Multi-scale Multi-patch Person Re-Identification with Exclusivity Regularized Softmax

Cheng Wang[a], Liangchen Song[b], Guoli Wang[b], Qian Zhang[b], Xinggang Wang[a,*]

[a]*School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China*
[b]*Horizon Robotics Inc., Beijing 100086, China*

Abstract

Discriminative feature learning is critical for person re-identification. To obtain abundant visual information from the input person image, we first propose a novel network that extracts multi-scale patch-level deep features. Then, we propose an improved softmax loss function for learning more compact and more discriminative feature vectors. Specifically, we integrate feature pyramid blocks and region-level global average pooling functions into the feature extraction network, introduce the well-established normalization techniques in face recognition algorithms into person re-ID, and penalize the redundancy in feature vectors by minimizing the $l_{1,2}$ norm of the weight matrix in the softmax layer. Experiments on three large-scale datasets under the standard settings show the effectiveness of the proposed method. Moreover, we report our cross-domain re-ID results by training re-ID models on source datasets and testing them on other target datasets.

Keywords: Person re-identification, deep learning, exclusivity regularized softmax

1. Introduction

The task of image-based person re-identification (re-ID) refers to retrieve images of some specific pedestrians from a database of images. Benefiting from the great success of deep Convolution Neural Networks (CNN) on various computer vision tasks,

[*] Corresponding author
Email address: xgwang@hust.edu.cn (Xinggang Wang)

re-ID methods based on deeply-learned features have outperformed human experts and shown impressive performance. Recently, by exploiting part information in images, several state-of-the-art methods can better discriminate the non-salient or infrequent detailed information [1]. However, due to the large variations in person images, learning deep features that can capture very detailed cues in an image is very difficult.

To better capture the detailed information, recently proposed methods can be roughly classified into two directions. One direction is to use local body-parts features which contain fine-grained information [1]. Directly training a network with only local part-level features can enforce the network to focus on those non-salient but discriminative components. However, using part-level features is based on the assumption that the parts should be accurately located. To achieve this goal, existing methods either incorporate pose estimation models [2], or decompose an image into parts and then compute classification loss separately [3]. All of the above approaches can effectively align parts, but there is no guarantee that all the alignments are exactly accurate. Thus, if trained with aligning parts, the model is at the risk of learning or memorizing irrelevant background noise in an image, which will lead to the deterioration of the generalization capability of the model [4].

The other direction combines the features computed at multiple intermediate layers of a network [5, 6]. The intuition behind is that the feature maps tend to of higher semantic level from the bottom to the top layers in a network [7]. Thus, the non-salient details, which are usually ignored by top layers but kept in the low-level features, are contained in the final multi-scale features. Although those rare yet crucial patterns are included in the feature, computing classification loss or metric loss such as triplet loss does not always result in focusing on those infrequent patterns if other patterns are discriminative enough. Therefore, simply forcing the feature having different scales is not sufficient for gaining the power of encoding the non-salient patterns.

In this paper, we propose to fuse the two above ideas for better mining the local non-salient details. To reach this goal, we design a multi-scale multi-patch network. Specifically, we convert the intermediate features in a network to the same size and then add them up to get a multi-scale feature. Then, a global pooling branch and a part pooling branch [1] are connected to the multi-scale feature.

Apart from a novel network for extracting better features, we incorporate the normalization technique into the training process of our network. The normalization technique is effective and well-established in face recognition algorithms [8, 9], which is similar to person re-ID in terms of the image retrieval aspect. More precisely, for the final softmax layer, we normalize the weights and the embedding feature vectors when computing softmax loss. Besides, for gaining a higher discriminative ability to recognize unseen person images, we further reinforce the irrelevance among the column vectors of the weight matrix via exclusivity regularization. The reason for regularizing the irrelevance is that we hope to get a feature vector with fewer redundancy elements in it. If every element in the feature vector is uncorrelated and hence representative for some specific attributes of a person image, the generalization ability of the feature encoding network is higher. To do so, for the weight matrix W in the softmax layer, we borrow the idea in improvement work on SVM [10] and penalize the $l_{1,2}$ of it. As shown in Fig. 1, after applying the normalization, the angles between the column vectors are larger and thus less correlated.

