Mask Separated Feature Learning for Person Re-Identification

Jinbeom Kim, GyeongHwan Kim
Sogang University
Seoul, Korea
mmi.sogang.ac.kr

Abstract

The local feature is widely used to overcome the lack of the global feature representation. But most works in person re-identification task train two features with same spatial information. In this paper, we propose the mask separated feature learning for person re-identification called MSReID. Our method can be implemented after any CNN backbone network such as ResNet50. Instead of generating mask with semantic annotations, MSReID is trained with only ID class labels. Our experiments have proven that MSReID outperforms the previous state of the arts on Market1501, DukeMTMC-reID, and CUHK03-NP datasets. For example, on CUHK03-L dataset, we obtain the best performance of Rank-1/mAP=$84.6\%/81.9\%$ *with a large margin (over than* 3.5%*). In addition, the ablation studies have shown that the mask separated feature learning scheme is helpful than the results when the local feature is used only without the generated mask.*

1. Introduction

The purpose of person re-identification (ReID) is to determine that the person who is taken from multiple non-overlapping cameras is the same person. Given a query image, it compares all the images in the gallery to find the closest image. In general, ReID focuses on the cropped person images which are obtained by pedestrian detector, not the entire frame of videos.

ReID has been improved significantly with advances of deep convolutional neural network (CNN). Despite the previous works, it remains a challenging problem. First, ReID has no pre-defined class in practical cases. This means that ReID needs to be learned in a different way than general classification. Second, since cropped images are captured by multiple cameras, the illumination conditions are different. And some of human body parts can be visible on one camera, but not on others. Even whithin a same camera, body parts are often occluded by background clutter, other people, and wearing stuffs. Also cropped images are often misaligned. These circumstances remove the significant features of an image and makes ReID task more difficult. Third, a human has flexible joints and this makes the different actions. Hence, the posture of the person can be changed dramatically even in a short moment. It incudes a large intra-class variation.

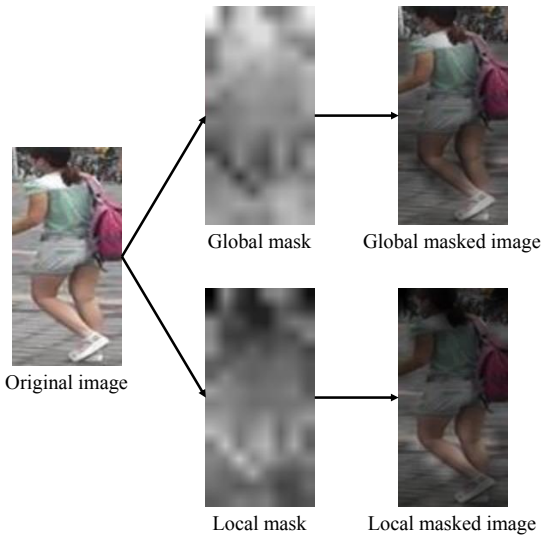

Figure 1: The visualization of the global mask and the local mask. The local mask is obtained by reversing the global mask. Each mask is resized to compare the spatial attention with the given image. For the original image, the global mask focus on the head parts especially, whereas the leg parts is concentrated on the local mask.

To overcome these problems, many efforts [29, 32, 31, 30, 10, 19, 16, 4, 20, 26] haven been focused on aligning human body parts. Among them, stripe based methods are widely used for separating each body parts because detected human bounding boxes often stand from camera. In [30, 4, 20], feature map extracted from an image are divided based on the row. The feature distance between the other feature maps are measured and obtained the relation

between the row feature maps. This makes it easy to remove the background on the incorrectly captured bounding box. Some works [29, 31, 10] have approached by comparing each body part rather than simply dividing the feature map in the row direction. Even [31] use 24 body parts images, including head, hands, upper body, lower body, *etc*.

