_t^is are used to calculate the Overlapped Activation Penalty L_{oap} so that the activation regions in different branches are non-overlapping. In (b), the identification losses L_{id}^is are summed to obtain L_{id}. The W^is in branches do not share parameters and the $\odot$ in (a) denotes element-wise multiplication.

3. Methods

This section presents the technical details on the CAMA model. As shown in Fig. 2, the CAMA model is a multi-branch (MltB) neural network including a backbone and a number of ordered branches. Without the need of extra parts localizations and exhaustive searching of informative regions, we propose to utilize a technique on visual explanation of deep learning, i.e., the CAM [55], to indicate the locations of informative parts and the richness of features embedded in the Re-ID model (Sec. 3.1). The new extended branches are guided by the a novel loss function called OAP to discover discriminative features from those regions less activated by the previous branches (Sec. 3.2). Finally, a visualization technique termed RAM is proposed to interpret the ranking results of a query image(Sec. 3.4).

3.1. Baseline Model

The forward propagation of the ID-discriminative embedding (IDE) model specified in [50] is as follows. Firstly, the input image passes through the CNN to obtain a tensor $T \in \mathbb{R}^{h\times w\times d}$, which can be interpreted as dense $h \times w$ grids of d-dimensional local features $T(x, y) \in \mathbb{R}^{1\times d}$ of spatial location (x, y), or dense d channels of feature map $T_k \in \mathbb{R}^{h\times w}$, then global average pooling (GAP) is applied on T to obtain a feature vector, finally a fully-connected (FC) layer is used to transfer the feature vector into the class scores $S \in \mathbb{R}^{1\times C}$. Here C is the number of training classes. The above procedure can be formulated as

$$S = \mathrm{FC}(\mathrm{GAP}(T)) \tag{1}$$

The work [45] proposes to add a batch normalization (BN) layer [15] after the global average pooling layer, which can be formulated as

$$S = \mathrm{FC}(\mathrm{BN}(\mathrm{GAP}(T))) \tag{2}$$

where BN is the vanilla batch normalization [15] for a 1D input. The Eq. (2) is denoted as *IDE+BN* in this paper. Since *IDE+BN* achieves better performance over the *IDE*, we adopt *IDE+BN* as the baseline. It is noted that the BN and the GAP are linear transformations, thus the order of these two transformations can be exchanged without changing the final results. Thus Eq. (2) is further formulated as

$$\begin{aligned} S &= \mathrm{FC}(\mathrm{BN}(\frac{\sum_{x,y} T(x,y)}{h\times w})) \\ &= \mathrm{FC}(\frac{\sum_{x,y} \mathrm{BN}(T(x,y))}{h\times w}) \\ &= \mathrm{FC}(\frac{\sum_{x,y} F(x,y)}{h\times w}) = \mathrm{FC}(\mathrm{GAP}(F)) \end{aligned} \tag{3}$$

where $F(x, y) = \mathrm{BN}(T(x, y))$ and $F \in \mathbb{R}^{h\times w\times d}$ is a tensor. Note that without this re-formulation, the following mentioned CAM can not be conveniently integrated with the BN augmented baseline, i.e., Eq. (2).

Class Activation Maps [55]. In Eq. (3), after applying GAP on F, a feature f is obtained. Then a FC layer $W \in \mathbb{R}^{d\times C}$ is used to transfer f into the class scores $S \in \mathbb{R}^{C}$. Since $W_c \in \mathbb{R}^{d\times 1}$ is a weight vector that generates a score S_c for class c ($c \in \{1, 2, ..., C\}$). We can obtain S_c by

$$\begin{aligned} S_c = f \cdot W_c &= \sum_{k=1}^{d} W_{c,k} \times f_k \\ &= \sum_{k=1}^{d} W_{c,k} \frac{1}{h \times w} \sum_{x,y} F_k(x,y) \\ &= \frac{1}{h \times w} \sum_{x,y} \sum_{k=1}^{d} W_{c,k} F_k(x,y) \end{aligned} \quad (4)$$

where $W_{c,k}$ denotes the k-th element of W_c and (x, y) indicates the spatial location. The class activation map of class c is defined as M_c, where $M_c(x, y) = \sum_{k=1}^{d} W_{c,k} F_k(x, y)$. $M_c(x, y)$ indicates the score that local feature $F(x, y) \in \mathbb{R}^{1 \times 1 \times d}$ contributes to S_c.

