Mask-guided Contrastive Attention Model for Person Re-Identification

Chunfeng Song[1,3] Yan Huang[1,3] Wanli Ouyang[4] Liang Wang[1,2,3]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
[4]University of Sydney
{chunfeng.song, yhuang, wangliang}@nlpr.ia.ac.cn wanli.ouyang@sydney.edu.au

Abstract

Person Re-identification (ReID) is an important yet challenging task in computer vision. Due to the diverse background clutters, variations on viewpoints and body poses, it is far from solved. How to extract discriminative and robust features invariant to background clutters is the core problem. In this paper, we first introduce the binary segmentation masks to construct synthetic RGB-Mask pairs as inputs, then we design a mask-guided contrastive attention model (MGCAM) to learn features separately from the body and background regions. Moreover, we propose a novel region-level triplet loss to restrain the features learnt from different regions, i.e., pulling the features from the full image and body region close, whereas pushing the features from backgrounds away. We may be the first one to successfully introduce the binary mask into person ReID task and the first one to propose region-level contrastive learning. We evaluate the proposed method on three public datasets, including MARS, Market-1501 and CUHK03. Extensive experimental results show that the proposed method is effective and achieves the state-of-the-art results. Mask and code will be released upon request.

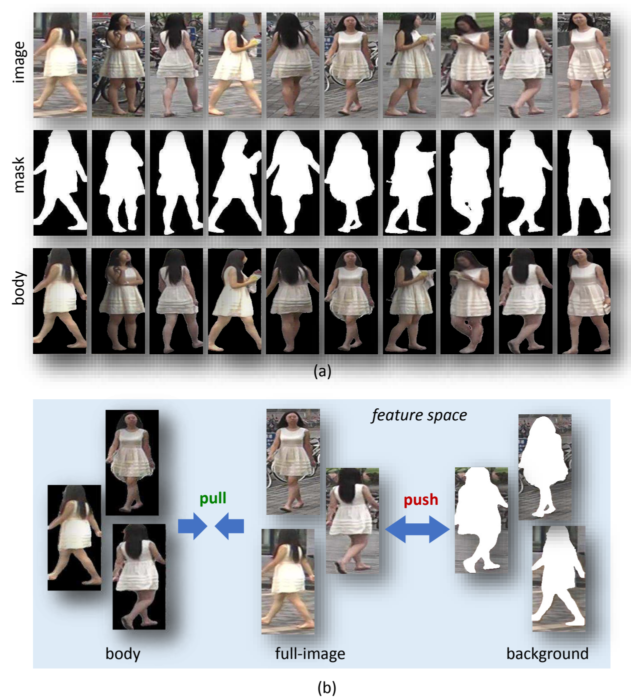

Figure 1. Illustration of the binary mask and region-level triplet loss for contrastive feature learning. (a) Examples of RGB images and their corresponding masks. The third row shows the body regions extracted directly with the masks. (b) The proposed region-level triplet loss can restrain the features learnt from different regions, i.e., pulling the features from the full image and body region close, whereas pushing the features from backgrounds away.

1. Introduction

Person Re-identification (ReID) plays an important role in various surveillance applications, such as pedestrian retrieval and public security event detection. In general, for a given probe person, ReID is to identify the same person across multiple cameras. It is still a challenging problem due to various body poses, view of cameras, illumination, and cluttered backgrounds. In the past years, numerous of methods [47, 39, 25, 36, 26, 37, 43, 34] have been proposed to address this problem. Most of previous methods directly learn features from the whole image, which contains not only the person body, but also the background clutters. Recently, several deep learning based methods are proposed to learn identity features from the body parts which are generated by either part region detection [24], or pose and key-points estimation [32, 23, 49, 44]. These methods have been proved effective through extracting features exactly from the body region rather than the background regions in the person image. It indicates that removing the background clutters in person image is helpful for improving the performance of person ReID.

Another solution to handle background clutter is to obtain the human body region by segmentation. Fortunately, with the rapid development of deep learning based image segmentation methods including FCN [28], Mask R-CNN [17] and the building of large scale human segmentation datasets [38, 31], we can obtain much better body mask now, as shown in Figure 1 (a). The generated binary segmentation masks are pretty good, which can accurately remove the backgrounds in person images. The method applied for generating the masks will be introduced in our related work.

The binary body mask can contribute to person ReID in two respects. **Firstly, the mask can help removing the background clutters in pixel-level.** This can greatly improve the robustness of ReID models under various of background conditions. **Secondly, the mask contains body shape information which can be regarded as the important gait features.** It has been proved that the body mask is robust to illumination, cloth colors, and thus is useful for identifying a person [35].