In this paper, we propose the mask separated feature learning for person Re-identification (MSReID). As far as we know, there are no previous work which train the global feature and the local feature separately. To separate two features, we design our network to generate mask without any pixel level annotations and only class labels are fed during training. In other words, our network doesn't need any additional segmentation netowkr or annotations. The feature map is masked with generated probability mask in the global branch, whereas the mask is reversed and the local feature map is multiplied with this reversed mask. This learning scheme leads that the global feature and the local feature are trained with different spatial attention. As shown in Figure 1, when the input image is given, different regions are concentrated in the global mask and the local mask. In our experiments, we have shown that mask separated feature learning is helpful to discriminate semantic information than using only the local feature or the mask.

In following sections, we first review the related works including mask guided methods, local and global based approaches, and attention modules. And we present our proposed method and following experiments to evaluate MSReID on popular person ReID datasets. Finally, the ablation studies are conducted to compare the effect of each modules which are used in our network.

2. Related works

2.1. Mask guided approaches

In practical cases, detected bounding boxes are often misaligned and the background clutter is included in the bounding box. Only things we focus are foreground, especially human body. To avoid the effect of the background, recent works [16, 10, 19] use additional information such as human body segmentation. In [16], the mask is generated by segmentation network from an original image and the masked image is created to extract foreground. The masked image and original image is concatenated and these two images are fed into the ReID network to obtain the feature vector. [10] use two CNN networks. one of them is used to get the global feature map and the other is used for the probability map of human body parts including foreground, head, upper body, lower boddy, shoes. For each body parts, the global feature map is multiplied with probability map and five local features are obtained by global average pooling. [19] also use the binary mask for foreground. By using this mask, the foreground and the background image is created and the network is trained in such a way that the foreground and the original image pull each other and push with the background image.

2.2. Local and global based approaches

Combining the global and the local feature can achieve the better performance in recent works [29, 3, 31, 13, 16, 2, 36, 26]. [29] propose the network which consist of the localization, sampling, and feature extraction modules. Localization and sampling module classify the body parts and feature vectors of each body parts are obtained by feature extraction module. In [3], the batch dropblock method are proposed and it is applied only to the local feature map. Since the occlussion is often occurred in ReID, the dropblock is helpful to regularize the training results. [13] and [36] propose the network architecture which have the small number of the weight parameters. These architectures use the local features because the global feature alone cannot perform well.

2.3. Attention based approaches

The attention mechanism is widely used in vision tasks. In recent works, various attention methods are used in ReID. [32] use the affinity matrix in feature map to calculate the relation between the features and perform spatial and channel attention. In affinity matrix, a background with relatively low relation can be removed. [11] and [1] adopt reinforcement learning method for different purposes. [11] adjust misaligned bounding box so that feature vector can be trained in a better form. In contrast, [1] adopt reinforcement learning methods to make the attention module achieve the better performance. In [13], the new attention module are proposed and it is called harmonic attention (HA). HA consist of two parts, including soft attention and hard attention. Soft attention use channel wise pooling and squeeze and excitation methods to obtain spatial and channel attention. Otherwise, And hard attention adopt the spatial transformer network (STN) to localizing the foreground image. These methods is helpful to find the more important regions and achieves the better performance.

3. Proposed Method

In this section, we present the model with two feature branch which are separated by self-trained mask. Without any additional annotations to the mask, the network generates the mask that allows two features to be trained with different spatial information. Overall architecture of MSReID is illustrated in Figure 2.

3.1. Backbone network

We adopt ResNet50 [5] for the backbone network which are commonly used in recent works [14, 3, 1, 2]. When images are fed into ResNet50, the feature map of the images