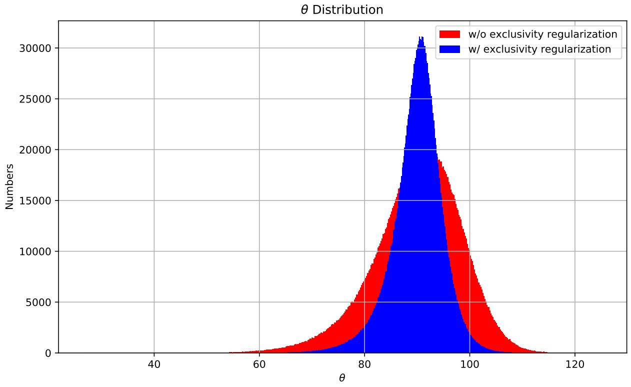


Figure 1: After applying the normalization, the angles between the column vectors are larger and thus less correlated. The figure is plotted according to our final models trained on Market-1501 with or without exclusivity regularization.

To sum up, our main contributions are as follows.

- To better capture the detailed cues in person images, we propose a novel multi-scale multi-patch network. The fused high-resolution feature map provides more details for learning better person representation.

- We propose a novel exclusively regularized softmax loss function for learning a more compact feature embedding space, which is used to match testing images in person re-ID tasks.
- To evaluate the effectiveness of our proposed method, we test it on three large-scale datasets under the standard settings and obtain very competitive results.

2. Related Works

2.1. Person Re-ID

With the help of deep learning, the person re-ID task has achieved notable progress. Recent works on person re-ID can be roughly divided into two types: 1) Some methods focus on designing a delicate model to make features contain sufficient information. Their proposed methods are mainly built on attention mechanism [5, 11], part-based [12, 3, 1] methods, attribute learning [13], and multi-level feature fusing methods [7, 14, 11]; 2) Other methods focus on designing an effective ranking loss [15, 16, 17, 18] to make features more discriminative. Both ways are trying to get a powerful feature extractor finally.

The main idea of designing a delicate model is to combine global features with local features extracted by deep models efficiently. IDE [19] is a decent baseline for many later researches. It learns a global feature from a given input image and ignores the local patterns. A number of later works use additional blocks to acquire local features via attention mechanism, part-based method, multi-level feature fusing method or the combination of the above methods. [11] utilizes attention mechanism on the Inception [20] model. In [11], hard regional attention, soft spatial attention, and channel attention are employed in a harmonious way. It combines both attention mechanisms and multi-scale methods, thus resulting in a notable effect. MLFN [7] focuses more on middle features which contain more detailed information. Besides, a feature selection step on the intermediate feature of multi-level is proposed in [7]. Similarly, SafeNet [14] utilizes multi-level features by a novel scale normalization block to balance features channels from different scales. SafeNet then connects scaled multi-level features together as the final encoder. JLML [21], PCB [1] and MGN [22] propose to slice the last feature map output from the backbone network into several parts. For each part, it is supervised by an independent classification loss. After the training finished, they concatenate features extracted by the sliced parts. All the mentioned methods above achieve considerable improvements by extracting extra information from the basic model and combining them with the global feature.

Apart from building a delicate deep model that is trained with a classification loss function, researches invent various objective functions to minimize the distance of intra-class samples while maximizing the distance of inter-class samples in feature embedding space. Treating data by triplets [15] inspires a variant of triplet loss into person re-ID. It takes a sampling way called online hard examples mining into the training stage, which makes the model be trained in an end-to-end way. The triplet loss mainly pays attention to obtain correct orders on the training set. Thus, it still suffers from a weaker generalization capability from the training set to the testing set. Quadruplet [17] takes a further step and samples several quadruplets instead of triplets in a mini-batch. In [17], they find that quadruplets can help the model generate outputs with a larger inter-class variation and a smaller intra-class variation when compared to the triplet loss. However, in Triplet[15] and Quadruplet [17], positive pairs, and negative pairs are sampled randomly in a mini-batch.

In this paper, we combine multi-level feature fusing and part-based methods in our deep model. Meanwhile, we add some modifications to the original classification loss function to make it have the ability to minimize the distance of intra-class samples while maximizing the distance of inter-class samples in a manner. Our work confirms the effectiveness of multi-scale multi-patch deep feature learning; to further enhance the deep features, some advanced technologies, such as deep ensemble learning [23], weakly-supervised learning [24], neural architecture search [25, 26, 27, 28, 29] and few-shot learning [30], can be adopted.