The most straightforward way to utilize the binary body mask is to directly mask the background in the images. With the binary mask, the masked image only contains the body region which is expected to perform better than using the whole image. However, in our experiments, we find the performance of masked images is even slightly worse compared with the one using the original images (refer to Section 4.3 for more details). This result means that directly removing the background with binary mask in a 'hard' manner is not a good choice, which may affect the structured information and smoothness of an image. In addition, the wrongly segmented masks may contain lots of backgrounds or lose some important body parts which will greatly impact the performance. In this case, removing the backgrounds in the feature-level may be a better solution.

To address this problem, we explore to utilize the binary mask to reduce the background clutters in the feature-level. We propose a mask-guided contrastive attention model (MGCAM) to learn features contrastively from the body and background regions. As shown in Figure 1 (b), in the feature space, the features learnt from the body region and the full image should be similar, whereas the features learnt from the background and the full image should be different. To this end, the proposed MGCAM first produces a pair of contrastive attention maps under the guide of the binary body mask. The contrastive attention maps are then added to CNN features to generate body-aware and background-aware features, respectively. Note that our region-level triplet loss is applied on region features from the same image rather than other triplet loss [12] on features from different images.

To learn body shape related features from the binary body mask, we propose to take it as an additional input accompanied with the original RGB image to construct a 4-channeled image. In this way, the CNN model can learn the appearance feature from the RGB channels and learn the body shape feature from the mask channel. So this method works in a relatively 'soft' manner. Even in the worst case, i.e., the mask is totally wrong, the CNN model still can learn features from the RGB channels. Our experiments have proved this method can improve the performance.

The contributions of this paper can be summarized as follows:

- To reduce the background clutters in person images with mask, we design a contrastive attention model which is guided by the binary mask. It can generate a pair of body-aware and background-aware attention maps, which can be used to produce features of body and background.
- We further propose a region-level triplet loss on the features from full image, body and background. It can force the model-learnt features to be invariant to background clutters.
- We explore to take the body mask as an additional input accompanied by the RGB image to enhance the ReID feature learning. The binary mask has two main advantages: 1) it can help reduce the background clutters, and 2) it contains identity related features such as body shape information.

2. Related Work

In this section, we first review some related works in person ReID, especially those deep learning based methods, then we introduce some segmentation approaches related to our method, finally we briefly describe some recent visual attention mechanisms.

Person ReID. Recently, deep learning based person ReID approaches have achieved great success [10, 24, 33, 54, 44] through simultaneously learning the person representation and similarity within one network. These methods usually learn the ID-discriminative Embedding (IDE) feature [48] via training a deep classification network. In addition, some works try to introduce the pair-wise contrastive loss [14], triplet ranking loss [54] and quadruplet loss [8] to further enhance the IDE feature. To combine the classification and pair-wise loss, Chen et al. attempt to apply a multi-task model to simultaneously learn classification and ranking tasks [9]. There are also some works trying to

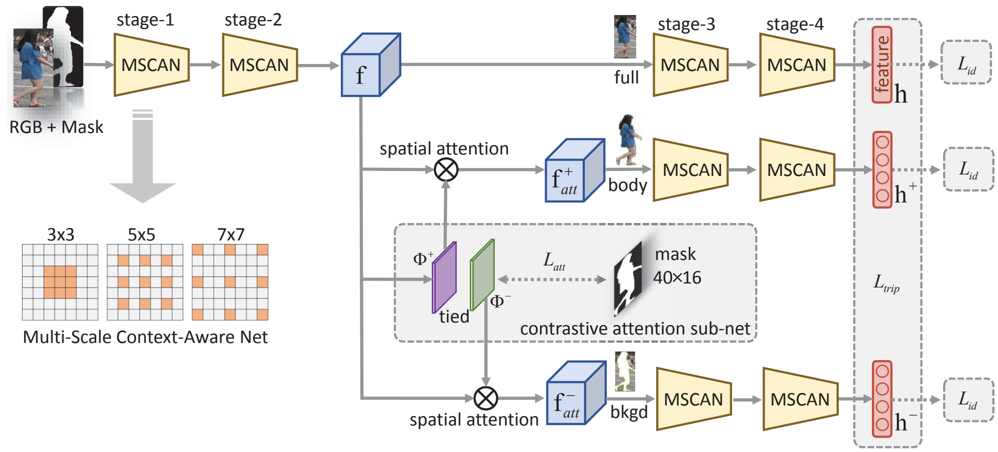

Figure 2. Framework of proposed Mask-guided Contrastive Attention Model (MGCAM) for person ReID. It contains four multi-scale context-aware stages and a fully-connected layer to learn final features. There are three main streams, i.e., the full-stream, the body-stream and the background-stream. In the middle is the contrastive attention sub-net which can generate a pair of body-aware and background-aware attention maps under the guide of binary mask. A region-level triplet loss is implemented on the features learnt from three streams.

implement the multi-scale context [24] or multi-resolution method [29] in person ReID. Note that above methods simply take the whole image as input which may be greatly affected by the background clutters and pose variations. To deeply learn the representations of pedestrian, several body region or part based methods are proposed. Xiao et al. try to combine the person detection and identification model [40]. Li et al. propose a two-stream model to jointly learn the global and part features [24]. Inspired by recent progress in pose estimation [13, 3], several pose based person ReID methods are proposed [44, 32, 23, 49]. Those methods have been proved effective for person ReID, shown that removing backgrounds is helpful for identifying person.

