IEICE TRANS. INF. & SYST., VOL.E102–D, NO.11 NOVEMBER 2019

PAPER

Improving Slice-Based Model for Person Re-ID with Multi-Level Representation and Triplet-Center Loss*

Yusheng ZHANG[†a)], ***Nonmember***, **Zhiheng ZHOU**[†b)], ***Member***, **Bo LI**[†c)], **Yu HUANG**[†], **Junchu HUANG**[†], ***and*** **Zengqun CHEN**[†], ***Nonmembers***

SUMMARY Person Re-Identification has received extensive study in the past few years and achieves impressive progress. Recent outstanding methods extract discriminative features by slicing feature maps of deep neural network into several stripes. Still there have improvement on feature fusion and metric learning strategy which can help promote slice-based methods. In this paper, we propose a novel framework that is end-to-end trainable, called Multi-level Slice-based Network (MSN), to capture features both in different levels and body parts. Our model consists of a dual-branch network architecture, one branch for global feature extraction and the other branch for local ones. Both branches process multi-level features using pyramid feature alike module. By concatenating the global and local features, distinctive features are exploited and properly compared. Also, our proposed method creatively introduces a triplet-center loss to elaborate combined loss function, which helps train the joint-learning network. By demonstrating the comprehensive experiments on the mainstream evaluation datasets including Market-1501, DukeMTMC, CUHK03, it indicates that our proposed model robustly achieves excellent performance and outperforms many of existing approaches. For example, on DukeMTMC dataset in single-query mode, we obtain a great result of Rank-1/mAP = 85.9%(+1.0%)/74.2%(+4.7%).

key words: *person re-identification, multi-level, body parts, triplet-center loss, combined loss*

1. Introduction

Person Re-Identification(Re-ID) aims at retrieving identical people across surveillance videos captured from different cameras under challenging situations, such as posture, occlusion, illumination, background clutter, detection failure, etc. With the prosperity of deep convolution network, mainstream Re-ID methods are designed to describe each pedestrian with a concise but expressive vector and then match them in a task-specific metric space like Euclidean metric space, where the feature vectors of an identical person are expected to have smaller distances than that of different people.

Different from traditional classification problems, there is no overlap between the pedestrian identities in the training datasets and that in testing datasets. Still, in training phase, most models will use classification loss to strengthen extracted features' distinctiveness. Therefore in testing phase, the vectors we get through the network are considered to be expressive enough to separate one person from another.

Existing holistic feature extraction methods merely focus on salient global regions, which is not robust enough for pedestrian retrieval, therefore researchers turn to develop slice-based methods and these methods can be divided into four types, according to variant part locating ways: (i)methods with strong structural information, such as prior empirical knowledge about human bodies or poses; (ii)methods with region proposal; (iii)methods with feature enhanced by attention mechanism on salient partitions; (iv)methods with partitions divided into given stripes in features maps. However, there exists some limitation. First, pose or view of point variation can affect the reliability of prior part locating methods. Second, such pre-located methods only focus on specific parts with fixed semantic messages, but may miss some apparent discriminative information. Last but not least, most of them are non-end-to-end models, which need additional external datasets to train, increasing the inconvenience and difficulty of feature learning. Zheng et al.[1] uses Spatial Transform Network(STN) to align input images, and Li et al.[2] divides body parts roughly into head-shoulder, upper-body and lower-body using STN as well. [3] trains region proposal model on external pose dataset and in [4], authors develop attention mechanism. Still there is something to do like incorporating holistic features with local ones. [5] proposes a spatial-channel parallelism network for both holistic and partial person Re-ID to solve occlusion problems.

Also, most methods extract features from the last layer of deep nerual network, ignoring information of former layers. Unlike object detection or classification tasks, Re-ID needs compact and exclusive representation, while features from the last layer may lack supplementary information. Hence, fusing features learned at different layers can help obtain more robust and exclusive representation. Originally, [6] introduces pyramidal feature hierarchy for object detection, predicting target location on feature maps with different resolutions. And Zhang et al.[7] proposes a method utilizing mid-level attributes, using additional attribute labels.

Although in training phase we regard Re-ID as a classification task, each identity has only a few samples of differ-

Manuscript received March 7, 2019.
Manuscript revised July 10, 2019.
Manuscript publicized August 19, 2019.

[†]The authors are with the School of Electronic and Information Engineering, South China University of Technology, No. 381, Wushan Road, Guangzhou, China.