2.2. Normalized Softmax

The normalization technique is widely used in face recognition algorithms. In [31], they study the effect of normalizing both features and weights of the last fully-connected layer. In their formulation, a loss can be calculated by $\cos\theta$, where θ represents the angle between the feature vector and the column vectors of the weight matrix in the FC layer. In the following works [8, 9], they introduce a margin between normalized features and normalized weights. Specifically, loss can be computed by $\cos(m\theta)$ [8], $\cos(\theta + m)$ [9] and $\cos(\theta - m)$ [32] respectively. More recently, inspired by the normalization technique in face recognition, [33] improve the softmax via normalization, which is still different from ours due to the exclusivity used in our formulation.

2.3. Exclusivity Regularization

In [10], they first define the concept of (relaxed) exclusivity to manage the diversity between base learners in an ensemble framework. Besides, they extend the traditional primal SVM to a novel ensemble version named Exclusivity Regularized Machine. Besides, they derive a practical formulation of relaxed exclusivity and show that it can be converted into optimizing $l_{1,2}$ norm. For re-ID methods, $l_{1,2}$ norm is employed by [21] to enforce sparsity. They sparsify the global feature representation with a group LASSO. However, they impose the regularization on the embedding feature space and their motivation is to get a sparse feature vector. In [34], $l_{1,2}$ norm is applied to fuse multi-view predictions for robust and consistent performance. In our work, we impose the regularization on the weight matrix in the final FC (softmax) layer.

3. Our Proposed Method

In this section, we first describe our proposed multi-scale multi-patch network, which is capable of learning part-level features from different scales. Next, we introduce the normalized softmax loss as well as the variants used in our network. Also, we briefly describe the exclusivity regularization and how it is applied in our work.

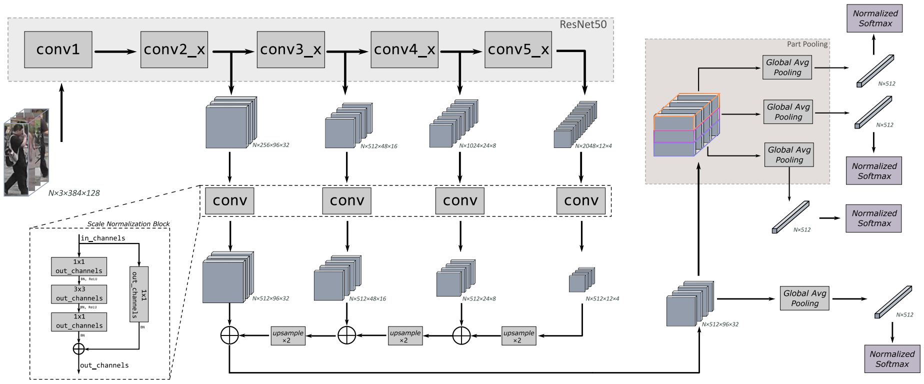
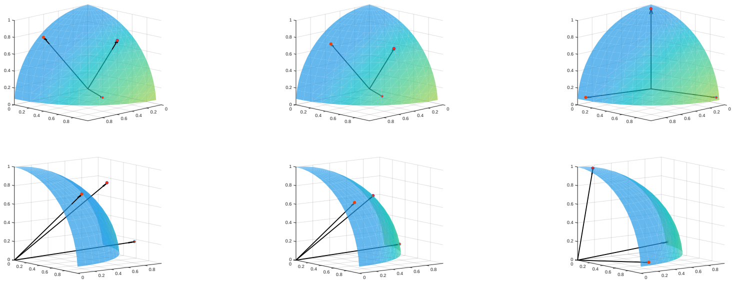


Figure 2: Network architecture of our proposed Multi-scale Multi-patch Network. We combine features from different scales together by first unifying the number of channels with the scale normalization block. Then we map the features to the same size via bilinear upsampling. After adding the features, we employ a global average pooling and part pooling to get the embedding vector. Finally, a normalized softmax layer and the cross-entropy loss are used for training the whole network.

