Dense-Scale Feature Learning in Person Re-Identification

Li Wang[1,2,3], Baoyu Fan[1,2,*], Zhenhua Guo[1,2], Yaqian Zhao[1,2]
Runze Zhang[1,2], Rengang Li[1,2], Weifeng Gong[1,2]

[1] Inspur Electronic Information Industry Co.,Ltd.
[2] State Key Laboratory of High-end Server Storage Technology
[3] College of Control Science and Engineering, China University of Petroleum(East China), Qingdao, China
{li.wang.upc}@foxmail.com,
{fanbaoyu,guozhenhua,zhaoyaqian,zhangrunze,lirg,gongwf}@inspur.com

Abstract. For mass pedestrians re-identification (Re-ID), models must be capable of representing extremely complex and diverse multi-scale features. However, existing models only learn limited multi-scale features in a multi-branches manner, and directly expanding the number of scale branches for more scales will confuse the discrimination and affect performance. Because for a specific input image, there are a few scale features that are critical. In order to fulfill vast scale representation for person Re-ID and solve the contradiction of excessive scale declining performance, we proposed a novel Dense-Scale Feature Learning Network (DSLNet) which consist of two core components: Dense Connection Group (DCG) for providing abundant scale features, and Channel-Wise Scale Selection (CSS) module for dynamic select the most discriminative scale features to each input image. DCG is composed of a densely connected convolutional stream. The receptive field gradually increases as the feature flows along the convolution stream. Dense shortcut connections provide much more fused multi-scale features than existing methods. CSS is a novel attention module different from any existing model which calculates attention along the branch direction. By enhancing or suppressing specific scale branches, truly channel-wised multi-scale selection is realized. To the best of our knowledge, DSLNet is most lightweight and achieves state-of-the-art performance among lightweight models on four commonly used Re-ID datasets, surpassing most large-scale models.

1 Introduction

Person re-identification (Re-ID) intends to automatically identify an individual across non-overlapping camera views deployed at different times and locations. The pedestrian image varies dramatically between different cameras, due to the complexity of the realistic environment, such as light modification, posture variability, variety of view, scale change, partial occlusion. So, how to obtain and

[*] Corresponding author.

select the most beneficial multi-scale features becomes critical to improve the performance of person re-identification.

The pedestrian image contains a plethora of multi-scale information. For example, for the pedestrian in Figure 1(a), their dressing is very similar (both wear grey top and black shorts), but their somatotype is not identical. So, the large-scale body features become a crucial discriminating factor in this case. For the pedestrians in Figure 1(b), they are very similar in dress and bodily form, but their shoes and shorts are varied. So, small-scale information has become a vital identification factor.

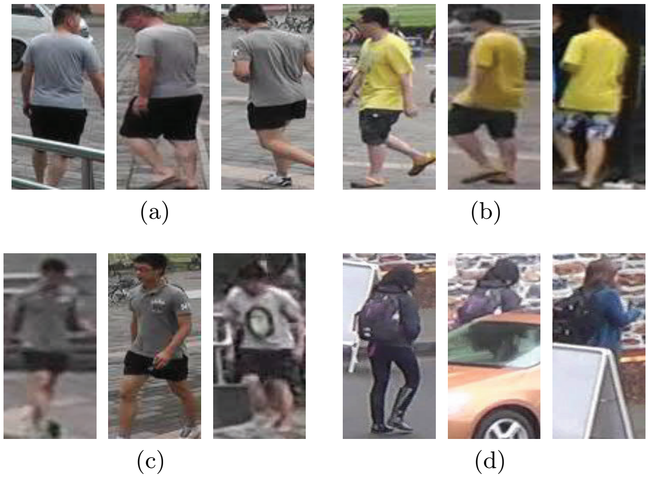
(a)
(b)
(c)
(d)

Fig. 1: The pedestrian image contains a plethora of multi-scale information. Larger-scale features and smaller-scale features can identify pedestrians in subfigure (a) and (b). However, only multiple-scales are not enough, and the fusion of multi-scale information is more discriminative (T-shirt with logo features that are more discriminative in subfigure (c)). Moreover, as multi-scale information becomes more abundant, selecting the discriminative information becomes more important(as shown in sub-figure (d), how key features can be filtered under occluded conditions becomes critical).

Under many circumstances, only multiple-scales are not enough and the person can only be successfully matched through fusion between multi-scale information. For example, for the pedestrian in Figure 1(c), the impostor needs to be distinguished through the logo on T-shirt, and without T-shirt as context, the logo no longer has discriminating power. Therefore, only by combining a variety of scale information can the most favourable pedestrian features be obtained.