Suppose that t is the target class of the input image, M_t is the class activation map of target class which indicates the decision evidence of the CNN model as identifying the input image. In this paper, we explore the usage of CAM for person Re-ID in twofold. 1) We utilize CAM to locate discriminative visual cues, based on which a learning principle aiming to augment the CAM is proposed to enhance the Re-ID model. 2) Inspired by the CAM, we propose the RAM for interpreting the ranking results in the test stage.

3.2. Overlapped Activation Penalty

For each training image I with label t, where t is the index of the target class. I passes through N branches respectively to obtain CAM, i.e., $M^i \in \mathbb{R}^{h \times w \times C}$, for each branch i. The t-th channel of M^i, i.e., $M_t^i \in \mathbb{R}^{h \times w}$, corresponds to the activation map of class t, and we use it to further obtain a^i to specify the image regions that i-th branch focuses on. Here, we traverse all spatial locations (x, y) of M_t^i with the Sigmoid function to obtain a mask $a^i \in \mathbb{R}^{h \times w}$.

$$a^i(x,y) = \frac{1}{1 + \exp(-(M_t^i(x,y) - \sigma^i))} \quad (5)$$

where the threshold σ^i is the k-th largest element of M_t^i. The Sigmoid function maps the elements larger than σ^i in M_t^i approximately to 1 while others to 0.

Since we aim to guide different branches to activate different image regions, the non-zero regions of a^i in different branches should be non-overlapping. To achieve that, the overlapped activation penalty is proposed to measure the area of the overlapped regions of $a^i, i \in \{1, 2, ..., N\}$, which is defined as follows:

$$L_{oap} = \frac{1}{N} \sum_{x,y} \left(a^1 \odot a^2 \odot \cdots \odot a^N\right) \quad (6)$$

where $\odot$ denotes element-wise multiplication and N is the number of branches.

3.3. Objective function

After global average pooling over the class activation maps M^i, we obtain the class scores S^i in branch i, which is further normalized by softmax function into a probability distribution $y^i \in \mathbb{R}^C$. The identification loss in branch i is calculated as the Cross Entropy between the predicted probability y^i and the ground-truth.

$$L_{id}^i = -\log(y_t^i), \quad i \in \{1, 2, \ldots, N\} \quad (7)$$

where t is the index of the target class, and the $L_{id}^i, i \in \{1, 2, \ldots, N\}$, are summed to obtain the identification loss of the CAMA model, i.e., $L_{id} = \sum_{i=1}^{N} L_{id}^i$.

The final objective function for the CAMA model is the weighted summation of L_{id} and L_{oap}.

$$L_{total} = L_{id} + \alpha L_{oap} \quad (8)$$

where α is the trade-off weight and we use $\alpha = 1$ in all experiments below. L_{oap} prefers that the activated regions of different branches are non-overlapping, while L_{id} guides the CAMA model to activate the discriminative image regions rather than the background.

In the test stage, the feature vectors in all branches $\{f^1, f^2, ..., f^N\}$ generated by applying global average pooling on the feature maps $\{F^1, F^2, ..., F^N\}$ are concatenated to obtain the final image representation f, i.e., $f = [\hat{f}^1; \hat{f}^2; ...; \hat{f}^N]$, where $\hat{f}^i$ denotes the L2 normalization of f^i.

3.4. Ranking Activation Map

Since the original CAM can not be implemented on the person IDs that are unseen during the training stage. To better interpret the ranking results, we propose the RAM, which can reveal the associated visual cues between the query and the gallery images. Here, we describe the procedure for generating ranking activation map.