3.1. Multi-scale Multi-patch Network

Our proposed network architecture is shown in Fig. 2. In the network, we use ResNet-50 [35] as backbone, which is employed by a number of state-of-the-art methods [22, 1]. Following the notations in ResNet, we extract feature maps with different scales from the network. Specifically, features from `conv2_x` (4× downsampling), `conv3_x` (8× downsampling), `conv4_x` (16× downsampling) and `conv5_x` (32× downsampling) are extracted. Since our motivation is to exploit features from different scale, we need to properly combine all these features together. Inspired by previous work [14] that efficiently merging features from different scales, we adopt their proposed scale normalization block in our network. However, in [14], their method is to concatenate those features, while we experimentally found that adding up those features achieves a better result when put multi-scale and multi-patch together. Thus, as shown in the figure, we remove ReLU layer before the output in the scale normalization block, due to the upsampling step.

After getting the feature that contains information from different scales, we then split the feature into two directions. The first direction is performing global average pooling directly on the whole feature map, which follows the vanilla re-ID networks that do not exploit part features. The second direction is to employ part pooling [1], which is a simple and strong baseline. All of the feature vectors after pooling are then connected to the normalized softmax layer.



(a) Softmax

(b) Normalized Softmax

(c) Normalized Softmax with Exclusivity Regularization

Figure 3: Illustration of the difference among vanilla softmax, normalized softmax with or without exclusivity regularization. After normalization, the learned features are projected onto the surface of a sphere. With exclusivity, the directions of projections tend to be less correlated. So the elements are more independent from each other and hence more representative.

3.2. Normalized Softmax and Exclusivity Regularization

3.2.1. Normalized Softmax

The softmax loss function is widely used in the re-ID task. It is computed by

$$L_{\text{softmax}} = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}. \tag{1}$$

where $x^i \in \mathbb{R}^d$ denotes the feature vector of the i-th sample with length d and x^i belongs to y_i-th class. $W_j \in \mathbb{R}^n$ denotes the j-th column of weights $W \in \mathbb{R}^{d\times n}$ in the classification layer and b_i is the bias of it. Besides, m and n represent mini-batch size and class number respectively.

Inspired by the success of feature normalization in face recognition algorithms [9, 31, 8], we add normalization on both feature and weights of the classification layer. Observe that $W_j^T x_i = \|W_j^T\|\|x_i\| \cos\theta_j$, where θ_j is the angle between feature vector

x_i and weight column vector W_j^T. After normalization on feature and weights, we have $W_j^T x_i = \cos\theta_j$. Then, $\|x_i\|$ is re-scaled to s. In this paper, we use $s = 15$. Thus, now the softmax loss function can be written as:

$$L_{\text{norm}} = -\frac{1}{m}\sum_{i=1}^{m} \log \frac{e^{s\cos\theta_{y_i}+b_{y_i}}}{e^{s\cos\theta_{y_i}+b_{y_i}} + \sum_{j=1,j\neq y_i}^{n} e^{s\theta+b_j}}. \tag{2}$$

3.2.2. Exclusivity Regularization

Recall that the output logits of softmax are computed by $l^i = W^T x^i$ and $l^i \in \mathbb{R}^n$. For each element j in l^i, we have $l_j^i = W_j^T x^i$. Then l_j^i can be viewed as the mapped value of x_i on the direction W_j^T. Thus, if we want to get a low redundant feature vector, any two directions, i.e., $W_{j_1}^T$ and $W_{j_2}^T$, are supposed to be less correlated as shown in Fig. 3. Since the column vectors W_j^T are normalized, the angle between two vectors can be expressed by $\langle W_{j_1}^T, W_{j_2}^T\rangle$. To this end, our target becomes

$$\min_{W} \sum_{j_1,j_2} \langle W_{j_1}^T, W_{j_2}^T\rangle. \tag{3}$$

In [10], they prove that the minimization problem (3) can be achieved by simply minimizing $\|W^T\|_{1,2}^T - \|W\|_F^2$. In our case, since we have already normalized each column vector in W^T, $\|W\|_F^2$ now becomes a constant and equals n^2. Therefore, our final loss function is

$$L = L_{\text{norm}} + \lambda \|W^T\|_{1,2}^T, \tag{4}$$

where λ is a tunable parameter.

4. Experiments

4.1. Datasets and Settings

In the experiments, three large scale datasets, i.e., Market-1501 [19], DukeMTMC-reID [36] and CUHK03 [37] are utilized for evaluating our method and comparing with the state-of-the-art methods. Specifically, we use the cumulative matching characteristics (CMC) and mean Average Precision (mAP) metrics. The details of the three datasets are given as follows.