*The work is supported by National Key R&D Program of China (2018YFC0309400), National Natural Science Foundation of China (61871188), Guangzhou city science and technology research projects (201902020008).

a) E-mail: 201720111186@mail.scut.edu.cn
b) E-mail: zhouzh@scut.edu.cn
c) E-mail: leebo@scut.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2019EDP7067

ent views. With merely classification loss, it is hard to learn generalized representation. Consequently metric learning is necessary. Metric learning aims at learning about the similarity between dual images, that is, specifically in Re-ID, similarity of images belong to an identical person should be larger than that of different people.

Motivated by the improvement of multi-level feature pyramid and metric learning on slice-based model, in this paper, we propose a strategy combining both global and local information, creatively absorbing the merit of utilizing multi-level features and adding an objective function: triplet-center loss. With our proposed MSN model, for instance, performance on DukeMTMC increase to 85.9% rank-1 accuracy(+1.0%) and 74.2% mAP(+4.7%), surpassing many of mainstream methods.

The remainder of paper is organized as follows. Section 2 offers related work of person re-identification methods such as multi-level representations, slice-based models and metric learning. We describe our model in Sect. 3. Experimental results are compared and discussed in Sect. 4, followed by implementation details in Sect. 5. In the end, we make a conclusion on this paper in Sect. 6.

2. Related Work

This method focuses on learning deep partitioned and multi-level features for improving the Re-ID accuracy as well as designing appropriate objective function, which has tight relationship with part-slicing methods, multi-level representations and metric learning. Thus we briefly review some related aspects with Re-ID.

2.1 Slice-Based Models

Recently some slice-based methods have pushed the Re-ID performance to a new level. [8] horizontally slices the feature maps and merges these local features with bidirectional LSTM, combining with global features. While [9] also uniformly slices maps horizontally, concatenating local features as final representation, applying Refined Part Pooling(RPP) module to optimize part features' mapping validation, it is proved to have outstanding performance with high-level identification rates and mean average precision.

Still global information and auxiliary messages hidden in slightly shallow layers can be utilized to promote slice-based models.

2.2 Multi-Level Representation

Both utilizing feature maps from various layers, Zhang et al.[7] proposes obtaining auxiliary "mid-level" attributes from various blocks of ResNet-50 while Res50M [10] proposes "multi-level" representations extracted from different layers in the last block of ResNet-50. The latter performs better for containing adequate high dimensional semantic information than the former.

Although Res50M is designed for holistic Re-ID, "multi-level" representations should also be applied to slice-based model. Further more, with metric learning strategy, performance of Res50M is better than the author claims (from Rank-1/mAP=80.43%/63.88% to 82.0%/67.5% in DukeMTMC-reID dataset).

2.3 Metric Learning

Common loss function used in training Siamese architectures include Contrastive loss, Triplet loss [11], Triplet loss with batch hard mining (TriHard loss) [12], Quadruplet loss [13] and MSML [14]. These losses compel the network to contract intra-class distance of positive pair (the Anchor sample and the Positive sample) and increase inter-class distance of negative pair (the Anchor sample and the Negative sample). Other than these functions mentioned, the Center Loss [15] simultaneously learns a center for deep features of each identity and penalizes the distances between features and their corresponding class centers, which focuses on reducing intra-class variations but omits inter-class variations, leading to probable overlap between diverse classes. Recently, [16] proposes a variant of center loss called Triplet-center Loss (TCL), which has significant improvement in multi-view 3D object retrieval, contributing to both intra- and inter-class variations, eliminating the defect of center loss.

2.4 Data Augment

Data augmentation is introduced to make network more robust. Jon et al.[17] adopts an image "cut-out" strategy consisting of adding random noise to image regions in random size to augment the data. With the help of GAN, [18] expands the original training set without collecting extra data while Zhong et al.[19] generates camera style adaptation samples for Re-ID.

3. Multi-Level Slice-Based Network

In this section, we begin describing the dual-branch network architecture for multi-level feature learning. Then we describe the elaborate objective function that provides supervision to enforce the features to be correctly clustered.

3.1 Architecture of MSN

The dual-branch structure of Multi-level Slice-based Network (MSN) is shown in Fig. 1. With many alternative networks designed for image classification tasks as options, e.g., VGG Network, Google Inception and ResNet, we employ ResNet-50 as the backbone of our network not only for its competitive performance in some Re-ID methods [20], [9], but also considering its special architecture suitable for our framework: the feature maps of the last three layers have the similar structure, which is convenient to partition stripes and achieve feature fusion.