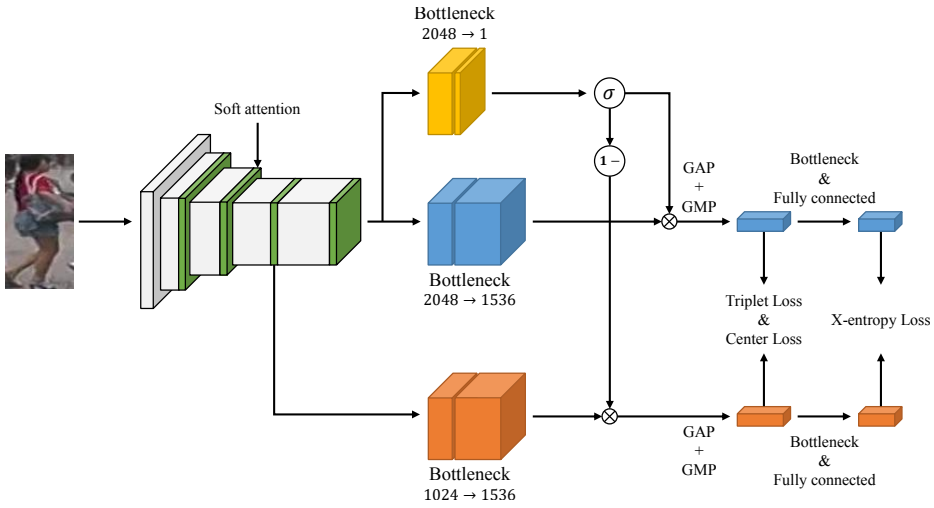

Figure 2: Overall architecture of MSReID. The global and the local feature are trained separately by generated mask. Global average pooling (GAP) and global max pooling (GMP) is applied each feature map to make feature vector. We use triplet loss and center loss for ReID, and cross entropy loss for ID classification.

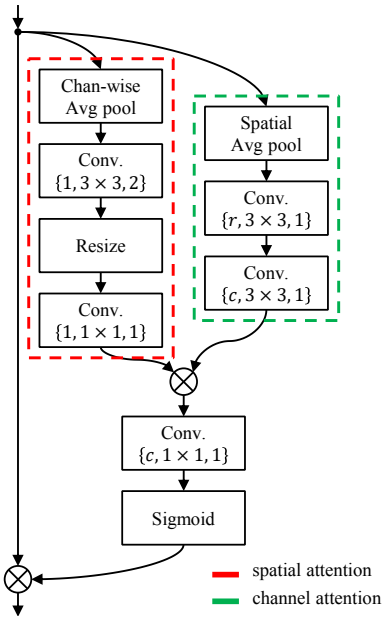

Figure 3: The structure of soft attention module proposed in [13]. It consist of spatial attention and channel attention. Red and green dashed box indicates spatial and channel attention, respectively. The ReLU and Batch Normalization (BN) [8] is applied after each convolution layer. We change the channel-wise average pooling into the weighted average pooling which is simply implemented with a non-biased fully connected layer.

are 32× downsampled. Since ReID often takes low resolution images (*e.g.* 256 × 128), it is hard to extract semantic information with usual ResNet. [21] first removed downsampling from the last layer of ResNet to solve this problem. We also change last stride from 2 to 1. This induces additional computation cost, but better performance.

3.2. Soft attention module

Since the attention module was first used in machine translation [22], attention has also been used in computer vision applications [25, 15, 28]. In person ReID, [13] proposed Harmonious Attention (HA). It is consist of two parts, hard attention and soft attention. Hard regional attention acts in a similar way to Spatial Transformer Network (STN) [9] because it learns a transformation matrix with scale and translation factors. It removes the background in the given cropped image so that network focuses on the human body parts. Soft attention in HA is divided into two parts, spatial and channel attention. Figure 3 shows the structure of soft attention module. In spatial attention, feature map is first reduced into 1 channel. Unlike in [13], we modify the average pooling into weighted average pooling. After reducing the feature map to 1 channel, a convolution layer is applied. The structure of the channel attention module is similar to squeeze-and-excitation block in [7]. To combine the spatial attention and channel attention, 1×1 convolution layer and sigmoid function is used to adjust the soft attention.

The soft attention module is placed after each layer of ResNet except the input 7×7 convolution layer. Empirically, the output result of which the model has the attention after the input layer is worse than of which it doesn't.