Furthermore, besides the ability to obtain pedestrian features at multi-scale scales, it is also needed to select more discriminative information from an enormous number of scale features and remove redundant features. As shown in

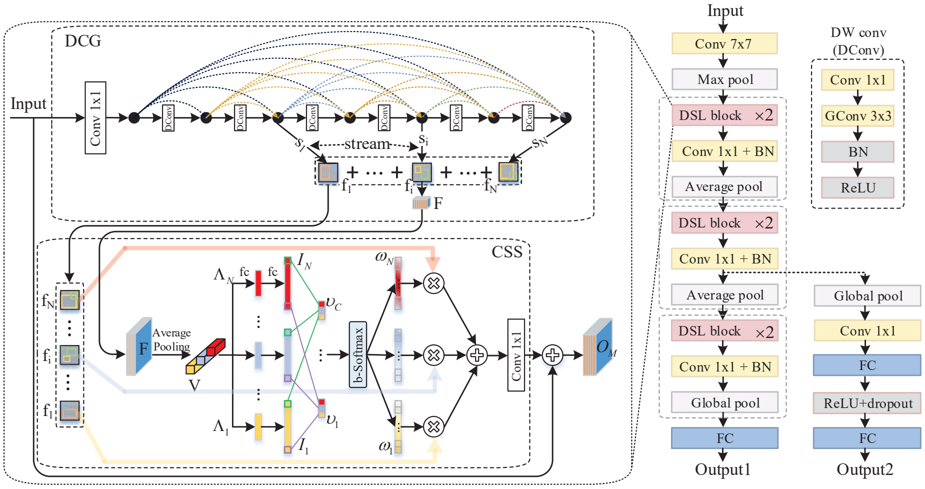

Fig. 2: The structure of dense-scale feature learning network.

Figure 1(d), the leftmost pedestrian image is not occluded and the pedestrian image in the middle is partially occluded, so the backpacks and tops of pedestrians become crucial identification information. It is of critical importance to filter features and remove redundant features for Re-ID tasks.

To solve the problem of multi-scale acquisition and effective scale selection, many of scholars put forward various solutions. In terms of multi-scale learning, [1, 2] that improve the multi-scale ability by using filters with different size, however the capacity will be significantly limited only by changing the convolution kernels size. OSNet [3] has designed a multi-branch block to obtain features with different scales. However, the multi-scale learning ability is greatly affected by the number of branches. A few attempts at learning multi-scale features also exist [4, 1]. Yet, none has proposed a good solution to fully learn effective scale information.

In terms of scale selection, SENet [5] adopts the strategy of feature recalibration to suppress or enhance the corresponding channel in the feature map. OSNet designed an attention structure (Aggregation Gate) as same as SENet, which applies it to multiple independent branches. However, the Aggregation Gate is actually trained to identify critical channels under the intra-scale features, and inter-scale features selection among multi-branches is implicitly realised only by sharing aggregation gate weights. However, for person re-identification, intra-scale feature selection would not be enough, and inter-scale feature selection is also critical because integrate multiple key scales features as contextual information can enhance the discrimination.

To this end, we proposed the dense-scale feature learning network (DSLNet), a simple yet efficient multi-scale representation network which can extract more luxurious scale features while obtaining the most favourable scale features, and it is also extremely lightweight.

As shown in Figure 2, The core of the dense-scale feature learning network is the DSL block which consists of two parts: Dense Connection group (DCG) and the Channel-Wise Scale Selection module (CSS).

In DCG, multi-scale feature mining is realized through the stack of depthwise separable convolutions, and multi-scale feature fusion is realized through dense connection. Additionally, DCG introduces multiple streams and outputs the corresponding feature maps which provide an enormous number of scale features.

The structure of DCG has the following three advantages: First, provide rich multi-scale fusion features. The small scale feature information can flow to the large scale feature in a dense style, and then fuse information through addition, which provides multi-scale fusion information with the largest capacity. Secondly, output fusion information in a more sophisticated way. Different convolution layers contain various types of fusion features. DCG attaches multiple streams and outputs feature maps in early layers which can provide more abundant multi-scale fusion features in a more advanced way. Thirdly, being lightweight. DCG uses deep separable convolution (DW conv) which reduces the parameters under the premise of preserving the accuracy, and the dense connection style also makes more effective use of the stacked convolution layers.

DCG provides rich multi-scale features also introduces a lot of noise information. Too rich scale features without a selection will confuse the discrimination. So we designed the channel-wise scale selection module (CSS) which dynamically select the inter-scale features in multiple streams. As shown in Figure 2, CSS learns the weighted attention with fused multi-scale features because the weights of each scale should not be calculated independently. Second, establish independent sub-network for each stream to learn the corresponding channel weights. Third, the truly inter-scale channel selection mechanism. Different from [3, 5], CSS changes sigmoid activation for each stream with a unified softmax activation function and calculates the softmax attention among corresponding channels in different streams. This change will lead to a realization of inter-scale feature selection. The benefit of these changes will be further discussed in the experiment section.