Suppose that F_q and F_g correspond to the feature maps of a query and a gallery image respectively. The feature representation is obtained by $f_q = \mathrm{GAP}(F_q), f_g = \mathrm{GAP}(F_g)$, where GAP denotes global average pooling. Then L2 normalization is performed on f to obtain $\hat{f}$, i.e., $\hat{f} = \frac{f}{\sqrt{\langle f, f \rangle}} = \frac{f}{\|f\|}$. The Euclidean distance between $L2$ normalized f_q and f_g is defined as

$$\begin{aligned} \mathrm{d}(f_q, f_g) &= \sqrt{\langle \hat{f}_q - \hat{f}_g, \hat{f}_q - \hat{f}_g \rangle} \\ &= \sqrt{\langle \hat{f}_q, \hat{f}_q \rangle + \langle \hat{f}_g, \hat{f}_g \rangle - 2\langle \hat{f}_q, \hat{f}_g \rangle} \\ &= \sqrt{2 - 2\langle \hat{f}_q, \hat{f}_g \rangle} = \sqrt{2 - 2\frac{\langle f_q, f_g \rangle}{\|f_q\| \, \|f_g\|}} \end{aligned} \quad (9)$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product of two vectors. Since the similarity is inversely proportional to the distance, we

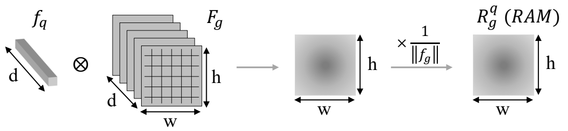


Figure 3: Illustration of Ranking Activation Map (RAM), where $\otimes$ denotes spatial attention. f_q is the feature of query image and F_g denotes feature maps of a gallery image.

can observe from Eq. (9) that the similarity between query and gallery image should be proportional to $\frac{\langle f_q, f_g \rangle}{\|f_g\|}$, which can be further formulated as

$$\begin{aligned}\frac{\langle f_q, f_g \rangle}{\|f_g\|} &= \frac{\langle f_q, \frac{1}{h \times w}\sum_{x,y} F_g(x,y)\rangle}{\|f_g\|} \\ &= \frac{\sum_{x,y}\langle f_q, F_g(x,y)\rangle}{h \times w \times \|f_g\|} \\ &= \frac{\sum_{x,y}\frac{\langle f_q, F_g(x,y)\rangle}{\|f_g\|}}{h \times w} = \frac{\sum_{x,y} R_g^q(x,y)}{h \times w} \qquad (10)\end{aligned}$$

where $R_g^q(x,y) = \frac{\langle f_q, F_g(x,y)\rangle}{\|f_g\|}$ indicates the score that the spatial grid (x,y) of g contributes to the final similarity between q and g. Specifically, R_q^q is obtained by replacing f_g with f_q in Eq. (10). We call $\{R_q^q, R_g^q | g \in \{1, 2, ...N_G\}\}$ the ranking activation maps correspond to the query image q, where N_G is the number of gallery images. By simply upsampling the RAMs to the size of the corresponding images, we can visualize the importance of the image regions leading to the ranking result.

Based on the above approach, we show some examples of the RAMs generated using baseline model in Fig. 4. Here, we only show top 10 images in the ranking results. In Fig. 4(a), the RAM of query image (1st column) indicates that the most salient feature is related to the green pocket. For another query, the most salient region corresponds to the red backpack (Fig. 4 (b)). The RAMs of top 10 gallery images highlight the regions of green pocket and parts with red color for the two queries respectively, which are semantically consistent with the corresponding query images.

4. Experiments

4.1. Datasets and Evaluation Metrics

Experiments are performed on DukeMTMC-reID [32], Market-1501[49], CUHK03 [25]. We adopt the Cumulative Matching Characteristic (CMC) [12] and mean Average Precision (mAP) [49] for performance metrics. All the experimental evaluations follow the single-query setting [49]. **Market1501** contains 1,501 identities captured by 5 high-resolution and one low-resolution cameras with different viewpoints, 12,936 images from 751 identities as used for training, 3,368 query images and 19,732 gallery images from another 750 identities for testing. **DukeMTMC-reID** contains 1,404 identities, 16,522 images for training, 2,228 query images, and 17,661 gallery images. Training and test sets both consist of 702 identities, and person bounding boxes are manually cropped. **CUHK03** consists of 13,164 images of 1,467 persons, and each identity only appears in two disjoint camera views. We adopt the new training/testing protocol proposed in [53], in which 767 identities are used for training and 700 for testing. CUHK03 offers both labeled and detected bounding boxes, we perform experiments on both of them.