Table 1 lists the configuration of the two branches. In

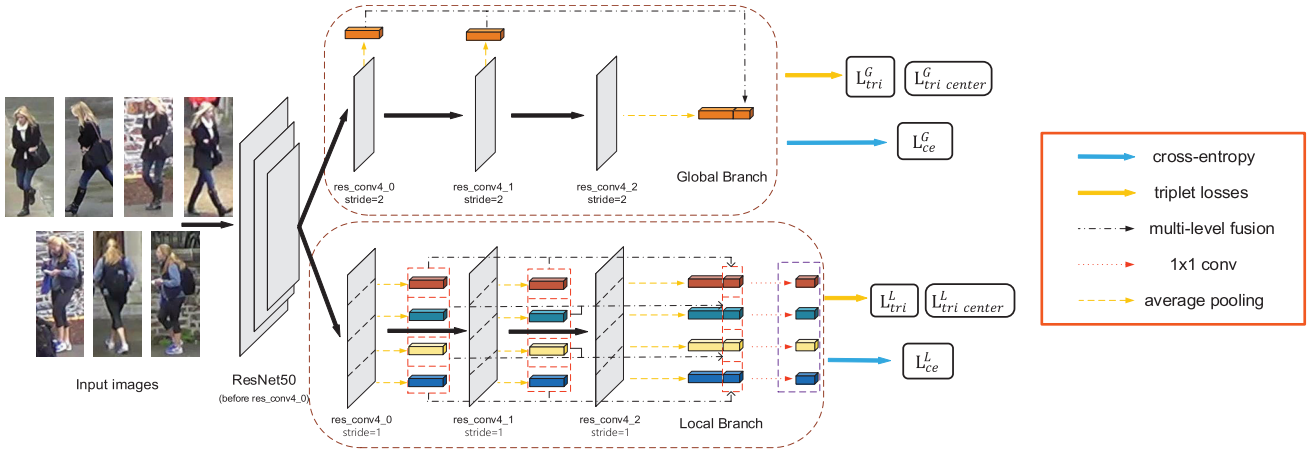


Fig. 1 MSN architecture. The ResNet-50 backbone is split into two branches after res_conv4_0 residual block: Global Branch, Local Branch. When evaluating, the dimension-reduced local feature vectors are concatenated together with global vector as the final representation of a pedestrian image. The 1×1 convolutions for dimension reduction and fully connected layers for identity prediction in each branch are independent, which have exclusive weights. Each path from the feature to the specific loss function represents an independent supervised back propagation. Best viewed in color.

Table 1 Comparison of the settings for two branches in MSN. The size of input images is set to 256×128. "Branch" refers to the name of branches. "Slicing Number" refers to the number of partitions on last three feature maps on each branch. "Mapping Size" refers to the size of last three output feature maps from each branch. "Mid Dim", "Final Dim" and "Reduction Dim" refers to the dimensionality and number of features for the different stages. "Features" means the symbols for the feature representations.

Branch	Slicing Number	Mapping Size	Mid Dim	Final Dim	Reduction Dim	Features
Global	1	8×4	2048×2	2048×1	-	$\mathbf{g}^G_{m_i}\|^1_{i=0}$, $\mathbf{g}^G_f$
Local	4	16×8	2048×8	2048×4	384×4	$\mathbf{g}^{L_j}_{m_i}\|^1_{i=0}\|^3_{j=0}$, $\mathbf{g}^{L_j}_f\|^3_{j=0}$

the upper branch, which is named **Global Branch** and employs original ResNet-50's parameters, we connect a global average pooling(GAP) layer in parallel after each convolution layer of the last three. We define features $\mathbf{g}^G_{m_i}|^1_{i=0}$ mid-level features and vector $\mathbf{g}^G_f$ final-layer feature then fuse mid-level features through a fusion module composed of a fully connected layer(FC) with batch normalization(BN) and ReLU to reduce 4096-dim vector to 1024-dim.This branch captures global feature representation by concatenating dimensionality-reduced vector $\mathbf{f}^G_m$ and final-layer feature $\mathbf{g}^G_f$ without partition, and so it is called the Global Branch.

The other branch is called **Local Branch**. One of notable variances between the two branches is that in the Local Branch we modify the default stride of *res50_conv4_0* layer from 2 to 1, by which we are capable to receive more adequate response feature maps containing more info. The mapping size doubles from 8×4 to 16×8. The other difference is that we adopt the horizontal slicing operation on response local maps from the last three convolution layers. Feature maps are equally partitioned into 4 independent parts (from 16×8 to 4×8×4). Each stripe employs similar but independent feature fusion operation as that in Global Branch. Also we apply a 1×1 convolution layer for reducing dimension after each 3072-dim vector (to 384 dimensions), supervising the four local representation processes with combined objective function.