Additionally, we go one step further. We proposed an enhanced hard triplet loss (E-TriHard), which reduces the distance gap between intra-class and interclass while reducing the absolute value of the inter-class. The experimental results (as shown in Table 3) verify that using E-TriHard loss during the fine-tuning phase will further improve the performance of DSLNet.

2 RELATED WORK

2.1 Deep Person Re-ID

In recent years, the person re-identification algorithm based on deep learning has achieved state-of-the-art performance and become a critical research direction.

Person re-identification can be divided into two steps: representation learning and metric learning. They are closely related and rely on each other.

In [6–9], feature learning methods based on body parts are proposed to learn salient pedestrian features. However, this kind of approach is restricted due to the limitation of variable body parts spatial distribution and features misalignment problem. The global feature learning methods are developed by [10–12]. However, the global feature tends to overlook detailed information about the image, which restricts the expression ability. In order to obtain more discriminative information, the multi-scale feature representation method is proposed. [13, 3] achieve multi-scale feature mining by designing multiple branching network structures. However, only stacking branches or changing convolution kernels has limited capacity.

For metric learning, the main research direction focuses on optimizing the loss function, including Contrastive Loss [14], Triplet Loss [15], Quadruplet Loss [16], lifted loss [17], and so on. The development of the hard sample mining methods, including hard triplet mining [18], margin sample mining [19], has gained great success in person Re-ID and attracted more attention.

In this paper, we proposed a dense-scale representation network. Compared with the above representation learning methods, DSLNet can learn more rich multi-scale features in a granular way through dense-scale feature fusion and reuse. In addition, we proposed an enhanced hard triplet loss function, which can enlarge the gap between the intra-class and inter-class while keeping the absolute value of the intra-class distance low.

2.2 Attention Mechanisms

With the rise and development of deep learning, attention model is widely used in various fields, such as natural language processing [20], action recognition [21]. The attention-driven approaches to power networks to acquire discriminative human representation are thus received widespread attention.

In [22–25], spatial attention modules are proposed to learn attention regions or to extract features at salient spatial locations. However, spatial attention methods depend on the specific network model structure and have weak generality. Channel-wise attention modules are proposed to enhance the channel-wise feature map representation. Squeeze-and-excitation network [5] compresses feature maps according to the channel-wise dimension to learn the discriminative representation of channels. OSNet [3] takes a similar attention structure to SENet and pushes it into multiple branches. In this paper, we propose a novel channel-wise scale selection module that dynamically weights the discriminative features in multiple streams from the fused dense-scale information.

3 Proposed Algorithm

In this section, we present DSLNet, a novel and efficient multi-scale learning neural network for the person Re-ID task. Firstly, we elaborate on the important components of DSLNet basic block(DSL block). Then we present the detailed architecture of DSLNet.

3.1 Dense Connection Group

In the person Re-ID task, only multiple scales are far from enough, and the feature can be more discriminative through fusion between multi-scale information.

To achieve multi-scale feature fusion, we extend the residual structure by introducing a dense connection group (DCG), The structure of the DCG is shown in Figure 2.

Given an input x_0, DCG consists of L feature extractors, and each feature extractor implements a nonlinear transformation $H_i(\cdot)$, the output of DCG can be expressed as follows:

$$x_i = H_i(x_0 + x_1 + \cdots\cdots, x_{i-1}) \quad i \in [1, L] \tag{1}$$

where x_i represents the output of DCG on various scales. Each filter will fuse the output of the former, and the receptive field will increase while the multi-scale features are merging. Due to the dense connection, each layer contains the information of all previous layers, which implements feature fusion with rich scales.

Moreover, to better balance the trade-off between scale mining capability and calculation cost, we carefully select the number of filters. Meanwhile, DSL block adopts the DW conv, further reducing the FLOPS and parameters while maintaining the accuracy.

3.2 Channel-Wise Scale Selection module

The dense connection group provides extremely rich multi-scale fusion features also introduces many redundant features. In fact, only a limited number of scales may be needed for the final authentication. So, how to extract the most discriminative features from the rich scale information is very critical. Therefore, we introduce the attention mechanism to realize the discriminative multi-scale feature selection.

As shown in Figure 2, the dense-scale channel-wise selection module (CSS) consists of three stages: dense-scale feature fusion, attention learning, channel-wise feature selection.

In the first stage, CSS obtains dense-scale features from the dense connection group (DCG) and then adopts element-wise addition to fuse the features. Next, CSS use the global average pooling layer to compress each channel. The process can be formulated as:

$$F = \sum_{i=1}^{N} s_i, F \in \mathbb{R}^{H \times W \times C} \tag{2}$$

$$\mathrm{V}_c = \sum_{w=1}^{W} \sum_{\mathrm{h}=1}^{H} F_c\left(i,j\right) / (H \times W) \tag{3}$$

where N represents the number of output streams in DCG, s_i represents the feature map of the $i-th$ stream in DCG. F stands for fused features. H and W denotes the width and height of F, c represents the $c-th$ channel of F. V represents the channel weight statistics and $V = [V_1, \cdot\cdot\cdot, V_c, \cdot\cdot\cdot V_C]$.