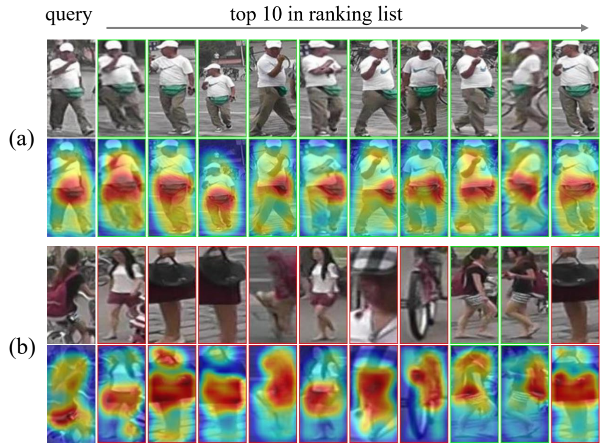


Figure 4: The RAMs highlight the associated visual features between query and gallery images for two ranking list in Market1501, i.e., the green pocket in (a) and red objects in (b). Images with green and red boundary denote true positive and false positive respectively.

4.2. Implementation Details

Model. We adopt the ResNet-50 that pre-trained on ImageNet [8] as the baseline model, which has a convolutional layer (named $conv1$) and four residual blocks, i.e., $conv2 \sim 5$. The common base model in Fig. 2 consists of $conv1 \sim 4$ and the $conv5$s in different branches do not share parameters. The classifier weights are randomly initialized. Note that we follow the setting in PCB [42] that removes the last spatial down-sampling operation in ResNet-50 to increase the spatial size of output feature maps.

Preprocessing. The input image size is fixed to $h \times w = 256 \times 128$ for all experiments on three Re-ID datasets. For data augmentation, standard random cropping and horizontal flipping are used during training.

Optimization. We use Pytorch [31] to implement the CAMA model. The Adam [21] optimizer is used with batch size of 32. We firstly fine-tuning classifier weights, i.e., W^i in Fig. 2, for 10 epochs with the learning rate gradually in-

	Methods	mAP	R-1	R-5	R-10
H	BoW [49] (ICCV15)	20.8	44.4	63.9	72.2
	WARCA [18] (ECCV16)	45.2	68.1	76.0	-
	KLFDA [20] (Arxiv16)	46.5	71.1	79.9	-
G	SVDNet [41] (ICCV17)	62.1	82.3	92.3	95.2
	MGCAM [38] (CVPR18)	74.3	83.8	-	-
	AOS [14] (CVPR18)	70.4	86.5	-	-
	PSE [33] (CVPR18)	69.0	87.7	94.5	96.8
	MultiScale [6] (ICCV17)	73.1	88.9	-	-
	MLFN [3] (CVPR18)	74.3	90.0	-	-
	GCSL [4] (CVPR18)	81.6	93.5	-	-
	SGGNN [35] (ECCV18)	82.8	92.3	96.1	97.4
	DGRW [34] (CVPR18)	82.5	92.7	96.9	98.1
L (+ G)	MSCAN [23] (CVPR17)	57.5	80.3	-	-
	DLPA [48] (ICCV17)	63.4	81.0	92.0	94.7
	PAN [51] (Arxiv17)	63.4	82.8	-	-
	PDC [39] (ICCV17)	63.4	84.1	92.7	94.9
	GLAD [44] (MM17)	73.9	89.9	-	-
	JLML [26] (IJCAI17)	65.5	85.1	-	-
	PABR [40] (ECCV18)	79.6	91.7	96.9	98.1
	HA-CNN [27] (CVPR18)	75.7	91.2	-	-
	PCB [42] (ECCV18)	81.6	93.8	97.5	98.5
	Proposed approach (N=2)	83.9	94.2	97.8	98.4
	Proposed approach (N=3)	**84.5**	**94.7**	**98.1**	**98.8**

Table 1: Market-1501 evaluation, where handcrafted feature based methods (H), global feature based methods (G) and methods employing local feature with or without global feature (L(+ G)) are compared. The best performances are in bold, - means no reported results are available and $N = 2$ denotes the proposed approach with 2 branches.

creased from 3×10^{-6} to 3×10^{-4} and then training the whole CAMA model for another 50 epochs with the initial learning rate 3×10^{-4} multiplied by 0.1 after every 20 epochs. On Market-1501 (12,936 training images), the baseline and the 3-branch CAMA model consumes about 3 and 4 hours respectively with a NVIDIA TITAN X GPU.