3.3. Mask separated features

In this work, we use multiple descriptors which is used in recent works [13, 3, 2, 19]. We use one global feature and one local feature. Local feature is came from third layer on ResNet50 and it has less refined information than the global feature. But local feature is trained without any spatial indicator such as mask. Hence, it makes the global and the local feature have the similar information on same spatial regions. To solve this problem, we use mask to train the global and the local feature with different spatial attention.

Unlike [10] and [16], MSReID does not use additional segmentation networks and segmentation annotation is required. We use the bottleneck block in [5] to generate mask without annotations and replace the activation function to sigmoid function. The range of the mask after sigmoid function is 0 to 1. The feature map after last layer of ResNet50 is multiplied by this mask and the global feature is obtained by taking global average pooling (GAP) and global max pooling (GMP) on the masked feature map. To leads the local feature is trained differently from the global feature, we reverse the mask and multiply it with the feature map of the third layer of ReNet50. Similar to global feature, GAP and GMP is applied to get the local feature.

3.4. Feature processing

We use the bottleneck block [5] to change the dimension of the global and the local feature to balance two triplet losses. In our network each dimension is set to be 1536. Each feature vector is trained separately and two vectors are concatenated in test stage. Hence, the total dimension of the feature is 3072. When measuring the similarity between two images, the form of feature map has disadvantage since Euclidean distance is sensitive to the spatial alignment of the feature maps. To remove them, global pooling is used in general. In our architecture, we use GAP and GMP. GAP can see the whole part of an image, but it also includes the background features. On contrary, GMP only focus on a specific part of the image. We use both GAP and GMP to complement each other.

Including the soft attention and the mask separated feature, we use a bottleneck layer with no bias before the classification layer which is used in [14]. Since classification and ReID are similar but different task, the bottleneck layer is used to make each task trained with different scale.

3.5. Loss function

To train our network, we use the hard batch triplet loss [6], center loss [27], and softmax cross-entropy loss with label smoothing [23], which are widely used in image based retrieval tasks.

Triplet loss use three input images, including query, positive, and negative. The purpose of triplet loss is to close the distance between query and positive, vice versa between query and negative. Batch hard triplet loss is formulated as

$$L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{K}\Big[m + \max_{p=i,\cdots,K} D(\boldsymbol{f}_a^i, \boldsymbol{f}_p^i) - \min_{\substack{j=1,\cdots,P \\ n=1,\cdots,K \\ i\neq j}} D(\boldsymbol{f}_a^i, \boldsymbol{f}_n^j)\Big]_+ \tag{1}$$

where $D(\boldsymbol{x}, \boldsymbol{y})$ denotes the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. And $\boldsymbol{f}_i^j$ corresponds to the i-th image of the j-th person in the mini batch. m is the margin of the triplet loss, and $[x]_+$ is a hinge loss which equals to $\max(0, x)$. If the negative distance is greater than m compared to the positive distance, the triplet loss becomes zero.

Since triplet loss is sensitive on miss-labeled data, it is hard to train a network with triplet loss only. To overcome this problem, we also use center loss. Center loss learns a center feature in Euclidean space. The center loss can be written as

$$L_{center} = \sum_{i=1}^{B} \|\boldsymbol{f}_i - \boldsymbol{c}_{y_i}\|_2^2 \tag{2}$$

where y_i denotes the label of the i-th image. $\boldsymbol{f}_i$, $\boldsymbol{c}_{y_i}$, B denote the i-th feature vector, center feature vector of y_i-th class, and the number of batch size, respectively. Center loss causes the feature vectors of each class to cluster and reduces intra-class variances of feature vectors.

Recently, classification loss is used in ReID. Unlike center loss, classification loss increase inter-class variance. we also use label smoothing regularization which is widely used. The softmax cross-entropy loss with label smoothing is computed as

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{B}\sum_{c=1}^{C}\log\left(p_i^k\right)\left((1-\epsilon)y_i^k + \frac{\epsilon}{C}\right) \tag{3}$$

where B and C denote batch size and the number of classes. $\epsilon \in (0, 1)$ is a smoothing parameter, p_i^k denotes the predicted probability of i-th data on the k-th class, and y_i^k indicates the ground truth $\in [0, 1]$.