For attention learning, CSS learns the channel weights for each DCG stream from the vector V. First, the vector V is split into N groups, then the CSS module establishes multiple sub-streams that contain two fully connected layers. $\mathit{\Lambda}_n$ will pass through the corresponding sub-stream to learn the channel weights I_n.

$$I_n = \varphi_n(\mathit{\Lambda}_n) \tag{4}$$

where $\varphi_n(\cdot)$ represents the mapping function formed by two fully connected layers, and n represents the $n-th$ stream in CSS. I represents the set of channel weights for all DCG streams, $I = [I_1, \cdot\cdot\cdot I_n, \cdot\cdot\cdot, I_N]$.

For channel-wise feature selection, CSS recombinants and normalizes the corresponding channel weights in I, so as to truly realize inter-scale feature selection. Specifically, CSS traverses I to extract the corresponding elements in the I_n and reconstructs a new vector υ (as shown in Figure 2), $\upsilon \in \mathbb{R}^{1\times N}$, and uses softmax layer to normalize υ. Finally, υ will replace the corresponding elements in I_n and get the renewed channel weight vector ω_n, $\omega_n \in \mathbb{R}^{1\times C}$. The process of recombining elements and normalization enables inter-scale feature selection, and all weight expressions are derived from the dense-scale fusion features.

Finally, the reconstruction weight ω_n is multiplied from the features of the corresponding CSS stream, the weighted features of all streams are finally fused, the output of the CSS can be formulated below:

$$D = \sum_{\mathrm{n}=1}^{N} \omega_n \odot F_n \tag{5}$$

where $\odot$ denotes the Hadamard product and D denotes the final learned weighted features.

3.3 Loss Function Design

In the process of training, loss function can supervise the learning of network, thus affecting the recognition performance of the model. Therefore, the selection and design of loss function plays an important role in image retrieval [26], face recognition [27] and person re-identification [28, 29]. We use two training methods: trained from scratch and fine-tuning from the ImageNet pre-trained models to evaluate the proposed algorithm. Different training processes use different loss functions.

For training from scratch, we use cross entropy loss to optimize the model. For fine-tuning, firstly, we use the cross entropy loss to train the model. Secondly, we optimize the hard triplet loss function and train the network with the improved loss function.

The hard triplet loss function can be expressed as:

$$L_{TriHard} = \frac{1}{N_t}\sum_{a=1}^{N_t}[\max_{y_p=y_a} d(f_a, f_p) - \min_{y_n\neq y_a} d(f_a, f_n) + m]_+ \tag{6}$$

where $[\cdot]_+$ represents $\max(\cdot, 0)$, N_t represents the number of triples in each batch, $d(\cdot,\cdot)$ stands for metric distance function and we adopt the Euclidean distance. m represents the margin of the hard triplet loss. f_p is positive sample features, f_n is negative sample features, f_a represents anchor sample features, $dist^a_{ap} = \max\limits_{y_p=y_a} d(f_a, f_p)$, $dist^a_{ap}$ represents the maximum intra-class distance of anchor samples. $dist^a_{an} = \min\limits_{y_n\neq y_a} d(f_a, f_n)$, $dist^a_{an}$ represents the minimum inter-class distance of anchor samples.

The hard triplet loss is widely used in person re-identification, however, it only considers the distance gap between $d(f_a, f_p)$ and $d(f_a, f_n)$ and ignores their absolute values [30]. To compensate for the drawbacks of the hard triplet loss, we add a regularization term to reduce the distance gap between intra-class and inter-class while reducing the absolute value of the inter-class. The enhanced hard triplet loss function(E-TriHard) is formulated as follows:

$$L_{E\_TriHard} = L_{TriHard} + \beta\frac{1}{N_t}\sum_{a=1}^{N_t}(\frac{dist^a_{ap}}{dist^a_{an}}) \tag{7}$$

The final loss function used in the training process is formulated as:

$$L_{final} = L_{soft\max} + \alpha L_{E\_TriHard} \tag{8}$$

where, $\alpha, \beta \in (0, 1]$, two hyper-parameters α and β are fixed in the experiments.

3.4 Network Architecture

Based on the DSLBlock, the structure of DSLNet is meticulously designed. As shown in Table 1, the stem of DSLNet is composed of a 7 × 7 convolution layer with a stride of 2 and a 3 × 3 maxpooling layer. Subsequently, we stack the DSLBlock layer-by-layer to construct the DSLNet. DSLNet contains three stages, with each stage including two DSL Blocks. The feature map can be downsampled with each stage. Additionally, during the training phase, we add an auxiliary branch at the second stage of the network. Auxiliary branch facilitates information flow to the early layers and relieves the gradient vanishing problem. Finally, we add global average pooling and fully-connected layer for training. The network structure and parameters of DSLNet are shown in Table 1.