4.3. Comparison with State-of-the-art Methods.

We compare the 2-branch and 3-branch CAMA with the state-of-the-art methods. The comparison methods can be separated into handcrafted feature based methods (H), deep learning methods with global feature (G) and deep learning methods employing local feature with or without global feature (L(+G)). The results show that the proposed method achieves the best performance. Note that: **1)** Compared to methods with part-level feature, our method exceeds PCB+RPP [42], which demonstrates the advantage of the proposed CAM-based multi-branches CNN approach to learn diverse features and enhance the discriminative capability of Re-ID models. **2)** Compared to methods with global feature, our methods outperform the MLFN, which uses a fusion architecture to fuse features of multiple semantic levels. Furthermore, the idea of MLFN is compatible with our idea of mining more discriminative features at the high-level, which will be further studied in the future.

	Methods	mAP	R-1
H	BoW [49] (ICCV15)	12.2	25.1
	LOMO+XQDA [28] (CVPR15)	17.0	30.8
G	SVDNet [41] (ICCV17)	56.8	76.7
	AOS [14] (CVPR18)	62.1	79.2
	PSE [33] (CVPR18)	62.0	79.8
	MultiScale [6] (ICCV17)	60.6	79.2
	MLFN [3] (CVPR18)	62.8	81.0
	GCSL [4] (CVPR18)	69.5	84.9
	SGGNN [35] (ECCV18)	68.2	81.1
	DGRW [34] (CVPR18)	66.7	80.7
L (+G)	PAN [51] (Arxiv17)	51.5	71.6
	JLML [26] (IJCAI17)	56.4	73.3
	PABR [40] (ECCV18)	69.3	84.4
	HA-CNN [27] (CVPR18)	63.8	80.5
	PCB [42] (ECCV18)	69.2	83.3
	Proposed approach (N=2)	72.0	84.8
	Proposed approach (N=3)	**72.9**	**85.8**

Table 2: DukeMTMC-reID evaluation. Rank-1 accuracies (%) and mAP (%) are reported, where $N = 2$ denotes the proposed method with 2 branches.

	Methods	labeled		detected	
		mAP	R-1	mAP	R-1
H	BoW+XQDA [49] (ICCV15)	7.3	7.9	6.4	6.4
	LOMO+XQDA [28] (CVPR15)	13.6	14.8	11.5	12.8
G	IDE-C+XQDA [53] (CVPR17)	20.0	21.9	19.0	21.1
	IDE-R+XQDA [53] (CVPR17)	29.6	32.0	28.2	31.1
	TriNet+Era [54] (Arxiv17)	53.8	58.1	50.7	55.5
	SVDNet [41] (ICCV17)	-	-	37.3	41.5
	MGCAM [38] (CVPR18)	50.2	50.1	46.9	46.7
	AOS [14] (CVPR18)	-	-	43.3	47.1
	MultiScale [6] (ICCV17)	40.5	43.0	37.0	40.7
	MLFN [3] (CVPR18)	49.2	54.7	47.8	52.8
L (+ G)	PAN [51] (Arxiv17)	-	-	34.0	36.3
	HA-CNN [27] (CVPR18)	41.0	44.4	38.6	41.7
	PCB [42] (ECCV18)	-	-	57.5	63.7
	Proposed approach (N=2)	64.2	66.1	61.0	64.3
	Proposed approach (N=3)	**66.5**	**70.1**	**64.2**	**66.6**

Table 3: CUHK03 evaluation with the setting of 767/700 training/test split on both the labeled and detected images. Rank-1 accuracies (%) and mAP (%) are reported.

4.4. Discussion with attention based mechanisms.

Although recent video-based work [24] has proposed to diversify attention maps, there are three main differences between [24] and the proposed model. 1) The motivation of [24] is to discover a set of discriminative body parts for avoiding the features from being corrupted by occluded region. While we aim to discover rich or even redundant discriminative features from limited training set so as to enhance the discriminative capability of Re-ID model on unseen test set. 2) On the learning process, [24] connects multiple attention modules with one single branch CNN, so the learned attention modules can be regarded as a set of part detectors relying exactly on the same set of CNN fea-