Segmentation Method. There are few works introducing segmentations into person ReID, due to the low quality and computation consuming [36]. With the rapid development of deep learning based image segmentation methods including the Fully Convolutional Networks (FCN) [28], CRF based methods proposed in [6], Mask R-CNN [17] and large scale human segmentation datasets [38, 31], now we can easily obtain much better body mask.

Visual Attention Mechanism. Visual Attention mechanism has achieved great success in computer vision field, such as object detection [5], image segmentation [7] and pose estimation [13]. It is efficient and effective via implementing a spatial attention map across each location of the features. Different from them, we introduce a mask-guided contrastive attention model which can generate a pair of attention maps to attend to the body and background regions in a person image, respectively. We might be the first one to introduce the binary mask guided contrastive attention model for person ReID.

3. Our Proposed Method

We propose the mask-guided contrastive attention model to learn features invariant to cluttered background and selectively learn representations within the body region. The overview of the proposed method is shown in Figure 2. There are two main components, the contrastive attention sub-net and the region-level triplet loss for contrastive feature learning. The first part can generate a pair of inverse attention masks which are used to the body-aware and background-aware feature learning. Whereas the second part restrains the distances between features from the full-stream, the body-stream and the background-stream.

3.1. Overall Architecture

There are variety of network structures introduced or proposed to learn features for person ReID, among which CaffeNet [22] and ResNet-50 [18] are mostly used two. In general, these deep networks should be first pre-trained on Image-net dataset [30] to initialize the large numbers of parameters. However, our method need to take 4-channeled inputs, i.e., the RGB-Mask, which is incompatible with these pre-trained models. Recently, a multi-scale context-aware network (MSCAN) has been proposed which can be trained from scratch [24]. MSCAN achieves the state-of-the-art performance on several person ReID datasets, outperforming the feautres learnt by pre-trained CaffeNet [22]. Therefore, we adopt the body-version MSCAN as our base network, details about MSCAN can refer to [24].

As shown in Figure 2, the adopted MSCAN contains four multi-scale context-aware stages and a fully-connected layer to fuse the learned features. There are three main streams in proposed mask-guided contrastive attention model

(MGCAM), i.e., the full-stream, the body-stream and the background-stream. The full stream learns features from the whole image, which is the same as the body-version MSCAN [24]. The body stream tends to learn the body features with a body-aware attention map. In the contrary, the background steam learns the background features with a background-aware attention map. Above attention maps are generated by the contrastive attention sub-net. Though the features of three streams are learnt from a same image, they are quite different from each other, especially the one learnt from backgrounds which contains almost none useful information related to the identity. In retrospect, a main goal of person ReID is to reduce the background clutters and concentrate on the body region. To this end, a triplet of constrains are added to restrain three features, pushing the background feature far from the whole feature and pulling the body feature close to the whole feature.

For a given person image and mask pair (RGB-M), the MGCAM first produces a middle feature map $f_{stage-2}$ after the second stage, with a size of 96×40×16. Then the sub-net produces a pair of contrastive attention maps with $f_{stage-2}$ as its source inputs. The contrastive attention maps are then added to the body-stream and background-stream, respectively to implement spatial attention. The full-stream directly takes the original feature map $f_{stage-2}$ without any operation. Both of the three streams finally compute a 128-dimension feature vector, representing the features learnt from full image, body, and background, respectively. We select the features of the full-stream for person ReID. In the following subsections, we describe the details of the two main parts of the proposed MGCAM.

3.2. Mask-guided Contrastive Attention Sub-net

In general, spatial attention model is to take the on-going feature as its input and produce a weighting map to carry out spatial-wise attention across the feature map. In this way, the network could attend the exactly spatial regions on the feature map that contribute most for training the model. Given an input sample RGB-M, the feature map after the second stage of MGCAM can be noted as $f_{stage-2}$. Taking $f_{stage-2}$ as inputs, the contrastive attention sub-net then produces a body-aware attention map which can be denoted as