Table 1: Architecture of DSLNet with input image size 256 × 128. c: The number of input channels. k: Operation type. s: Stride. n: Number of repetitions of this layer operation

Output	Layer	c	k	s	n
128 × 64	conv2d	60	7 × 7	2	1
64 × 32	max pool	60	3 × 3	2	1
64 × 32	DSL block	252	DW conv	1	2
64 × 32	conv2d	252	1 × 1	1	1
32 × 16	average pool	252	2 × 2	2	1
32 × 16	DSL block	384	DW conv	1	2
32 × 16	conv2d	384	1 × 1	1	1
16 × 8	average pool	384	2 × 2	2	1
16 × 8	DSL block	516	DW conv	1	2
16 × 8	conv2d	516	1 × 1	1	1
1 × 1	global average pool	516	16 × 8	1	1
1 × 1	fc	516	fc	1	1
Params	1.9M				
Flops	825.0M				

4 Experimental Results

4.1 Datasets and Evaluation Metrics

Four mainstream challenging Re-ID datasets are used to verify the proposed model, including Market-1501 [Zheng et al., 2015], DukeMTMC-Re-ID [Ristani et al., 2016], MSMT17 [Wei et al., 2018] and CUHK03 [Li et al., 2014]. Among them, Market-1501 includes 32,668 images of 1501 pedestrians, of which, 12,396 images of 751 identities were used for training and the rest for testing. Duke MTMC-Re-ID consists of 36,411 images of 1,812 identities, of them, 1,404 identities were captured by more than two cameras, and the rest only appeared in only one camera. Compared to other datasets, MSMT17 is a larger and more realistic Re-ID dataset which was published in 2018, it contains 126,411 pedestrian images of 4101 identities, 32621 of these images with 1041 identities were selected as the training set, and the rest 93820 images with 3060 other identities were used for testing. The CUHK03 dataset is composed of 14,097 pedestrian images of 1,467 identities, with each person having 9.6 images on average, and the CUHK03 dataset contains two subsets that provide hand-labeled and DPM detected bounding boxes respectively. We evaluate our proposed model on DPM detected subset.

4.2 Implementation Details

In our experiments, we employ two training methods: training from scratch and fine-tuning from the ImageNet pre-trained model.

For training from scratch, we use the stochastic gradient descent algorithm to optimize the model and epoch is set to 350. The learning rate is decayed using the cosine annealing strategy with the initialization value of 0.0015. In the fine-tuning stage, the AMS-Grad optimizer is used. The pre-trained weight is frozen at the first 10 epochs, and only the randomly initialized classifier can be trained. The epoch is set to 250. The learning rate decay strategy adopts the cosine annealing strategy, and the initial learning rate is 0.065. Two loss functions, cross-entropy loss and E-TriHard loss, are used in the fine-tuning stage. For E-TriHard loss, we set the hyperparameters α and β to 0.5 and 0.8, respectively.

In all experiments, the batch size is set to 64, and the weight decay is set to 5e-4. Images are resized to 256×128, and the corresponding data enhancement methods are adopted. In the verification stage, we delete the auxiliary branch and extract the 512-D features from the last fully-connected layer of the main branch and use the cosine distance for measurement. For all experiments, we use single query evaluation and simultaneously adopt both Rank-1 (R1) accuracy and the mean average precision (mAP) to evaluate the performance of DSLNet. All experiments are conducted based on the deep learning framework of PyTorch, and we use NVIDIA V100 GPU to train the model.

4.3 Performance Evaluation

Trained from Scratch Based on DSL block, we build a lightweight DSLNet which can obtain dense-scale information and realize channel-wise scale feature selection. For each DSL block, we stack six depth-wise separable convolutions in series to obtain various scale receptive fields and simultaneously add dense connections in DSL block for the fuse of multi-scale features at a granular level. We trained the proposed model from scratch and compared it with state-of-the-art models using the same training strategy. The results are shown in Table 2.

From Table 2, we can see that DSLNet outperforms the other methods in all datasets. More concretely, DSLNet achieves the best rank-1 value and mAP accuracy of 94.0% and 83.9% in Market1501 datasets, and 86.9%/74.8% on Duke. While OSNet, the second-best method, arrives at 93.6%/81.0% and 84.7%/68.6%, respectively. The gap is even more significant in the CUHK03 and MSMT17 databases. For R1, DSLNet outperforms OSNet by more than 6% improvement of R1 rate on CUHK03. For mAP, DSLNet beats OSNet by 6.8% on CUHK03 and 9.0% on MSMT17.