Models	Market1501				DukeMTMC-reID				CUHK03(detected)			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
IDE + BN	78.8	91.0	96.2	97.6	66.1	79.8	90.5	92.8	56.4	57.7	74.1	82.0
MltB + L_{id}	79.0	91.6	96.5	97.8	65.8	80.7	91.1	93.3	57.6	58.5	75.9	83.6
MltB + L_{id} + L_{oap}	**84.5**	**94.7**	**98.1**	**98.8**	**72.9**	**85.8**	**93.1**	**94.9**	**64.2**	**66.6**	**82.7**	**87.9**

Table 4: Component analysis of the proposed method on three datasets, where mAP, rank-1, rank-5, and rank-10 accuracies are reported. MltB denotes the multi-branch network with 3 branches. The IDE+BN model is formulated as Eq. (3).

tures. While we adopt a multi-branch network to extract more features from training data, where the CAM is used for diversifying the discriminative features not part detectors. 3) In test stage, the attention modules learned by [24] or other attention based methods, e.g., [17], DLPA [48] and the RPP in [42], are also used to calculate the attention maps of test images for weighting the CNN features. While the learned CAM coefficients cannot be used in test stage, as the test person IDs are unseen in the training set. So only the multi-branch CNN networks are adopted to extract features without any additional weighting operations. Thus, from the above three aspects, our method is significantly different from [24] and other attention based methods.

4.5. Effect of Overlapped Activation Penalty.

In this section, we investigate the effect of each component of our method by conducting analytic experiments on three person Re-ID datasets. The results are presented in Table. 4. The difference between MltB+L_{id} and the baseline model (IDE+BN) is that MltB+L_{id} extends a series of ordered branches. However, MltB+L_{id} achieves only a small margin over the baseline model, which indicates that the visual features captured by different branches are almost the same. We can see from the results that MltB+L_{id} + L_{oap} achieves significant improvement over MltB+L_{id}, which validates the powerful capability of the proposed overlapped activation penalty to force the new branch to focus on non-overlapping image regions so as to discover diverse and discriminative visual features.

4.6. Parameters Analysis

In this section, we carry out experiments to study the effect of the threshold σ^i in Eq. (5) and the number of branches N, where σ^i is the k-th largest value of corresponding activation map M_t^i.

Influence of k. The larger k means the larger the activated regions reserved by each branch after the Sigmoid function Eq. (5). Since the last spatial down-sampling operation in ResNet-50 is removed, the CAMA model outputs activation map with $128(16 \times 8)$ spatial grids. As illustrated in Fig. 5, the mAP and rank-1 accuracies fluctuate in a small range when k is less than 26, but sharply decrease when k is greater than 26 ($\frac{26}{128} \simeq 20\%$). That is because when k becomes too large, the overlapped activation penalty will enforce the current branch to activate non-discriminative image regions, e.g., the background, as the most discriminative image regions have been reserved by the previous branch.

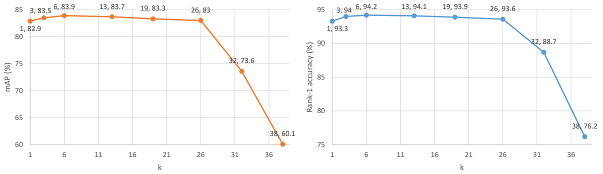


Figure 5: Influence of k, where the number of branches is set to 2. The (k, mAP) and (k, rank-1) results are reported.

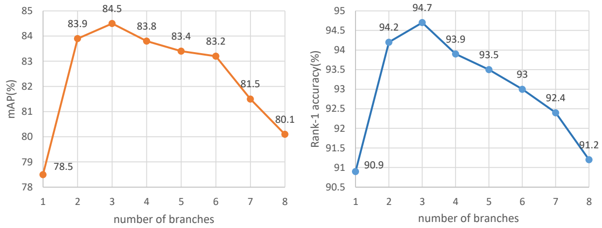
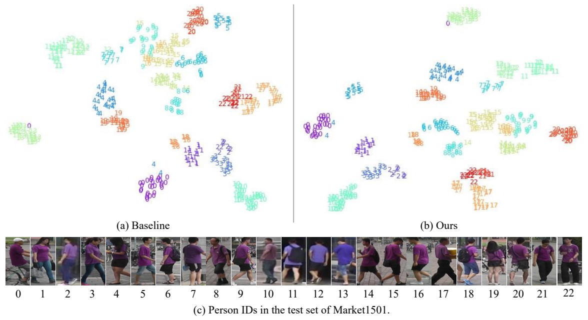


Figure 6: Impact of number of branches N, where mAP and Rank-1 accuracy are compared.