$$\Phi^{+} = \sigma(\mathrm{W} * \mathrm{f}_{stage-2} + \mathrm{b}) \tag{1}$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function, W and b mean the convolutional filter weights and bias. Unlike the attention model in previous works [7, 5], this attention model works in a 'soft' manner with the sigmoid function which is similar with [13]. We then generate an inverse attention map Φ^{-} to attend the contrastive feature. To ensure that Φ^{+} and Φ^{-} can constitute a contrastive attention pair, for each location (i, j) in this pair of attention maps should meet the constraint:

$$\Phi^{-}(i,j) + \Phi^{+}(i,j) = 1 \tag{2}$$

Consequently, we apply this pair of attention maps to the feature $f_{stage-2}$ to produce a pair of contrastive features:

$$\mathrm{f}^{+}_{att} = \mathrm{f}_{stage-2} \otimes \Phi^{+} \tag{3}$$

$$\mathrm{f}^{-}_{att} = \mathrm{f}_{stage-2} \otimes \Phi^{-} \tag{4}$$

where $\otimes$ means the spatial weighting operation. The positive attention map is expected to have high scores in the body region whereas the negative one has low scores. However, as the positive and negative attention maps play equal roles if without other constrains, it is not guaranteed the positive one can learn body-aware map. To give a clear hint, we introduce the body mask to guide the attention map via adding a Mean Squared Error (MSE) loss between the positive attention map and corresponding body segmentation mask:

$$L_{att} = \sum_{i=1}^{I}\sum_{j=1}^{J} \left\| M_{(i,j)} - \Phi_{(i,j)} \right\|_2^2 \tag{5}$$

where M is the body mask which is pre-generated from person image with proper segmentation method (refer to Section 4.1 for details) and resized into the same size of the attention map. Therefore, the mask-guided contrastive attention sub-net can generate contrastive features associating with the body and background separately.

3.3. Region-Level Triplet Loss for Contrastive Feature Learning

With the contrastive attention maps described in last subsection, we further introduce a region-level triplet loss to enhance contrastive feature learning. After the attention operation, features from three main streams can be denoted as f_{full}, f^{+}_{att} and f^{-}_{att} . They are then sent to the following two MSCAN stages to produce the final 128-dimensional feature vectors, noted as h_{full}, h_{body}, and h_{bkgd}, respectively. With this triplet of features, we take h_{full} as the anchor sample, h_{body} be the positive sample, and h_{bkgd} be the negative sample. Then the region-level triplet loss can be defined as

$$\begin{aligned} L_{trip} = &\|\mathrm{h}_{full} - \mathrm{h}_{body}\|_2^2 \\ &+ \max\{(m - \|\mathrm{h}_{full} - \mathrm{h}_{bkgd}\|_2^2), 0\} \end{aligned} \tag{6}$$

where m is a margin parameter which is empirically set to 10 in the experiments. With the minimization of this loss, in the feature space, features from full-stream and body-stream will get close to each other whereas the feature from background will be away. As a result, the feature of the full-stream will become invariant to background clutters and be more aware to body regions, which can enhance the performance in person ReID task.

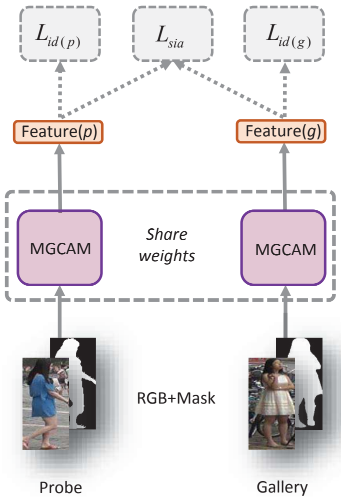

Figure 3. The Siamese network takes a pair of samples as its inputs. The weights are shared across the two branches. The Siamese loss and identity loss are jointly implemented on the features learnt from the probe and gallery.

3.4. Objective Function

We adopt the soft-max regression on the final layers of three streams to predict the identities of persons. For simplicity, we denote the total cross-entropy identity loss of three streams as L_{id}. In addition, we also introduce the Siamese network to pull the features of same instance close and separate the features of different persons, as shown in Figure 3. The two branches of Siamese network can share weights. It should be mentioned that the Siamese network learns pair similarity at instance-level, which is quite different from the proposed region-level loss. Given a pair RGB-M of person p and g, their final features of the full-stream are noted as $\mathrm{h}(p)$ and $\mathrm{h}(g)$, then the loss of the Siamese network can be defined as

$$L_{sia} = \begin{cases} \|\mathrm{h}(p) - \mathrm{h}(g)\|_2^2, & p = g \\ \max\{(m - \|\mathrm{h}(p) - \mathrm{h}(g)\|_2^2), 0\}, & p \neq g \end{cases} \quad (7)$$

where m is a margin parameter which is empirically set to 10 in our experiments. Following the previous work in [41], we also jointly train the network under Siamese loss and identity loss to further improve the performance of person ReID. Taking the region-level loss in MGCAM into consideration, the total loss of a pair of samples (p, g) can be denoted as

$$L_{all} = L_{id(p,g)} + \lambda \cdot L_{sia} + \alpha \cdot L_{trip(p,g)} + \beta \cdot L_{att(p,g)} \quad (8)$$

where λ, α and β are the hypermeters, which are respectively set to 0.01, 0.01 and 0.1 in our experiments. As MGCAM can also be trained without Siamese network, we evaluate both versions of them in our experiments.