During test phase, global feature vector is concatenated with all local feature vectors, adequately utilizing the complement information of different scales.

3.2 Combined Loss Function

To learn discriminative features of pedestrians, specific objective functions are adopted to supervise different parts of network in training procedure.

On one hand, in training phase, we regard the Re-ID representation task approximately as a multi-identity classification problem with few samples. Nevertheless, adopting initial form of cross entropy loss may raise over-fitting and reduction of adapting ability [21], for model becoming too confident about its predictions. Thus we adopt classification loss function with label smoothing encouraging the model to be less confident:

$$
\begin{aligned}
p(i) &= \frac{e^{\mathbf{W}_{yi}^T \mathbf{f}}}{\sum_{k=1}^{K} e^{\mathbf{W}_{yk}^T \mathbf{f}}} \\
\hat{q}(k) &= (1-\epsilon)\delta_{k,y} + \epsilon/K \\
L_{cels} &= \sum_{k=1}^{K} log(p(k))\hat{q}(k)
\end{aligned}
\tag{1}
$$

where **W** corresponds to a weight matrix, with size of vector's dimension N and number of training identities K. Among them $p(k)$ means the predicted probability of identity k, while $\hat{q}(k)$ refers to the modified ground-truth distribution of the identity, which is a mixture of ground-truth distribution $\delta_{k,y}$(equals 1 when $y=k$ and 0 otherwise) and a fixed distribution $\mu(k)=1/K$, with weights $1-\epsilon$ and ϵ.

On the other hand, we treat Re-ID tasks as metric learning problems. Thus the embedding are trained with metric losses to enhance clustering performance, in which batch-hard triplet loss and triplet-center loss are used as follows:

$$L_{triplet}=\sum_{i=1}^{B}\sum_{a=1}^{T} [\max_{p=1...T}\|\mathbf{f}_a^{(i)}-\mathbf{f}_p^{(i)}\|_2-\min_{\substack{n=1...T\\ j=1...B\\ j\neq i}}\|\mathbf{f}_a^{(i)}-\mathbf{f}_n^{(j)}\|_2+\alpha]_+ \tag{2}$$

where $\mathbf{f}_a^{(i)},\mathbf{f}_p^{(i)},\mathbf{f}_n^{(i)}$ are dimension-reduced features in two branches captured from anchor, positive and negative samples respectively, and α is a margin hyper-parameter to control the gap of intra and inter distances.$[\bullet]_+$ means maximizing operation compared to value 0. In per training batch, there are B selected identities and T images with each identity. With such definition, candidate triplets consist of hardest positive and negative sample pairs, improving the robustness of metric learning.

Further more, the triplet-center loss is defined as:

$$L_{tc}=\sum_{i=1}^{B}\sum_{a=1}^{T}[\|\mathbf{f}_a^{(i)}-\mathbf{c}^{(i)}\|_2^2-\min_{j\neq i}\|\mathbf{f}_a^{(i)}-\mathbf{c}^{(j)}\|_2^2+\beta]_+ \tag{3}$$

where $\mathbf{c^{(i)}}$ means the learnt center feature of identity i, and β is another margin hyper-parameter to control the gap of corresponding-center and irrelevant-center distances. Batch-hard triplet loss truly pushes the two hardest feature samples away but triplet-center loss compel the features of identical person to cluster around corresponding class centers and get far away from the most confusing class centers.

Ultimately, the combined loss is defined as:

$$L=L_{cels}+\lambda_1 L_{triplet}+\lambda_2 L_{tc} \tag{4}$$

in which λ_1 and λ_2 are scale factors.

4. Experiment

To evaluate our proposal, we conduct extensive experiments on three large-scaled datasets accepted by mainstream methods.

4.1 Datasets

The three datasets are: Market 1501, DukeMTMC-reID and CUHK03.

Market 1501 This dataset consists of images of 1,501 pedestrians captured from 6 different cameras, which are cropped with bounding-boxes predicted by DPM-detector. The whole dataset is divided into training set with 12,936 images of 751 persons and testing set with 3,368 query images and 19,732 gallery images of the other 750 persons.

DukeMTMC-reID It is a subset of the DukeMTMC which is designed for person re-identification. It contains 36,411 images of 1,812 persons from 8 different high-resolution cameras. 16,522 images of 702 persons are randomly selected from the dataset as the training set, and the remaining 702 persons are divided into the testing set that contains 2,228 query images and 17,661 gallery images.