Furthermore, DSLNet is the most light-weighted model which only has 1.9M parameters. OSNet has similar parameter amount to MobileNetV2, both of which are 2.2M. DSLNet has created an elegant and effective backbone network, which uses fewer parameters to achieve the best performance.

Fine-tuning from ImageNet In order to highlight the significance of the proposed DSLNet for person Re-ID task, we compare it with some recent remarkable works. We conduct pre-training of DSLNet on the ImageNet dataset

Table 2: **Trained from scratch.**

Method	Venue	Params(M)	GFLOPs	Duke		Market1501	
				R1	mAP	R1	mAP
MobileNetV2 [31]	CVPR'18	2.2	**0.2**	75.2	55.8	87.0	69.5
BraidNet [32]	CVPR'18	-	-	76.4	59.5	83.7	69.5
HAN [28]	CVPR'18	2.7	1.09	80.5	63.8	91.2	75.7
OSNet [33]	ICCV'19	2.2	0.98	84.7	68.6	93.6	81.0
DSLNet	***ours***	**1.9**	0.82	**86.9**	**74.8**	**94.0**	**83.9**

Method	Venue	Params(M)	GFLOPs	CUHK03		MSMT17	
				R1	mAP	R1	mAP
MobileNetV2 [31]	CVPR'18	2.2	**0.2**	46.5	46.0	50.9	27.0
HAN [28]	CVPR'18	2.7	1.09	41.7	38.6	-	-
OSNet [33]	ICCV'19	2.2	0.98	57.1	54.2	71.0	43.3
DSLNet	***ours***	**1.9**	0.82	**63.8**	**61.0**	**75.0**	**52.3**

and then use the pre-trained weight to conduct fine-tuning on the Re-ID dataset. All results are summarized in Table 3.

From Table 3, it can be seen that DSLNet achieves higher R1/mAP than the other methods on four mainstream datasets. For R1, DSLNet achieves the highest rate of 73.6% on CUHK03 and 89.5% on Duke, while OSNet arrives at the rate of 69.1%/88.6% respectively. For mAP, DSLNet beats OSNet by 3.7% on CUHK03 and 3.6% on Duke. On MSMT17, which is the largest one among the four commonly used Re-ID datasets, DSLNet outperforms OSNet by a significant margin. Concretely, DSLNet achieves the R1/mAP of 80.2%/57.3%, respectively, while OSNet just arrives at 78.7%/52.9%. Adding E-TriHard loss to the training process will further improve the performance of DSLNet. On Market1501 and Duke, DSLNet (E-TriHard) achieves the R1/mAP of 95.1%/87.3% and 90.4%/78.5%. On CUHK03 and MSMT17, DSLNet (E-TriHard) arrives at 76.8%/72.4% and 82.1%/59.4%. The performance on Re-ID benchmarks, especially on Market1501 and Duke, has been saturated lately. Therefore, the improvements obtained are significant.

Furthermore, we can also see that DSLNet achieves the best results with the smallest model. DGNet, IANet, CAMA, st-ReID adopted the backbone based on ResNet50, which involved the parameters amount of more than 23.5M, VA-reID [37] employed the SeResNeXt backbone network with the parameter amount of more than 46M. In comparison, our model is dozens of times smaller than theirs. These experimental results validate the efficiency and robustness of DSLNet, which is due to the multi-scale feature extraction and fusion ability of DSLNet.

4.4 Ablation Study

In order to verify the influence of different components of DSLNet, we conduct related ablation experiment on the CUHK03 dataset. We verify the influence of DCG/CSS components on the performance, respectively.

Table 3: **Fine-tuning from ImageNet**. †: reproduced by us.

Method	Venue	Backbone	Params (M)	Duke		Market1501	
				R1	mAP	R1	mAP
IANet [34]	CVPR'19	ResNet	> 23.5	87.1	73.4	94.4	83.1
DGNet [35]	CVPR'19	ResNet	> 23.5	86.6	74.8	94.8	86.0
st-ReID [36]	AAAI'19	ResNet	> 23.5	**94.4**	**83.9**	**98.1**	**87.6**
VA-reID [37]	AAAI'20	SeResNeXt	> 46.9	91.6	84.5	96.2	91.7
LUO [30]	TMM'19	ResNext	46.9	90.1	79.1	95.0	88.2
Auto-ReID [38]	ICCV'19	ResNet	13.1	-	-	94.5	85.1
OSNet [33]	ICCV'19	OSNet	2.2	88.6	73.5	94.8	84.9
DSLNet	***ours***	DSLNet	**1.9**	89.5	77.1	94.5	85.1
DSLNet *(E-TriHard)*	***ours***	DSLNet	**1.9**	90.4	78.5	95.1	87.3
Method	Venue	Backbone	Params(M)	CUHK03		MSMT17	
				R1	mAP	R1	mAP
CAMA [39]	CVPR'19	ResNet	> 23.5	66.6	64.2	-	-
IANet [34]	CVPR'19	ResNet	> 23.5	-	-	75.5	46.8
DGNet [35]	CVPR'19	ResNet	> 23.5	65.6	61.1	77.2	52.3
Auto-ReID [38]	ICCV'19	ResNet	13.1	73.3	69.3	78.2	52.5
OSNet [33]	ICCV'19	OSNet	2.2	$69.1^{\dagger}$	$65.7^{\dagger}$	78.7	52.9
DSLNet	***ours***	DSLNet	**1.9**	73.6	69.4	80.2	57.3
DSLNet *(E-TriHard)*	***ours***	DSLNet	**1.9**	**76.8**	**72.4**	**82.1**	**59.4**