MSReID use three losses as follow:

$$L_{total} = L_{cls} + L_{triplet} + \lambda_c L_{center} \tag{4}$$

where λ_c is a weight of center loss.

4. Experiments

4.1. Datasets

For evaluations, we selected three commonly used in person ReID datasets, Market1501 [33], DukeMTMC-ReID [17], and CUHK03-NP [12, 34]. The datasets are splited into training, query, and gallery data. For testing,

Type	Method	Feature dim	Backbone	Market1501		DukeMTMC-ReID		CUHK03-L		CUHK03-D	
				Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Mask guided	SPReID [10]	12288	inception	92.5	81.3	84.4	71.0	-	-	-	-
	MaskReID [16]	256	inception	90.4	75.4	78.9	61.9	-	-	-	-
Attention based	ABD [2]	2048	resnet	95.6	88.3	**89.0**	78.6	-	-	-	-
	Mancs [24]	2048	resnet	93.1	82.3	84.9	71.8	69.0	63.9	65.5	60.5
	RGA-SC [32]	2048	resnet	**96.1**	88.4	-	-	**81.1**	**77.4**	**79.6**	**74.5**
	SCAL [1]	2048	resnet	**95.8**	**89.3**	**89.0**	**79.6**	74.8	72.3	71.1	68.6
	HACNN	1024	hacnn	91.2	75.7	80.5	63.8	44.4	41.0	41.7	38.6
Others	AlignedReID [30]	4096	resnet	91.8	79.3	-	-	-	-	-	-
	PCB [20]	12288	resnet	93.8	81.6	83.3	69.2	-	-	-	-
	HPM [4]	3840	resnet	94.2	82.7	86.6	74.3	-	-	-	-
	DSA [31]	-	resnet	95.7	87.6	86.2	74.3	78.9	75.2	78.2	73.1
	MGN [26]	768	resnet	95.7	86.9	88.7	78.4	68.0	67.4	66.8	66.0
	BDB [3]	1536	resnet	95.3	86.7	**89.0**	76.0	79.4	76.7	76.4	73.5
	BagOfTricks [14]	2048	resnet	94.5	85.9	86.4	76.4	-	-	-	-
	LocalCNN [29]	512	local cnn	91.5	77.7	82.2	66.0	58.7	53.8	56.8	51.6
	OSNet [36]	512	osnet	94.8	84.9	88.6	73.5	-	-	-	-
-	MSReID (ours)	3072	resnet	**95.8**	**88.8**	**90.2**	**80.6**	**84.6**	**81.9**	**82.7**	**79.5**

Table 1: Performance comparisions with state of the arts methods on Market1501, DukeMTMC-ReID, and CUHK03-NP datasets. Bold and red numbers denote the best performance and underlined numbers denote the second best. MSReID achieves the best performance on DukeMTMC-ReID and CUHK03-NP, while the second best on Market1501. Our results are measured from the average of 8 different training results.

Datasets	Market1501	DukeMTMC	CUHK03-NP	
			Labeled	Detected
Identities	1501	1404	1467	1467
Images	32271	36411	14096	14097
Camearas	6	8	2	2
Train IDs	751	702	767	767
Test IDs	750	702	700	700

Table 2: The detailed statistics of three datasets.

each image in the query measures distances with all data in the gallery. Unlike the original CUHK03 dataset [12], recently a new protocol of CUHK03 has been proposed and it become the most challenging dataset because it use only two cameras. More detailed statistics is provided in Table 2. We use the cumulative matching characteristics (CMC), especially Rank-1 score, and mean average precision (mAP) as evaluation metrics to compare state of the art methods.