The number of branches N. According to Fig. 6, the CAMA model achieves the best mAP and rank-1 performance when N reaches to 3 on Market1501. The proposed approach does not always perform better with the increase of N, which indicates that the number of discriminative visual cues of person ID is finite. As N increasing to 3, the CAMA model is able to discover new cues on the images. When N is too large, the activated region of the new branch cannot satisfy the constraints of L_{id} (discriminative) and L_{oap} (non-overlapped with the image regions activated by the old branches) simultaneously. In this case, the new branch is harmful to the optimization of the overall network, thus the performance will reach to a peak value at a certain value of N, as illustrated in Fig. 6.

4.7. Why does CAMA Work?

We aim to make the multiple branches of the CAMA model focusing on different regions of the input image during the training stage. However, do these branches pro-


(a) Baseline

(b) Ours

(c) Person IDs in the test set of Market1501.

Figure 7: tSNE visualization of the baseline model and our method on the Market1501 test set. Different numbers indicate different identities (Zoom in for best view).

branch	1st	1st+2nd	1st+2nd+3rd
mAP	79.0	82.7	84.5
rank-1	90.9	93.5	94.7

Table 5: Quantitative analysis on Market1501. For the proposed approach with 3 branches, *1st+2nd* denotes only the 1st and 2nd branch are used for testing.

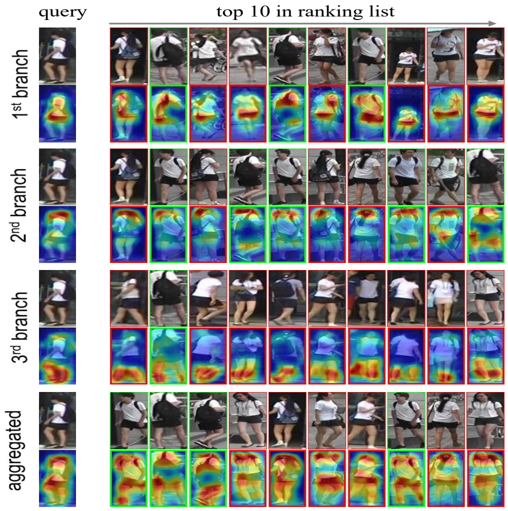


Figure 8: Qualitative analysis. The ranking results and RAMs of different branches and their aggregation result. The features from different branches are indeed complementary. Images with green and red boundary denote true positive and false positive respectively.

duce different visual features in the test stage? Qualitative and quantitative analysis are conducted on the proposed approach with 3 branches.

Qualitative analysis. In Fig. 8, the RAMs reveal the associated visual cues between query and gallery images for each branch. We can observe that for the same input query image, the features learned by different branches are indeed complementary. Specifically, the features learned by 1st branch, i.e., f_q^1, are most related to the black shorts which also activates the 1st ranking list, while the 2nd and 3rd branches focus on the head and the lower body respectively. It indicates that the proposed CAMA model indeed captures diverse discriminative visual cues, thus the concatenated version of features from different branches produces a better ranking result. The RAMs of i-th branch in Fig. 8 are generated by replacing f_q in Eq. (10) with f_q^i.

Quantitative analysis. With the aim of capturing rich visual cues for person Re-ID, the branches in the CAMA model are forced to activate different discriminative image regions. Table. 5 indicates that the mAP and rank-1 accuracy perform better as more number of branches are used for testing. Specifically, the first branch of the CAMA model only achieves 79.0% mAP and 90.9% rank-1 accuracy, as more branches are combined for testing, the mAP, and rank-1 accuracy gradually rise to 84.5% and 94.7% respectively.

Furthermore, we choose a number of person IDs with similar appearance from the test set of Market1501 to visualize the feature distribution by t-SNE [29]. These persons are wearing purple clothes with small inter-person variation as shown in Fig. 7 (c). By comparing Fig. 7 (a) and (b), we can observe that for identities that are hard distinguished by the baseline model, the proposed approach can better distinguish them, e.g. the 9-th, 15-th, and 16-th identities.