Table 1. The details of three datasets used in experiments.

Datasets	MARS[48]	Market-1501[50]	CUHK03[25]
# identities	1,261	1,501	1,467
# boxes	1,191,003	32,668	14,096
# cameras	6	6	2
# resolution	128×256	64×128	vary

3.5. Feature Extraction

As introduced in above subsections, the features of the full-stream are learnt with both the restrains from the region-level triplet loss and the instance-level siamese loss, whereas the features from the other two streams are only used to guide the feature learning of the full stream. Therefore, we take the 128-dimensional feature vector generated from the full-stream as the representation for each sample. This feature is effective for person ReID in three folds: 1) It is invariant to background clutters due to the help of proposed MGCAM. 2) It may contain the body shape features learnt from the mask. 3) It is more discriminative via joint learning with the siamese loss and identity loss.

4. Experiments

In this section, we describe the experimental details and testify the effectiveness of proposed MGCAM on three widely used ReID databases.

4.1. Datasets

We evaluate the proposed method on three large-scale public person ReID datasets, including MARS [48], Market-1501 [50] and CUHK03 [25], details of them are shown in Table 1. Both of the three datasets contain more than one thousand identities and large numbers of images which are close to the practical application.

MARS[48] is the current largest sequence-based person ReID dataset, containing 1,261 identities with each identity captured by six cameras. There are 20,478 video sequences and 1,191,003 bounding boxes which are generated by a DPM detector [16] and a GMMCP tracker [15]. All images are with a resolution of 128×256. Following [48], we use 625 identities for training and the rest 631 identities for testing.

Market-1501[50] contains 1,501 identities which are captured by cameras from 6 different viewpoints. There are 32,668 pedestrian images which are labeled by bounding boxes with a DPM detector [16]. Each person has 3.6 images on average at each viewpoint. The dataset is spilt into two parts: 751 identities are used for training and the rest 750 identities are used for testing. In the testing phase, following the same setting of [46], 3,368 hand-drawn person images are selected as probe set to query the correct identities across the testing set.

CUHK03[25] contains 1,467 identities which are captured by several surveillance cameras. Each identity is cap-

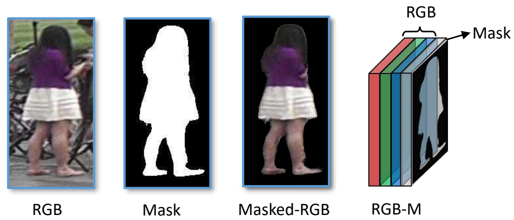

Figure 4. Different inputs for person ReID. With the binary mask, we can generate a synthetic RGB+Mask pair (RGB-M in short), including three RGB channels and one mask channel.

tured from two disjoint cameras. There are 14,096 person images in total and each identity has roughly 4.8 images in each viewpoint. This dataset provides two types of bounding boxes annotations, including the manually annotated bounding boxes, and DPM-detected bounding boxes. We evaluate the proposed method on both types of them. Following [52], we adopt the new training/testing protocol to split the dataset into two balanced parts: 767 identities are in the training set and the rest 700 identities are in the testing set.

4.2. Implementation Details

Base Model Selection. As described in Section 3, there are variety of network structures introduced or proposed to learn features for person ReID, among which CaffeNet [22] and ResNet-50 [18] are the mostly used two. In general, these deep networks should be pre-trained on Imagenet [30] due to their large numbers of parameters. However, our methods need to take a 4-channel RGB-M inputs (shown in Figure 4) which is incompatible with these pre-trained models. Recently, a multi-scale context-aware network (MSCAN) has been proposed which can be trained from scratch [24]. It achieves state-of-the-art performance on several person ReID datasets. Therefore, we adopt the simplest version MSCAN-body as our base network. More details of MSCAN can refer to [24].

Data Pre-processing. For each image, we first generate a binary segmentation mask corresponding to the body and background region with a FCN [28] based segmentation model which is trained on labeled human segmentation datasets such as [38, 31]. Most masks are satisfying even for the images with complex backgrounds. There are also some failures caused by the wrongly detected images. Besides the RGB images and masks, we generate two kinds of inputs as shown in Figure 4. The first one is to directly mask the original RGB images to remove the background regions, noted as Masked-RGB. The second one is to keep both the RGB and the mask channels to compose a synthetic RGB+Mask pair, noted as RGB-M. Therefore, we get four kinds of image-like inputs. We first resize each inputs into 160×64, then normalize them via subtracting the mean

Table 2. Evaluate the inputs on the MARS dataset. All the results are measured with the XQDA distance metric.