4.2. Experiment settings

For training, the input images are resized to 256×128. Hence, the feature map of an each image is size of 16×8. And we adopt three different augmentation methods, including random horizontal flipping, padding and random cropping, and random erasing (RE) [35]. RE augmentation is more effective in person ReID since person are often occluded with different objects.

In hard triplet embedding, we set $P = 32$ and $K = 3$ to train our network. And the margin m is set to be 0.3 for the Market1501 and DukeMTMC-ReID datasets, and 1.2 for the CUHK03-NP dataset. We choice different margin empirically because CUHK03-NP has half images compared to others and the number of cameras is much smaller.

For the optimization, we set λ_c and ϵ to be 5×10^{-4}

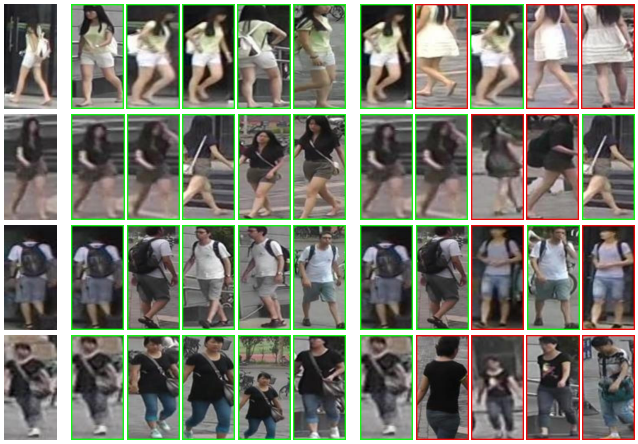

Figure 4: Top 5 ranking for query images on Market1501. The images in the leftmost column are the query images. The image from 2nd to 6th column and from 7th to 11th column show the ranking results of MSReID and BagOfTricks, respectively. The correct images which have the same id as the given query are highlighted by green borders.