5. Conclusion

In this work, we propose a CAMA model to discover discriminative and diverse visual features for person Re-ID which can enhance traditional global representation. The proposed overlapped activation penalty can be implemented flexibly in an end-to-end training framework. Moreover, we introduce RAM to visualize the associated visual features between query and gallery images in a ranking list. With the help of CAM and RAM, we show that the CAMA model indeed acquires more discriminative features, which gives qualitative validations of the learned Re-ID model clearly and gives some insights into the interpretability of person Re-ID model. In this work, we show that it is a promising way to enhance the Re-ID model from a respect of interpretable CNN. In the future, we will extend the idea on more recognition tasks, e.g., zero-shot learning (ZSL).

6. Acknowledgement

This work is joinly funded by the National Key Research and Development Program of China (2016YFB1001005), the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006), and the National Natural Science Foundation of China (Grant No. 61473290, 61673375 and 61721004).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proc. CVPR*, 2015. 2

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016. 2

[3] X. Chang, T. M. Hospedales, and T. Xiang. Multi-level factorisation net for person re-identification. In *Proc. CVPR*, 2018. 2, 6

[4] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proc. CVPR*, 2018. 2, 6

[5] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proc. CVPR*, 2017. 2

[6] Y. Chen, X. Zhu, S. Gong, et al. Person re-identification by deep learning multi-scale representations. 2017. 6

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. CVPR*, 2016. 2

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 5

[9] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015. 2

[10] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 2

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. NIPS*, 2014. 2

[12] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS Workshop*, 2007. 5

[13] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[14] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang. Adversarially occluded samples for person re-identification. In *Proc. CVPR*, 2018. 2, 6

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 3

[16] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Proc. NIPS*, 2015. 2

[17] M. Jiang, Y. Yuan, and Q. Wang. Self-attention learning for person re-identification. In *Proc. BMVC*, 2018. 7

[18] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *Proc. ECCV*, 2016. 6

[19] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *Proc. CVPR*, 2018. 2

[20] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. I. Camps, and R. J. Radke. A comprehensive evaluation and benchmark for person re-identification: Features. *Metrics, and Datasets. arXiv preprint*, 2016. 6

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012. 2

[23] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proc. CVPR*, 2017. 2, 6

[24] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proc. CVPR*, 2018. 6, 7

[25] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. CVPR*, 2014. 2, 5

[26] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *Proc. IJCAI*, 2017. 2, 6

[27] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proc. CVPR*, 2018. 2, 6

[28] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proc. CVPR*, 2015. 6

[29] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 8

[30] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proc. CVPR*, 2015. 2

[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. *NIPS Workshops*, 2017. 5

[32] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proc. ECCV*, 2016. 5

[33] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. 2018. 2, 6

[34] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang. Deep group-shuffling random walk for person re-identification. In *Proc. CVPR*, 2018. 2, 6

[35] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang. Person re-identification with deep similarity-guided graph neural network. In *Proc. ECCV*, 2018. 6

[36] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proc. CVPR*, 2018. 2

[37] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual attention matching network for context-aware feature sequence based person re-identification. 2018. 2

[38] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *Proc. CVPR*, 2018. 2, 6

[39] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *Proc. ICCV*, 2017. 2, 6

[40] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee. Part-aligned bilinear representations for person re-identification. In *Proc. ECCV*, 2018. 2, 6

[41] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *Proc. ICCV*, 2017. 6

[42] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proc. ECCV*, 2018. 2, 5, 6, 7

[43] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *Proc. CVPR*, 2018. 2

[44] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM Multimedia*, 2017. 2, 6

[45] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou. Towards good practices on building effective cnn baseline model for person re-identification. *arXiv preprint arXiv:1807.11042*, 2018. 3

[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *Proc. ICPR*, 2014. 2

[47] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, 2014. 2

[48] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proc. ICCV*, 2017. 2, 6, 7

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proc. ICCV*, 2015. 1, 5, 6

[50] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 3

[51] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv preprint arXiv:1707.00408*, 2017. 6

[52] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. ICCV*, 2017. 2

[53] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proc. CVPR*, 2017. 2, 5, 6

[54] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 6

[55] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proc. CVPR*, 2016. 1, 2, 3