Validity of Dense Connection The dense connection group obtains different scale receptive fields by convolution layer stacking and realizes contextal information fusion by dense connection. To verify the effectiveness of the dense connection in DCG, we remove all dense connections from DSL blocks in ablation experiment and compare the impact on the final performance. Baseline stands for delete all dense connections in DSLNet. Add DC means adding dense connection to DSL blocks.

As shown in Table 4, by adding dense connections, the performance of the model is improved significantly, which benefits from the reuse of multi-scale features, and the fusion of context information. Add DC outperforms baseline model by more than 2% improvement of R1 rate and 1.8% improvement on mAP accuracy. It proves that the design of dense connection is reasonable and effective.

Validity of CSS Too rich features without an effective selection mechanism can undermine the classifier's discriminatory ability. Attention mechanism-based approaches can be effective in addressing the problem of feature selection. To verify CSS's validity, we introduce the attention model from [3, 5] into DSLNet for comparison experiments. As shown in Table 5, Baseline stands for removing CSS module from DSL blocks. SENet Attention represents adding the SENet attention structure to the baseline model. OSNet Attention means adding independent attention models to the baseline model for all streams, yet the weights of all attention models are shared. CSS is our proposed approach.

Table 4: Validity of Dense Connection

Model	Architecture	CUHK03	
		R1	mAP
1	Baseline	61.4	59.2
2	Add DC	**63.8**	**61.0**

Table 5: Validity of CSS

Model	Architecture	CUHK03	
		R1	mAP
1	Baseline	56.2	54.2
2	SENet Attention	62.2	60
3	OSNet Attention	62.5	59.9
4	CSS	**63.8**	**61.0**

From the experimental results, we can find that: 1) Adding the attention network will significantly improve the baseline model's performance, validate that the attention mechanism-based feature selection mode is essential for Re-ID tasks. 2) CSS model outperforms all other attention benchmarks by a clear margin. Compared with model 2 and model 3, with the introduction of CSS, R1/mAP can be improved by 1.6%/1.0% and 1.3%/1.1%, respectively. CSS fuses the features from multiple streams to dynamically adjust the weights of all channels in the DCG, truly realizing the role of channel-wise feature selection across streams. The experimental results verify the superiority of the CSS method.

4.5 Visualizations

Visualization of Learned Features To validate the effectiveness of DSLNet to represent and select multi-scale features, we extract the feature map of the last convolutional layer for visualization and observe whether the DSLNet focuses more on the key regions. We use the visualization method of [40], which summits the feature maps along the channel dimension and then performs a spatial euclidean normalization for a bright feature display. As shown in Figure 3, the rightmost column represents the DSLNet activation map. The middle column represents the activation map with all dense connections removed from the DSLNet. We can see that DSLNet mines more effective multi-scale information, as shown in the second example on the first line, where the logo on the pedestrian bag is activated, in the second example on the second line, where the pedestrian's shoes and handbag activate a larger and more pronounced area. Other examples in Figure 3 also show that DSLNet highlights more salient features of the same target. DCG provides the abundant of multi-scale feature combinations, and CSS modules enable efficient discriminating feature selection. These qualitative results confirm the ability of DSLNet for effective feature representation and selection.

4.6 Visual Retrieval results

To further demonstrate the robustness and effectiveness of DSLNet, we acquire the eight nearest retrieval results of query images for analysis.