CUHK03 This dataset consists of 14,097 images of 1,467 persons from 5 different pairs of cameras. Two types of subsets are provided in this dataset: dataset 'labelled'(pedestrian bounding boxes manually labelled) and dataset 'detected'(DPM-detected bounding boxes). With uniform test protocols as the other two datasets, the CUHK03 might be the most challenging dataset at present for person retrieving, because of its obviously less training images number(7,365) compared to DukeMTMC-reID(16,522) and Market-1501(12,936).

4.2 Evaluate Protocols

In our experiments, to evaluate the performances of Re-ID methods in a mainstream standard, we adopt mean average precision(mAP), and the cumulative matching characteristics(CMC) at rank-1, rank-5 and rank-10 on all the candidate datasets above, which are complementary to reflect the retrieving performance.

In this paper we deploy the single query evaluate protocol. During evaluation, we finally extract the concatenated features of query and gallery images to compute Euclidean distance and select the nearest 50 gallery samples for each query sample to compute mAP and CMC. For each query image, instances of identical person from the same camera in gallery images will be discarded during computation. Following experiments all deploy the same mode and testing samples of various datasets are fixed.

4.3 Comparison with State-of-the-Art Methods

We compare the performance of our proposed method with current state-of-the-art methods on all the candidate datasets to show advantage over all the existing competitors. Results

Table 2 Comparison of results on Market-1501, 'RPP' refers to implementing refined part pooling.

Methods	Rank-1	mAP
Mid-Level [10]	89.9%	75.6%
HA-CNN [22]	91.2%	75.7%
DuATM [23]	91.4%	76.6%
PCB [9]	92.3%	77.4%
GSRW [24]	92.7%	82.5%
DNN_CRF [20]	93.5%	81.6%
PCB+RPP [9]	**93.8%**	81.6%
MSN(Ours)	93.4%	**82.8%**

Table 3 Comparison of results on DukeMTMC-reID, where 'RPP' refers to implementing refined part pooling.

Methods	Rank-1	mAP
HA-CNN	80.5%	63.8%
Deep-Person [8]	80.9%	64.8%
MLFN [25]	81.2%	62.8%
Mid-Level	81.5%	66.6%
DuATM	81.8%	64.6%
PCB	81.9%	65.3%
PCB+RPP	83.3%	69.2%
Part-aligned [26]	84.4%	69.3%
DNN_CRF	84.9%	69.5%
MSN(Ours)	**85.9%**	**74.2%**

Table 4 Comparison of results on CUHK03-detected, 'RE' refers to implementing random erasing, 'RPP' refers to implementing refined part pooling.

Methods	Rank-1	mAP
HA-CNN	41.7%	38.6%
MLFN	52.8%	47.8%
Mid-Level	54.1%	52.1%
DaRE [27]	55.1%	51.3%
TriNet+RE [28]	55.5%	50.7%
PCB	61.3%	54.2%
PCB+RPP	63.7%	57.5%
MSN(Ours)	**64.8%**	**62.7%**

in detail are given as follow:

Comparison on Market-1501 The results on Market-1501 dataset is shown in Table 2. In evaluating protocol we adopted, our proposed MSN achieves highest 82.8% in mAP and the third best 93.4% in Rank-1, while competitive method PCB+RPP achieved the best rank-1 result, and GSRW achieved the second highest mAP.

Comparison on DukeMTMC-reID As is shown in Table 3, the performance of MSN is excellent on this challenging dataset. MSN achieves result of Rank-1/mAP=85.9%/74.2%, outperforming all the given methods in both rank-1 and mAP.

Comparison on CUHK03 According to Table 4, our MSN achieves Rank-1/mAP= 64.8%/62.7% on this extremely hard CUHK03 detected subset, still the best in both mAP and Rank-1. The reason why there is an tremendous gap between results of CUHK03-detected and the other two datasets is that the imbalance of dataset size and the few instances of each identity limit our model to learn discriminative message.

4.4 Qualitative Analysis on Effective Components

Part slicing As Fig. 2 shows, part-sliced deep network can learn preliminary prominence on different parts, according to potential semantic information. Such as focused thermal regions in Fig. 2(a), they give attention against view variation. In Fig. 2(b), although the two instances are not part-aligned, the responses of Region 1 can both focus on their heads.