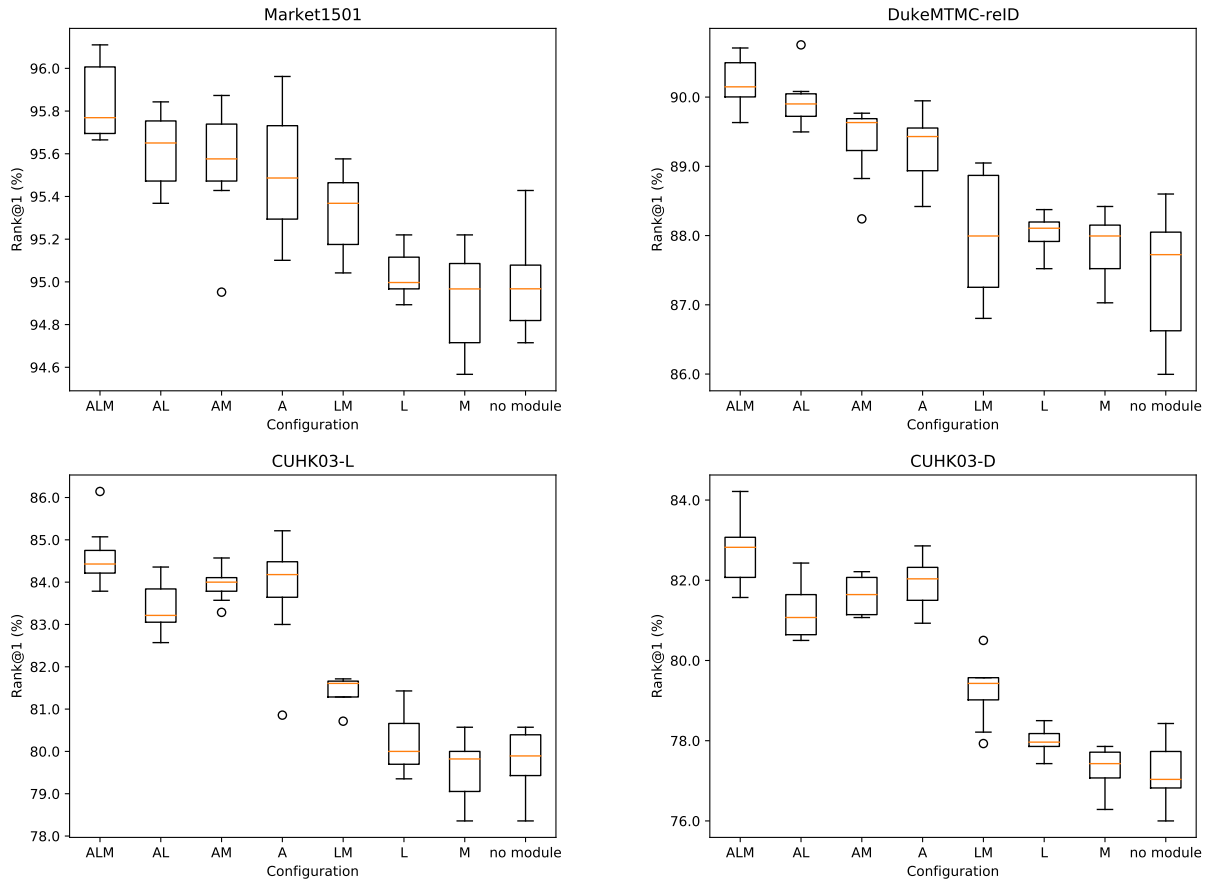

Figure 5: Box plot of the model with specific configuration. Each model is evaluated 8 times. The models with ALM is the best performance on each ReID dataset.

and 0.1, respectively. And we use Adam optimizer except center parameters. we also use learning rate warmup. For an initial 10 epochs, learning rate is linearly increased from 1.5×10^{-4} to 1.5×10^{-3}. During training, the learning rate is decreased 10 times every 60 epoch. Only center parameters are trained with SGD optimizer without learning rate warmup. Training is finished at 180 epoch.

All experiments were implemented on 1 Titan X Pascal GPU. We also used automatic mixed precision (AMP) to overcome the GPU memory insufficient since the memory is almost full on batch size of 64. Empirically, AMP does not affect training results.

4.3. Comparision with state of the arts

We compare our proposed method with current state of the art methods. Other methods are categorized into three groups, mask guided, attention based, and others. We repeated the experiment 8 times in the same setting, and the average value is reported in Table 1. Comparing the mask guided and attention methods, our MSReID use a mask which is similar to the mask guided method, but the mask is generated without any semantic annotations. It leads that global and the local features are trained separately. And we use soft attention which is proposed in [13].

For Market1501 dataset, we only report the single query (SQ) results without any special processing such as re-ranking. As shown in Table 1, we achieves the second best performance. In our eight experiments, we obtained an average of 95.8% Rank-1 and 88.8% mAP score. This result close to the state of the art method, RGA-SC [32]. Separately, MSReID obtained up to 96.0% Rank-1 and 89.6% mAP for Market1501, which outperforms all existing methods including RGA-SC [32]. Compared to BagOfTricks [14], it achieves a large margin, Rank-1 of 1.3% and mAP 2.9%.

For DukeMTMC-ReID dataset, our architecture outperforms other sate of the arts methods at least $1.2\%/1.0\%$ in Rank-1/mAP, compared to the second best method [1]. In our experiments, Rank-1 score of MSReID is up to 90.7%. Note that our MSReID is the first model that can achieve above 90% on Rank-1.