Methods	Inputs	Rank-1	mAP
MSCAN-body[24]	RGB	69.70	52.41
	Mask	29.34	12.83
	Masked RGB	68.13	51.49
	RGB-M	**71.26**	**55.44**
Ours(hard)	RGB-M	70.40	54.27
Ours	RGB	72.83	57.39
	RGB-M	**74.19**	**59.13**

values and scale them with a factor of 1/256. We also implement randomly mirror as basic data augmentation in the training phase.

Optimization. All the models are trained on Caffe framework [19]. We first train our model without Siamese network for roughly 7.5×10^4 iterations using an initial learning rate of 0.01, and decrease it after each 1.5×10^4 iterations. For each iteration, we randomly select 128 samples across the whole dataset for training. This well-trained model is noted as MGCAM. Then we add the Siamese network as shown in Figure 3 and fine-tune the whole model with an initial learning rate of 0.001. We gradually decrease the learning rate until the loss stop dropping. We note this well-trained model as MGCAM-Siamese. Finally, we evaluate both of the models and compare with previous state-of-the-art methods.

Evaluation Metrics. We use the 128-dim feature vector generated from the full-image stream as the representation for each person inputs. Then we compute the distance between probe and gallery samples with several classic metrics, including conventional Eulidean distance, XQDA [26], KISSME [21] and the recently proposed Re-ranking methods [52]. As Re-ranking method has several variations in [52], we take the XQDA+Re-ranking version as default in this paper. Finally, we use the Cumulative Matching Characteristic (CMC) [2] curve and Mean Average Precision (mAP) [25] to evaluate the performance of proposed methods on three datasets. Considering that ReID is a ranking problem, we report the *single-query* rank-1 cumulated matching accuracy following [52].

4.3. Effectiveness of Proposed Method

On the MARS dataset, we evaluate the effectiveness of the proposed method comprehensively. We first explore the influence with different inputs, then compare the proposed methods with the baseline model.

4.3.1 Evaluate the Effect of Mask

With the generated segmentation masks, we can train CNN models with them in different manners. As shown in Figure 4, there are four kinds of inputs: original RGB image, binary mask, masked RGB image and the synthetic RGB+Mask

Table 3. Evaluate the effectiveness of MGCAM on the MARS dataset. All methods take the RGB-M as their inputs.

Methods	Distance Metric	Rank-1	mAP
MSCAN-body [24]	Eulidean	71.21	54.92
	KISSME	67.22	47.47
	XQDA	71.26	55.44
	Re-ranking	72.32	66.01
Ours	Eulidean	74.29	59.59
	KISSME	70.96	51.26
	XQDA	74.19	59.13
	Re-ranking	76.01	70.13
Ours-Siamese	Eulidean	75.66	61.29
	KISSME	72.42	53.13
	XQDA	75.35	60.34
	Re-ranking	**77.17**	**71.17**

Table 4. Results on the MARS dataset.

Methods	Ref	Rank1	mAP
CNN+XQDA[48]	ECCV2016	65.3	47.6
MSCAN-body [24]	CVPR2017	68.23	51.82
SFT[54]	CVPR2017	70.6	50.7
IDE+XQDA [52]	CVPR2017	70.51	55.12
MSCAN-Fusion[24]	CVPR2017	71.77	56.05
IDE+XQDA+Rerank [52]	CVPR2017	73.94	68.45
Ours		76.01	70.13
Ours-Siamese		**77.17**	**71.17**

pairs (RGB-M). We train the baseline model MSCAN-body [24] with four kinds of inputs respectively and report the results in Table 2. We also compare the RGB images and RGB-M on proposed MGCAM. It is obvious that the RGB-M performs better than RGB in both models. Therefore, in the following evaluation and experiments, we take the RGB-M as default inputs for proposed methods. We can draw three conclusions from the results:

- Mask is useful. Only taking mask as inputs can achieve 29.34% rank-1 accuracy showing that the mask contains useful information associated with the identity, such as the body shape, the ratio between head and shoulders.
- The masked RGB images perform a little bad showing that removing the background in a hard manner is not a good choice. This may affect the structured information and smoothness of an image. It also results in completely failure in case of the mask is wrongly generated.
- The RGB-M performs the best indicating that it can keep both the appearance feature from RGB and body shape feature from mask. Taking the masks as additional inputs can enhance the CNN in two aspects: 1) Mask contains human shape feature and is robust to illumination and clothing colors. 2) Mask can provide apparent hints for CNN to distinguish human body and background regions in original RGB image.