Market-1501. [19] It contains 32,668 images of 1,501 identities captured by six camera views. The whole dataset is divided into a training set containing 12,936 images of 751 identities and a testing set containing 19,732 images of 750 identities. For each identity in the testing set, we select one image from each camera as a query image, forming 3,368 queries following the standard setting in [19].

CUHK03. [37] It contains 14,097 images of 1,467 identities. It provides the person bounding boxes detected using the deformable part model detector and the manually labeled person bounding boxes, which are called the detected dataset and the labeled dataset respectively. We conduct experiments both on the labeled dataset and the detected dataset. The dataset offers a 20-splits dividing, resulting in a training set with 1,367 identities and a testing set with 100 identities. Similar to [38], we also evaluate a division way with the training set of 767 identities and the testing set of 700 identities.

DukeMTMC-reID. [36] Similar to Market-1501, DukeMTMC-reID contains 36,411 images of 1,812 identities taken by 8 cameras, where only 1,404 identities appear in more than 2 cameras. The other 408 identities are regarded as distractors. The training set contains 16,522 images of 702 identities while the testing set contains 2,228 query images of 702 identities and 17,661 gallery images.

4.2. Implementation Details

We implement our method based on PyTorch [39]. For the backbone ResNet-50, we use the official ImageNet pretrained model. During training, we first resize training images to 384×128 for random crops and then the cropped person images are resized to 384×128 again, which is the same as the setting in [1]. Both random horizontal flip and random erasing [40] are used as data augmentation. Before feeding into our network, each image is subtracted by the mean value and divided by the standard deviation according to standard normalization procedure when using the pretrained model on ImageNet. We use Adam as our optimizer. The learning rate is set to 1e-3. It decays at epoch 70 as well as epoch 140 with a decay ratio of 0.1. The total training epoch is set to 150. The batchsize is set to 120. The PK sampling is used by sampling 15 images per identity and 8 identities per batch. In our multi-scale multi-patch network, 6 parts are used in part-pooling.

4.3. Comparisons with State-of-The-Art Methods

4.3.1. Competitors

We compared our methods against several existing state-of-the-art deep neural network based methods. At first, there are (a) four multi-scale methods (MLFN [7], SafeNet [14], HA-CNN [11] and Mancs [5]) and (b) two multi-patch methods (PCB [1] and MSCAN [3]). Then, we also compare our method with three recent representative works IDE [19], TriNet [15], CRAFT [12] and another work JLML [21] that also uses $l_{1,2}$ norm as regularization.

4.3.2. Evaluation on CUHK03

Tab. 1 shows the comparisons of our method against some existing methods on CUHK03. Note that several methods do not report their results under the setting of the 767/700 split. Therefore, we ignore these methods and only compare our method with the other methods with their reported results in their original paper. It is evident that our method outperforms existing methods in all categories on both labeled and detected bounding boxes. Our method surpasses recent multi-scale methods including HA-CNN, MLFN and Mancs, and a multi-patch method, i.e., PCB. Specifically, on the labeled set, our method outperforms the state-of-the-art method Mancs by 1.4% and 4.6% in Rank-1 and mAP respectively. Besides, on the detected set, the advantage of our method is more evident. On the detected set, for Rank-1 and mAP, our method is 4.6% and 6.7% higher than Mancs (multi-scale method) and 8.8% and 13% higher than PCB (multi-patch method) respectively.

4.3.3. Evaluation on Market-1501

We evaluated our proposed method on Market-1501 and the results are shown in Tab. 2. Tab. 2 shows the clear performance gain of the JLML method overall state-of-the-arts, but with less significant Rank-1 or mAP gains over other methods compared with the gains on the CUHK03 dataset. We think the reason is that Market-1501 is a

Table 1: Results on CUHK03. We use the 767/700 split setting.

Annotation	Labelled		Detected	
Metrics(%)	Rank-1	mAP	Rank-1	mAP
IDE(C) [19]	15.6	14.9	15.1	14.2
IDE(C)+XQDA [19]	21.9	20.0	21.1	19.0
IDE(R) [19]	22.2	21.0	21.3	19.7
IDE(R)+XQDA [19]	32.0	29.6	31.1	28.2
TriNet [15]+RE[40]	58.1	53.8	55.5	50.7
HA-CNN [11]	44.4	41.0	41.7	38.6
MLFN [7]	54.7	49.2	52.8	47.8
Mancs [5]	69.0	63.9	65.5	60.5
PCB [1]	-	-	61.3	54.2
Ours	70.4	68.5	70.1	67.2

simpler dataset and thus the results are kind of saturated. Compared with other methods, our method is competitive in both single query and multi-query settings. Specifically, on Market-1501, our method outperforms Mancs and PCB, which are two state-of-the-art methods dealing with local patterns.