CUHK03-NP dataset is much more challenging ReID

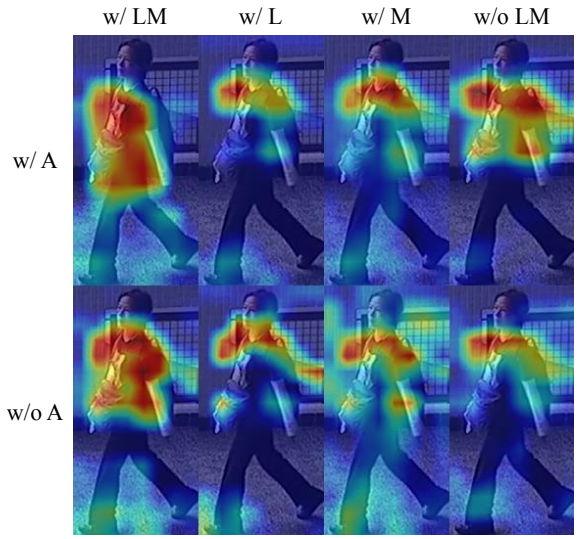

Figure 6: The attention map visualization of MSReID with GradCAM [18]. First row images are the result of models with soft attention module and second row is not. The column images indicate the results of the model with LM, L, M, and without LM, respectively.

benchmark than Market1501 and DukeMTMC-ReID becuase there are only 2 cameras and the number of train/test images are small compared with others. CUHK03-NP is divided into two types. One is manually labeled dataset and the other is detected with a pedestrian detector. CUHK03-D dataset is more realistic since captured images are not often aligned well. For two types of dataset, RGA-SC [32] is the second best as shown in Table 1. Our architecture obtains a large margin compared with other methods, over $3.5\%/4.5\%$ on labeled dataset and $3.1\%/5\%$ on detected dataset. It suggest that MSReID is more powerful on challenging dataset.

For qualitative results, we evaluate the top 5 ranking (Rank-5) results on Market1501. We tested with our MSReID and BagOfTricks [14] and the results are illustrated in Figure 4. Compared to BagOfTricks, our method distinguishes different clothes of the same color, easily. And it retrieves the different view images of the same person.

4.4. Ablation study

To analyze the effect of indivisual components in MSReID, we conducted ablation evaluations on Market1501 [33], DukeMTMC-reID [17], and CUHK03-NP [12].

We use three modules in MSReID, including the soft attention (A), the local feature (L), and the mask (M). To analyze the effect of each module, we trained various models with or without each module. There are 8 possible configurations from ALM to no module. As shown in Figure 5, we evaluate each model 8 times with same experiment settings to measure stable performance. The box plots show that the performance is better when used together than using the local feature (L) and mask (M) respectively on all ReID dataset. It suggest that our mask separated learning method is effective on image retrieval task. Moreover, the soft attention (A) increased the overall performance regardless of any module.

In addition, we visualized the attention map using GradCAM [18] to see how each module affected the original image. Figure 6 shows the attention map for each module. The attention maps of L and M are similar or smaller in the attention area than when the module was not used. But the attention map focus on the human body properly when two modules are applied together. For the soft attention, the background clutter in attention maps is removed and attention map is emphasized than original one which does not use attention module.

5. Conclusion

In this paper, we have proposed a mask separated feature learning to solve the person ReID task. MSReID is a simply implemented with backbone network but effective. Our experiments on three person ReID dataset show that MSReID achieved state of the art performance. Ablations show that the combining the mask and the local feature contribute to its performance than that of each separately. In the future, we will design the new training methods which the mask are trained with semi-supervised learning methods to guide the person body parts.

References

[1] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9637–9646, 2019.

[2] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8351–8361, 2019.

[3] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3691–3701, 2019.

[4] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[10] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[11] Xu Lan, Hanxiao Wang, Shaogang Gong, and Xiatian Zhu. Deep reinforcement learning attention selection for person re-identification. *arXiv preprint arXiv:1707.02785*, 2017.

[12] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[13] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[14] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[15] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[16] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864*, 2018.

[17] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

[18] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[19] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018.

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin 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)*, pages 480–496, 2018.

[21] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin 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)*, pages 480–496, 2018.

[22] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[24] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, 2018.

[25] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[26] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.

[27] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[29] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Local convolutional neural networks for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1074–1082, 2018.

[30] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[31] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019.

[32] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention. *arXiv preprint arXiv:1904.02998*, 2019.

[33] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[34] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[35] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[36] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019.