Table 5. Results on the Market-1501 dataset.

Methods	Ref	Rank1	mAP
BOW[50]	ICCV 2015	34.4	14.09
PersonNet [37]	arXiv 2016	37.21	18.57
WARCA [20]	ECCV 2016	45.16	-
SCSP [4]	CVPR 2016	51.9	26.35
DNS [43]	CVPR 2016	61.02	35.68
Gated [34]	ECCV 2016	65.88	39.55
Point-to-Set[53]	CVPR 2017	70.72	44.27
CCAFA [11]	TPAMI 2017	71.8	45.5
Consistent-Aware [27]	CVPR 2017	73.84	47.11
Spindle [44]	CVPR 2017	76.9	–
Re-ranking [52]	CVPR 2017	77.11	63.63
GAN [51]	ICCV 2017	78.06	56.23
MSCAN [24]	CVPR 2017	80.31	57.53
DLPAR [45]	ICCV 2017	81.0	63.4
Scalable [1]	CVPR 2017	82.21	68.8
DaF [42]	BMVC 2017	82.3	72.42
SVDNet [33]	ICCV 2017	82.3	62.1
Ours		83.55	74.25
Ours-Siamese		**83.79**	**74.33**

Table 6. Results on the CUHK03 dataset.

Methods	Ref	Labeled		Detected	
		Rank1	mAP	Rank1	mAP
BOW[50]	ICCV2015	7.93	9.29	6.36	6.39
LOMO[26]	CVPR2015	14.8	13.6	12.8	11.5
DaF [42]	BMVC2017	27.5	31.5	26.4	30.0
Re-rank [52]	CVPR2017	38.1	40.3	34.7	37.4
SVDNet [33]	ICCV2017	40.93	37.83	41.5	37.26
DPFL [10]	ICCV2017	43.0	40.5	40.7	37.0
Ours		49.29	49.89	46.29	46.74
Ours-Siamese		**50.14**	**50.21**	**46.71**	**46.87**

4.3.2 Evaluate the Effect of MGCAM

We further compare two versions of the proposed method with the baseline method [24] to evaluate their improvements. Experiments are conducted with four classic distance metrics, including Euclidean distance, XQDA [26], KISSME [21] and the recently proposed Re-ranking methods [52]. The comparison results are shown in Table 3. We report the results of proposed MGCAM with and without the Siamese loss, to testify the effectiveness of each component in contrastive attention model. For qualitative evaluation, we also visualize the learnt attention maps in Figure 5. Different from the binary mask which has a constant weight value for each spatial location, the soft attention map has large weights at the more important parts such as the head and colorful cloth, whereas has small weights for the less important parts such as legs and arms without clothes, which are less informative to identify a person. We can draw the following conclusions from the experimental results from Table 2 and Table 3.

- By comparing our MGCAM and the baseline model with two kinds of inputs in Table 2 and four kinds of distance metrics in Table 3, we can find that the proposed MGCAM is more effective. The results in Table

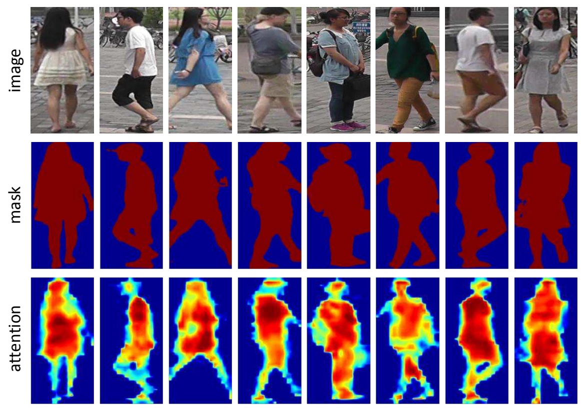

Figure 5. Visualization of the masks and learnt attention maps.

2 also shows that the hard mode which directly using the binary mask as the hard attention map performs worse than the soft mode.

- Siamese loss can further enhance the performance of MGCAM, showing the region-level triplet loss is compatible with the instance-level Siamese loss.
- The results shown in Table 3 also show that, even with the same feature, the results will vary when different distance metrics are adopted. Among which the recently proposed re-ranking method [52] performs the best. Thus we take this metric as default to compare with the state-of-the-art methods in the following subsection.

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

Above experiments have shown proposed MGCAM taking the RGB-M as inputs can achieve satisfying performance. To verify the generalization of our method, we compare with the state-of-the-art methods on three popular ReID datasets in the following parts.

MARS: On this dataset, we compare with several recent proposed state-of-the-art methods, including the pioneer method CNN+XQDA [48], the baseline method MSCAN [24], the SFT [54] method which jointly learn both the spatial and temporal representations in one framework, and the Re-ranking methods presented in [52]. Only single query is evaluated and compared on MARS. The overall experimental results are shown in Table 4. Our MGCAM-Siamese achieves 77.17% rank-1 accuracy and 71.17% mAP, outperforming the compared state-of-the-art methods. Note that our model is trained from scratch without any pre-training, showing our method is robust and effective.