4.3.4. Evaluation on DukeMTMC-reID

Similar to Market-1501, we compare our method with related methods and the results are depicted in Tab. 3. Our method performs better than other methods. Note that on DukeMTMC-reID, the multi-scale method SafeNet performs better than the multi-patch method PCB, which is different from the results on Market-1501. We presume that the difference reflects the intrinsic properties in a dataset. That is, in some datasets, those rare patterns can be better captured in a multi-scale manner, while in other datasets using multi-patch methods is more appropriate.

4.4. Hyperparameter Analysis

Since the hyperparameter λ plays a vital role in the whole loss function, we further examine the performance variation concerning different λ. As shown in Fig. 4, we

Table 2: Results on Market-1501.

Metrics(%)	Single Query		Multiple Query	
	Rank-1	mAP	Rank-1	mAP
CRAFT [12]	68.7	42.3	77.0	50.3
TriNet [15]	84.9	69.1	90.5	76.4
JLML [21]	85.1	65.5	89.7	74.5
SafeNet [14]	90.2	72.7	93.1	81.6
MLFN [7]	90.0	74.3	92.3	82.4
HA-CNN [11]	91.2	75.7	93.8	82.8
Mancs [5]	93.1	82.3	95.4	87.5
MSCAN [3]	76.2	53.1	84.0	62.9
PCB [1]	92.3	77.4	-	-
Ours	93.7	81.2	95.3	86.1

Table 3: Results on DukeMTMC-reID.

	Rank-1	mAP
PAN [41]	71.6	51.5
SVDNet [42]	76.7	56.8
Deep-Person [43]	80.9	64.8
SafeNet [14]	82.7	57.0
MLFN [7]	81.0	62.8
HA-CNN [11]	80.5	63.8
PCB [1]	81.8	66.1
Ours	84.4	70.4

Table 4: Evaluation of our proposed multi-scale multi-patch network. Ours best indicates the best results (with normalized softmax and exclusivity regularization) that are obtain in the setting of training and testing on the same dataset.

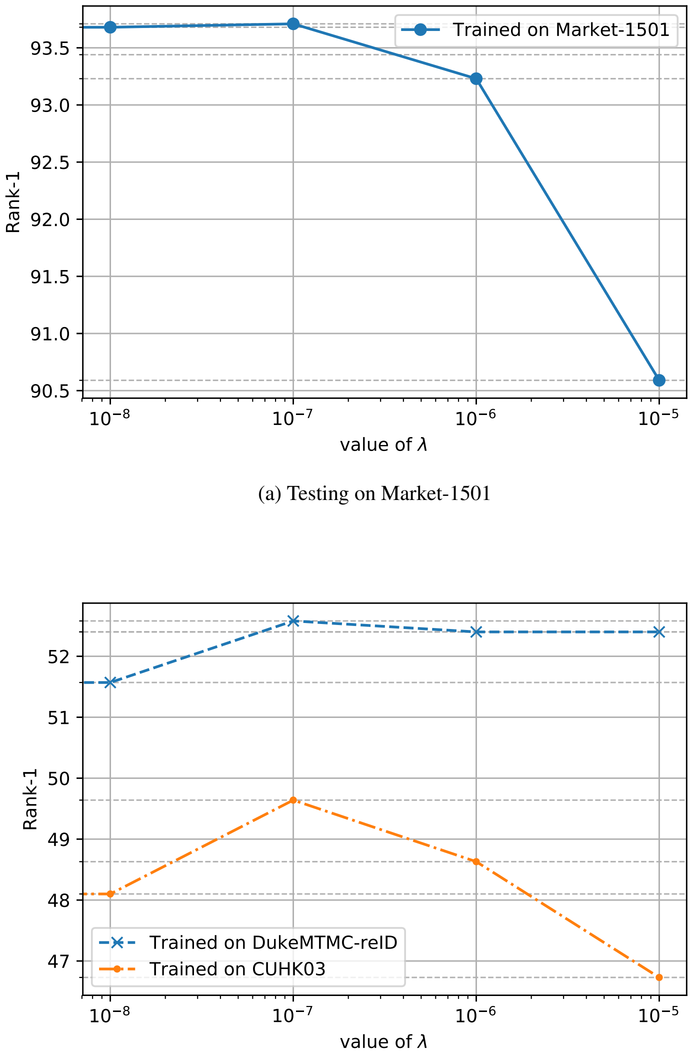
	Testing Dataset					
	Market-1501		DukeMTMC-reID		CUHK03	
Training Dataset	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Market-1501	92.5	78.4	22.4	12.2	2.4	3.2
DukeMTMC-reID	44.9	20.4	81.2	66.9	4.2	4.6
CUHK03	37.7	17.0	16.5	8.5	56.2	55.1
Ours best	93.7	81.2	84.4	70.4	70.1	67.2