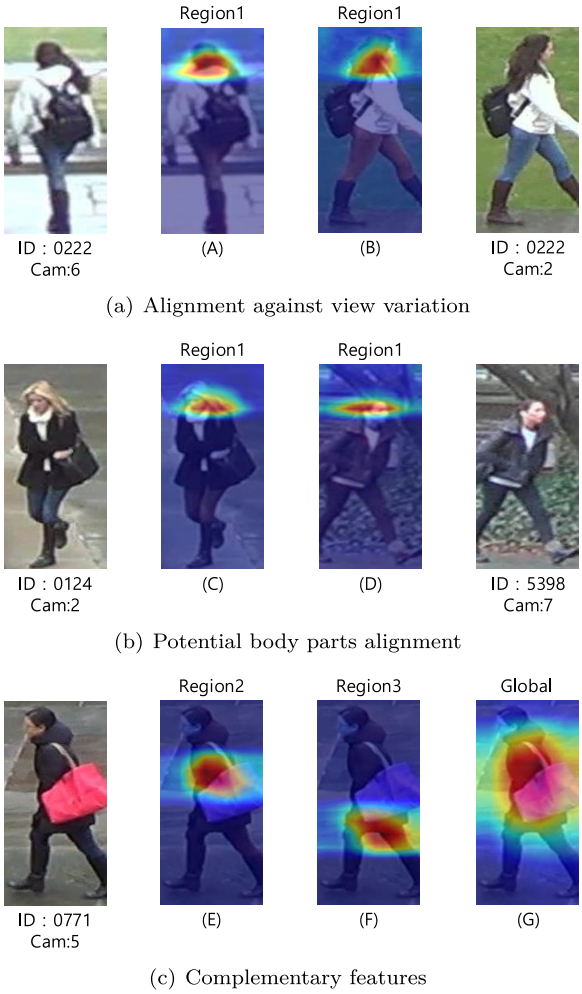


Fig. 2 Visualization of feature maps on final layer of proposed MSN's Local Branch in training phase using Grad-CAM [29]. Region 1, 2, 3 indicate the horizontally sliced feature stripes. Global indicates the feature map from Global Branch.

Dual Branch Dual-branch setting is very common for person Re-ID tasks. The Global Branch extracts discriminative representation focusing on the whole body while the Local Branch gives attention to semantic part patterns that the Global Branch may omits. Figure 2(c) reveals that the Global Branch cares about the region around the pink bag, when the Local Branch cares about different semantic parts. In 2nd to 4th rows of Table 5 we acknowledge the effectiveness of dual branch, with which the performance can reach great promotion especially in mAP.

Multi-level Feature Fig. 3 presents different weight-focusing regions of feature response maps from different layers in Local Branch of MSN, with a certain input image. The explanation why weight-focusing regions of final-layer are much bigger than that of fused layer0 is that our multi-level features have experienced dimensionality reduction operation, which is concatenated as auxiliary features. Observing the heatmaps, we can notice the multi-layer0 in deep block focus on slightly different parts compared to final-layer that may give attention to background, which of-

Fig. 3 Visualization of feature maps on two layers from last of proposed MSN's Local Branch in training phase. Region1 indicates the first horizontally sliced feature stripe. The 'final-layer' means the last layer and the 'multi-layer0' means the third layer from last.

Table 5 The influence of different branch setting. In Local Branch, $\mathbf{g}^L_{m_i}|^1_{i=0}$ indicate multi-level features, $\mathbf{f}^L_{m_i}|^1_{i=0}$ indicate middle-level features from shallow blocks and $\mathbf{g}^L_f$ refers to final-layer feature. All of them are evaluated on Market 1501 dataset.

Branch setting	Rank-1	mAP
Global only	90.6%	78.3%
Local only	92.7%	80.5%
Both	**93.4%**	**82.8%**
2	93.2%	82.7%
4	**93.4%**	**82.8%**
8	92.8%	82.0%
$\mathbf{g}^L_{m_i}|^1_{i=0}$	90.9%	78.2%
$\mathbf{g}^L_f$	90.6%	77.1%
$\mathbf{f}^L_{m_i}|^1_{i=0}$ and $\mathbf{g}^L_f$	91.3%	80.2%
$\mathbf{g}^L_{m_i}|^1_{i=0}$ and $\mathbf{g}^L_f$	**93.4%**	**82.8%**

fers complimentary information.

In some cases it is useful to incorporate multiple spatial granularities of information such as image-to-image synthesis problems (utilizing features from shallow blocks). However, in our case we aim to extract expressive representation for pedestrian matching, and features from shallow blocks lack semantic information. Incorporating such mid-level features damages semantic representation. According to 8th to 11th rows of Table 5, not only can we see that the joint usage of $\mathbf{g}^L_{m_i}|^1_{i=0}$ and $\mathbf{g}^L_f$ can get better performance than just using any of them, but also we can find that multi-level features from deep block perform exactly better than mid-level features from shallow blocks.