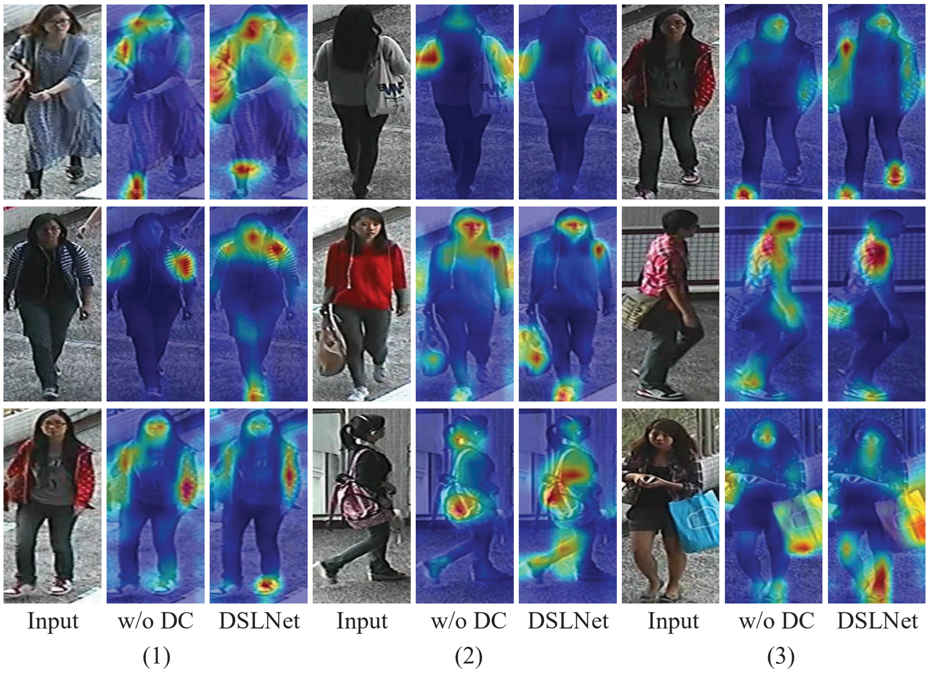

Fig. 3: Visualizations of the activation map. From left to right, each of the three images constitutes a set of comparison experiments, the right column represents the DSLNet activation map. The middle column represents the activation map with all dense connections removed from the DSLNet. We can see that with the addition of dense connections, DSLNet highlights more salient features.

We select the retrieval results of query samples under blur, occlusion, and illumination change. One can see that DSLNet can still get correct retrieval results in an unfavorable environment. The above experiments further prove the robustness and effectiveness of DSLNet, which benefits from the dense-scale feature representation and discriminative feature selection ability.

5 Conclusions

In this paper, we proposed DSLNet, an efficient multi-scale representation network which can extract dense-scale features while selecting discriminative multi-scale information. In the future, we will do further research to investigate the potential of DSLNet in other visual recognition tasks.

Acknowledgments

The paper is funded by the Key Scientific and Technical Innovation Engineer Project of Shandong (No. 2019JZZY011101).

References

1. Qian, X., Fu, Y., Jiang, Y.G., Xiang, T., Xue, X.: Multi-scale deep learning architectures for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 5399–5408
2. Chang, X., Hospedales, T.M., Xiang, T.: Multi-level factorisation net for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2109–2118
3. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3702–3712
4. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2017) 2590–2600
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7132–7141
6. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: 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) 480–496
7. Ustinova, E., Ganin, Y., Lempitsky, V.: Multi-region bilinear convolutional neural networks for person re-identification. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE (2017) 1–6
8. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: Glad: Global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. (2017) 420–428
9. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1077–1085
10. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 384–393
11. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 8514–8522
12. Dai, Z., Chen, M., Gu, X., Zhu, S., Tan, P.: Batch dropblock network for person re-identification and beyond. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3691–3701
13. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: 2018 ACM Multimedia Conference on Multimedia Conference, ACM (2018) 274–282
14. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: European conference on computer vision, Springer (2016) 791–808
15. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research **11** (2010)
16. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 403–412

17. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4004–4012
18. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
19. Xiao, Q., Luo, H., Zhang, C.: Margin sample mining loss: A deep learning based method for person re-identification. arXiv preprint arXiv:1710.00478 (2017)
20. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 21–29
21. Sudhakaran, S., Escalera, S., Lanz, O.: Lsta: Long short-term attention for egocentric action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9954–9963
22. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2285–2294
23. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2119–2128
24. Kim, W., Goyal, B., Chawla, K., Lee, J., Kwon, K.: Attention-based ensemble for deep metric learning. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 736–751
25. Lan, X., Wang, H., Gong, S., Zhu, X.: Deep reinforcement learning attention selection for person re-identification. arXiv preprint arXiv:1707.02785 (2017)
26. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European conference on computer vision, Springer (2016) 241–257
27. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 212–220
28. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 2285–2294
29. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1318–1327
30. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. arXiv preprint arXiv:1906.08332 (2019)
31. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 4510–4520
32. Wang, Y., Chen, Z., Wu, F., Wang, G.: Person re-identification with cascaded pairwise convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1470–1478
33. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: The IEEE International Conference on Computer Vision (ICCV). (2019)
34. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9317–9326

35. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 2138–2147
36. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8933–8940
37. Zhu, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Sun, X., Zheng, W.: Aware loss with angular regularization for person re-identification. In: AAAI. (2020) 13114–13121
38. Quan, R., Dong, X., Wu, Y., Zhu, L., Yang, Y.: Auto-reid: Searching for a part-aware convnet for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3750–3759
39. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1389–1398
40. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)