train our network on Market-1501 and then draw the curve of Rank-1 accuracy with different λ on the three different datasets. We can observe the following truths: (1) Basically, setting λ to 1e-7 is a good practice; (2) When λ is too large, the accuracy drop on the original dataset is higher than that on other datasets, which shows that the regularization term helps to reduce the generalization error but increases the empirical error. (3) The effectiveness of adding the regularization term varies among different datasets. In other words, the best parameters for a dataset needs to be carefully tuned with the help of a well-annotated validation set, which could be an obstacle in person re-ID based applications. Besides, When λ is 0, our model can achieve 93.41% and 80.65% in rank1 and mAP respectively on Market-1501.

4.5. Ablation Study

4.5.1. Multi-scale Multi-patch Network with Softmax

Since there are two components in our method: the multi-scale multi-patch network and the exclusivity regularized normalized softmax. In this part, we switch the normalized softmax layer into the standard softmax layer to examine the performance of our proposed network alone. We train our network on Market-1501 and then report the results on the test sets of the three datasets. As shown in Tab. 4, our proposed multi-scale multi-patch network is also very effective with the standard softmax. Specifically, for Market-1501 we show the single gallery results and for CUHK03 we show the results



(a) Testing on Market-1501

(b) Testing on DukeMTMC and CUHK03

Figure 4: Rank-1 accuracy on different datasets with different λ values in the loss function Eqn. (4).

on detected images. Compared with our best results reported in the above section, we can observe the performance gains of our improved softmax. However, by comparing the results on different test sets, we see that the results of testing on other datasets are unsatisfying, which limits the potential value of application since person re-ID is usually applied in cross-domain scenarios. Besides, when we take out the part pooling component, our model can achieve 93.2% and 81.1% in rank1 and mAP respectively on Market-1501. When we take out the multi-scale part, the remaining model can achieve in 93.2% and 81.3% in rank1 and mAP respectively on Market-1501.

4.5.2. Effectiveness of the Normalized Softmax and the Exclusivity Regularization

Similar to previous experiments, we report the cross-domain testing results on the datasets that are different from the training dataset, to better examine the generalization ability of the learned model. By comparing the results in Tab. 4, it is obvious that the proposed normalization and exclusivity regularization can reliably help a deep neural network learn more generalized deep embeddings. For example, the Rank-1 and mAP on Market-1501 are improved from 92.5%/78.4% to 93.7%(+1.2%)/81.2%(+2.8%) respectively. On CUHK03, the improvement is more significant, i.e., 13.9%/12.1% gain in Rank-1 and mAP respectively.

5. Conclusion

In this paper, inspired by the recent progress of focusing on non-salient parts in images, we propose a network to fuse multi-scale and multi-patch features. Then, we introduce the well-established normalization techniques in face recognition algorithms, for further improving the softmax layer and thus getting a more representative feature embedding. In addition, we propose to penalize the redundancy in feature vectors via exclusivity regularization, which is achieved by minimizing the $l_{1,2}$ norm of the weight matrix in the softmax layer. Moreover, for better understanding the challenges in re-ID applications, we report cross-domain re-ID results.

Acknowledgement

We thank the anonymous reviewers for their very useful comments and suggestions. This work was supported in part by the National Natural Science Foundation of China (NSFC) (No. 61733007 and No. 61572207).

References

[1] 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: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 501–518.

[2] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose invariant embedding for deep person re-identification, arXiv preprint arXiv:1701.07732.