Combined Loss with Triplet-center Loss We treat Re-ID as a multi-classification mission in training phase with lots of known identities and few corresponding samples that vary in position and posture. However, there is no overlap between test set and training set. Thus besides correct classification, we need corresponding features to cluster around so that in test phase features belong to same unknown identities can have shorter distance. Triplet loss enhances clustering through sample distance which is useful but not enough while triplet-center loss makes further efforts on clustering by learning identity centers. The 2nd to 6th rows of Table 6 demonstrate the promotion of accuracy performance using combined loss and proves the significance of triplet-center loss.

Table 6 Different performance with various settings of combined loss on DukeMTMC-reID dataset. 'CELS' means cross entropy label smoothing loss function, 'CE' means standard cross entropy loss function, 'Htri' means batch-hard triplet loss function, 'Tri-center' means triplet center loss function. In the second part of table, numbers in '()' indicates (λ_1, λ_2, initial LR for class center) in Local Branch.

Objective Function	Rank-1	mAP
CELS only	82.5%	68.2%
CELS + Htri	84.5%	71.9%
CE + Htri	80.2%	70.1%
CELS + Tri-center	85.1%	71.5%
CELS + Htri + Tri-center	**85.9%**	**74.2%**
(1, 0.01, 0.1)	83.1%	70.8%
(5, 0.01, 0.1)	84.0%	71.2%
(5, 0.05, 0.1)	84.4%	71.0%
(5, 0.01, 0.01)	85.4%	73.6%
(5, 0.02, 0.01)	**85.9%**	**74.2%**

Also, values of scale factors and learning rate of class center can make a great difference, experiment results in 7th to 11th rows of Table 6 also illustrate this. The reason why scale factor λ_2 of triplet-center loss is much smaller than that of triplet loss is that in the beginning of training our class centers are randomly initialized so the center distance can be very large. If we use similar scale as λ_1, the training process is hard to converge. Additionally, we can obviously observe effects using triplet-center loss that the performance on mAP improves more than that on Rank-1 accuracy, which proves the positive effects of triplet-center loss on average meaning.

5. Implementation

To accelerate the training phase, we load the pretrained weights of ResNet-50 on ImageNet to initialize the backbone and two branches of MSN. We set B = 32 and T = 4 to train our proposed model and the mini-batch selection for triplet losses is stochastic. Each triplet consists of farthest positive pair and hardest negative pair which is stochastic along with stochastic batch sampling. About the hyperparameter *margin* α and β for triplet loss and triplet-center loss, we set to 2 and 3. In the *label-smoothing regularization*, we set the $\epsilon = 0.1$. Finally, the scale factors λ_2 of the combined loss is set to 0.02. We choose Adam as the optimizer with weight decay 5e-4. As for the learning rate strategy, we set the initial learning rate to 2e-4, and decay the learning rate to 2e-5 and 2e-6 after training for 30 and 60 epochs. In addition, the learning rate of class centers is set to 0.01. The total training process lasts for 90 epochs. And during per epoch, we train two branches independently while freezing the other branch for these two branches have different partitions, thus the objective function should have different forms. We adopt $\lambda_1 = 1$ in the Global Branch and $\lambda_1 = 5$ in the Local Branch.

During evaluation, we ultimately extract the global feature concatenated with local dimensionality-reduced features to compute mAP and CMC in Euclidean metric space.

Our work is implemented on PyTorch 0.4.0 framework. All our experiments on different datasets follow the settings above.

6. Conclusion

This paper proposes a Multi-level Slice-based Network (MSN), a novel deep network for learning discriminative representations in person re-identification tasks using effective objective function. Both branches utilize the multi-level messages in slightly shallow layers with different partitions. Adequate experiments indicate that proposed MSN achieves excellent performance on several mainstream person Re-ID datasets and confirmed components' effectiveness.

References

[1] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," IEEE Trans. Circuits Syst. Video Technol., pp.1–1, 2018.

[2] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," 2017 IEEE Conference on Comput. Vis. Pattern Recognit., CVPR 2017, pp.7398–7407, Honolulu, HI, USA, July 2017.

[3] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," 2017 IEEE Conference on Comput. Vis. Pattern Recognit., CVPR 2017, pp.907–915, Honolulu, HI, USA, July 2017.