Market-1501: As this datasets is one of the most used large scale ReID dataset, we compare our approach with a series of state-of-the-art methods, and list the results in Table 5. The compared methods include the body part-based and pose-based methods, such as Spindle-Net [44], Deeply-Learned Part-Aligned Representations (DLPAR) [45], as well as the fusion version of MSCAN [24]. These methods tend to remove the background clutters and fusion the features of the body regions. Experimental results show that our method achieves satisfying results through simultaneously using RGB-M as inputs and the mask-guided contrastive attention mechanism.

CUHK03: For the CUHK03 dataset, we evaluate our methods on both the detected and labeled parts. Following the protocols in [52], we compare with the most recent state-of-the-art methods in terms of both rank-1 accuracy and mAP under single query. The compared methods include the Deep Pyramid Feature Learning (DPFL) [10], SVDNet [33], and two re-ranking methods: DaF [42] and Re-ranking [52]. As shown in Table 6, our method outperforms the compared methods with an obvious margin, showing the advantages of proposed method. Note that our method is using the same distance metric with Re-ranking [52]. Compared with the features learnt on ResNet-50 [18] model by Re-ranking [52], the features learnt on our methods improve both the rank-1 accuracy and mAP by at least 10 percent. It further shows the effectiveness of our method.

5. Conclusion

In this paper, we propose a novel method to extract discriminative and robust features invariant to background clutters. To address this problem, we first introduce the binary segmentation masks to construct synthetic RGB-Mask pairs as inputs, then we design a mask-guided contrastive attention model (MGCAM) to learn features separately from the body and background regions. Moreover, we propose a novel region-level triplet loss to restrain the features learned from different regions, i.e., pulling the features from the full image and body region close, whereas pushing the features from backgrounds away. Extensive experimental results show that the proposed method is effective and achieves the state-of-the-art results.

Acknowledgement

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015), Beijing Natural Science Foundation (4162058), and Capital Science and Technology Leading Talent Training Project (Z181100006318030). Wanli Ouyang is supported by SenseTime Group Limited. This work is also supported by grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer.

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017. 7

[2] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. The relation between the roc curve and the cm-c. In *IEEE Workshop on Automatic Identification Advanced Technologies*, 2005. 6

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[4] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 7

[5] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017. 3, 4

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 3

[7] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 3, 4

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

[9] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *AAAI*, 2017. 2

[10] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *CVPR*, 2017. 2, 7, 8

[11] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE TPAMI*, 2017. 7

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

[13] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017. 3, 4

[14] D. Chung, K. Tahboub, and E. J. Delp. A two stream siamese convolutional neural network for person re-identification. In *CVPR*, 2017. 2

[15] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, 2015. 5

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 5

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6, 8

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM ICM*, 2014. 6

[20] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *ECCV*, 2016. 7

[21] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 6, 7

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 6

[23] V. Kumar, A. Namboodiri, M. Paluri, and C. Jawahar. Pose-aware person recognition. In *CVPR*, 2017. 2, 3

[24] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 2, 3, 4, 6, 7, 8

[25] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 5, 6

[26] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 6, 7

[27] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *CVPR*, 2017. 7

[28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 3, 6

[29] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *CVPR*, 2017. 3

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *I-JCV*, 2015. 3, 6

[31] C. Song, Y. Huang, Z. Wang, and L. Wang. 1000fps human segmentation with deep convolutional neural networks. In *ACPR*, 2015. 2, 3, 6

[32] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017. 2, 3

[33] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 2, 7, 8

[34] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 1, 7

[35] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE TPA-MI*, 25(12):1505–1518, 2003. 2

[36] X. Wang and R. Zhao. *Person Re-identification: System Design and Evaluation Overview*. Springer London, 2014. 1, 3

[37] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arX-iv preprint arXiv:1601.07255*, 2016. 1, 7

[38] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan. Early hierarchical contexts learned by convolutional networks for image segmentation. In *ICPR*, 2014. 2, 3, 6

[39] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 1

[40] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 3

[41] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, 2017. 5

[42] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. In *BMVC*, 2017. 7, 8

[43] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 1, 7

[44] 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. In *CVPR*, 2017. 2, 3, 7, 8

[45] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 7, 8

[46] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 5

[47] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. *IEEE TPAMI*, 39(2):356–370, 2016. 1

[48] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 2, 5, 7, 8

[49] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017. 2, 3

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5, 7

[51] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 7

[52] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 6, 7, 8

[53] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, 2017. 7

[54] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017. 2, 7, 8