[3] D. Li, X. Chen, Z. Zhang, K. Huang, Learning deep context-aware features over body and latent parts for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 384–393.

[4] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, arXiv preprint arXiv:1611.03530.

[5] C. Wang, Q. Zhang, C. Huang, W. Liu, X. Wang, Mancs: A multi-task attentional network with curriculum sampling for person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 365–381.

[6] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2590–2600.

[7] X. Chang, T. M. Hospedales, T. Xiang, Multi-level factorisation net for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2018, pp. 2109–2118.

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2017, pp. 212–220.

[9] J. Deng, J. Guo, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, arXiv preprint arXiv:1801.07698.

[10] X. Guo, X. Wang, H. Ling, Exclusivity regularized machine: A new ensemble svm classifier, in: IJCAI-17, 2017, pp. 1739–1745.

[11] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2018, pp. 2285–2294.

[12] Y.-C. Chen, X. Zhu, W.-S. Zheng, J.-H. Lai, Person re-identification by camera correlation aware feature augmentation, T-PAMI 40 (2) (2018) 392–408.

[13] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recognition 95 (2019) 151–161.

[14] K. Yuan, Q. Zhang, C. Huang, S. Xiang, C. Pan, Safenet: Scale-normalization and anchor-based feature extraction network for person re-identification, in: IJCAI, 2018, pp. 1121–1127.

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

[16] Q. Xiao, H. Luo, C. Zhang, Margin sample mining loss: A deep learning based method for person re-identification, CoRR.

[17] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 403–412.

[18] W. Zhou, S. Newsam, C. Li, Z. Shao, Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval, Remote Sensing 9 (5) (2017) 489.

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116–1124.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI, Vol. 4, 2017, pp. 4278–4284.

[21] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: IJCAI, 2017, pp. 2194–2200.

[22] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM International Conference on Multimedia, ACM, 2018, pp. 274–282.

[23] X. Dong, L. Zheng, F. Ma, Y. Yang, D. Meng, Few-example object detection with model communication, IEEE transactions on pattern analysis and machine intelligence 41 (7) (2018) 1641–1654.

[24] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, A. L. Yuille, Pcl: Proposal cluster learning for weakly supervised object detection, IEEE transactions on pattern analysis and machine intelligence.

[25] X. Dong, Y. Yang, Searching for a robust neural architecture in four gpu hours, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1761–1770.

[26] X. Dong, Y. Yang, Network pruning via transformable architecture search, in: NeurIPS, 2019.

[27] X. Dong, Y. Yang, One-shot neural architecture search via self-evaluated template network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3681–3690.

[28] J. Fang, Y. Sun, Q. Zhang, Y. Li, W. Liu, X. Wang, Densely connected search space for more flexible neural architecture search, arXiv preprint arXiv:1906.09607.

[29] R. Quan, X. Dong, Y. Wu, L. Zhu, Y. Yang, Auto-reid: Searching for a part-aware convnet for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.

[30] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, Y. Yang, Progressive learning for person re-identification with one example, IEEE Transactions on Image Processing 28 (6) (2019) 2872–2881.

[31] F. Wang, X. Xiang, J. Cheng, A. L. Yuille, Normface: L2 hypersphere embedding for face verification, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM, 2017, pp. 1041–1049.

[32] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, W. Liu, Cosface: Large margin cosine loss for deep face recognition, arXiv preprint arXiv:1801.09414.

[33] X. Fan, W. Jiang, H. Luo, M. Fei, SphereReID: Deep Hypersphere Manifold Embedding for Person Re-Identification, arXiv preprint arXiv:1807.00537.

[34] X. Dong, Y. Yan, M. Tan, Y. Yang, I. W. Tsang, Late fusion via subspace search with consistency preservation, IEEE Transactions on Image Processing 28 (1) (2018) 518–528.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[36] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3774–3782.

[37] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 152–159.

[38] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3652–3661.

[39] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: NIPS Autodiff Workshop, 2017.

[40] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, arXiv preprint arXiv:1708.04896.

[41] Z. Zheng, L. Zheng, Y. Yang, Pedestrian Alignment Network for Large-scale Person Re-identification, arXiv preprint arXiv:1707.00408.

[42] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3820–3828.

[43] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-person: Learning discriminative deep features for person re-identification, arXiv preprint arXiv:1711.10658.