[4] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," IEEE Trans. Image Process., vol.26, no.7, pp.3492–3506, July 2017.

[5] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, and W. Jiang, "Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification," CoRR, vol.abs/1810.06996, pp.19–34, 2018.

[6] T. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, and S.J. Belongie, "Feature pyramid networks for object detection," 2017 IEEE Conference on Comput. Vis. Pattern Recognit., CVPR 2017, pp.936–944, Honolulu, HI, USA, July 2017.

[7] G. Zhang and J. Xu, "Person re-identification by mid-level attribute and part-based identity learning," Asian Conference on Machine Learning (ACML), Nov. 2018.

[8] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," CoRR, vol.abs/1711.10658, 2017.

[9] 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)," Computer Vision – ECCV 2018, ed. V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Cham, pp.501–518, Springer International Publishing, 2018.

[10] Q. Yu, X. Chang, Y. Song, T. Xiang, and T.M. Hospedales, "The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching," CoRR, vol.abs/1711.08106, 2017.

[11] V.K.B. G, G. Carneiro, and I.D. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions," 2016 IEEE Conference on Comput. Vis. Pattern Recognit., CVPR 2016, pp.5385–5394, Las Vegas, NV, USA, June 2016.

[12] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," CoRR, vol.abs/1703.07737, 2017.

[13] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," 2017 IEEE Conference on Comput. Vis. Pattern Recognit., CVPR 2017, pp.1320–1329, Honolulu, HI, USA, July 2017.

[14] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," CoRR, vol.abs/1710.00478, 2017.

[15] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part VII, pp.499–515, Amsterdam, The Netherlands, Oct. 2016.

[16] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3d object retrieval," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

[17] J. Almazán, B. Gajic, N. Murray, and D. Larlus, "Re-id done right: Towards good practices for person re-identification," CoRR, vol.abs/1801.05339, 2018.

[18] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," IEEE Int. Conf. Computer Vision, ICCV 2017, pp.3774–3782, Venice, Italy, Oct. 2017.

[19] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

[20] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2016 IEEE Conference on Comput. Vis. Pattern Recognit., CVPR 2016, pp.2818–2826, Las Vegas, NV, USA, June 2016.

[22] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

[23] 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," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

[24] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

[25] X. Chang, T.M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

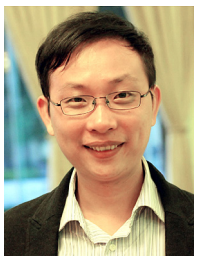
[26] Y. Suh, J. Wang, S. Tang, T. Mei, and K.M. Lee, "Part-aligned bilinear representations for person re-identification," Computer Vision - ECCV 2018 - 15th European Conference, Proceedings, Part XIV, pp.418–437, Munich, Germany, Sept. 2018.

[27] 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," The IEEE Conference on Comput. Vis. Pattern Recognit. (CVPR), June 2018.

[28] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," CoRR, vol.abs/1708.04896, 2017.

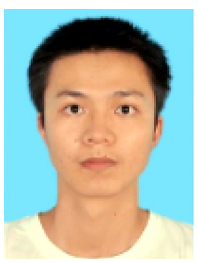
[29] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," Oct. 2017.

Yusheng Zhang recieved B.S. degree in Electronic information engineering from Hunan University. He is currently working toward the M.S. degree in Signal and information processing in South China University of Technology. His research interests are computer vision, image processing and machine learning.



Zhiheng Zhou received his M.S. degree in mathematical statistics and Ph.D. in Communication and information system from South China University of Technology. He is a professor with South China University of Technology. His research interests include pattern recognition and image engineering.

Bo Li received the B.E. degree in information management from Xidian University(XDU), and the M.E. degree in computer science and engineering and the Ph.D. degree in signal processing from the South China University of Technology. He is an Associate Professor with the School of Electronic and Information Engineering, South China University of Technology. His research interests include image/video processing, saliency model, NMF and image registration.



Yu Huang recieved B.S. degree in information engineering from South China University of Technology. He is currently working toward the M.S. degree in Signal and information processing in South China University of Technology. His research interests are computer vision, image processing and machine learning.

Junchu Huang recieved B.S. degree in information engineering from South China University of Technology. He is currently working toward the Ph.D. degree in School of Electronic and Information Engineering in South China University of Technology. His research interests are transfer learning and image processing.

Zengqun Chen recieved B.S. degree in information engineering from South China University of Technology. He is currently working toward the M.S. degree in Signal and information processing in South China University of Technology. His research interests are computer vision, image processing and machine learning.