Mask-guided Contrastive Attention Model for Person Re-Identification

Chunfeng Song\textsuperscript{1,3}, Yan Huang\textsuperscript{1,3}, Wanli Ouyang\textsuperscript{4}, Liang Wang\textsuperscript{1,2,3}
\textsuperscript{1}Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR)
\textsuperscript{2}Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Institute of Automation, Chinese Academy of Sciences (CASIA)
\textsuperscript{3}University of Chinese Academy of Sciences (UCAS)
\textsuperscript{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.

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 gener-
ated 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 per-
formance of person ReID.

Another solution to handle background clutter is to ob-
tain 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 seg-
mentation masks are pretty good, which can accurately re-
move the backgrounds in person images. The method ap-
plied for generating the masks will be introduced in our re-
lated 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 impor-
tant 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 com-
pared with the one using the original images (refer to Sec-
tion 4.3 for more details). This result means that directly re-
move the background with binary mask in a ‘hard’ man-
ner 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 bina-
ry mask to reduce the background clutters in the feature-
level. We propose a mask-guided contrastive attention mod-
el (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 differen-
t. 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 im-
age rather than other triplet loss [12] on features from dif-
ferent 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 back-
  ground clutters.
- We explore to take the body mask as an additional in-
  put 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 clut-
  ters, 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 per-
son 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 ReI-
D approaches have achieved great success [10, 24, 33, 54,
44] through simultaneously learning the person represen-
tation and similarity within one network. These methods
usually learn the ID-discriminative Embedding (IDE) fea-
ture [48] via training a deep classification network. In addi-
tion, 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 clas-
sification and pair-wise loss, Chen et al. attempt to ap-
ply a multi-task model to simultaneously learn classification
and ranking tasks [9]. There are also some works trying to
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 features 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 stream 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 \( f_{\text{stage} - 2} \) after the second stage, with a size of 96×40×16. Then the sub-net produces a pair of contrastive attention maps with \( f_{\text{stage} - 2} \) 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 \( f_{\text{stage} - 2} \) 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. In the subsequent operation, features from three main streams can be denoted as \( f_{\text{full}}, f_{\text{att}} \) and \( f_{\text{att}}^\text{bkgd} \). They are then sent to the following two MSCAN stages to produce the final 128-dimensional feature vectors, noted as \( h_{\text{full}}, h_{\text{body}}, \) and \( h_{\text{bkgd}} \), respectively. With this triplet of features, we take \( h_{\text{full}} \) as the anchor sample, \( h_{\text{body}} \) be the positive sample, and \( h_{\text{bkgd}} \) be the negative sample. Then the region-level triplet loss can be defined as

\[
L_{\text{trip}} = \| h_{\text{full}} - h_{\text{body}} \|_2^2 + \max \{ (m - \| h_{\text{full}} - h_{\text{bkgd}} \|_2^2), 0 \}
\]

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.

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

\[
\begin{align*}
\Phi^+ (i, j) &= \sigma (W * f_{\text{stage} - 2} + b) \\
\Phi^- (i, j) &= \Phi^+ (i, j) = 1 \\
\end{align*}
\]

\[
\begin{align*}
f^+_\text{att} &= f_{\text{stage} - 2} \otimes \Phi^+ \\
f^-\text{att} &= f_{\text{stage} - 2} \otimes \Phi^-
\end{align*}
\]

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:

\[
L_{\text{att}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \| M_{(i,j)} - \Phi (i,j) \|_2^2
\]
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 \( h(p) \) and \( h(g) \), then the loss of the Siamese network can be defined as

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

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

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

where \( \lambda, \alpha \) and \( \beta \) are the hyperparameters, 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.

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 \( 128 \times 256 \). 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 split 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.

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 $160 \times 64$, then normalize them via subtracting the mean 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 $7.5 \times 10^4$ iterations using an initial learning rate of 0.01, and decrease it after each $1.5 \times 10^4$ 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
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.

### 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

### Table 1. Distance Metrics

<table>
<thead>
<tr>
<th>Methods</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>71.21</td>
<td>54.92</td>
</tr>
<tr>
<td>KISSME</td>
<td>67.22</td>
<td>47.47</td>
</tr>
<tr>
<td>XQDA</td>
<td>71.26</td>
<td>55.44</td>
</tr>
<tr>
<td>Re-ranking</td>
<td>72.32</td>
<td>66.01</td>
</tr>
</tbody>
</table>

### Table 2. Results on the MARS dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Distance Metric</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+XQDA</td>
<td>ECCV2016</td>
<td>65.3</td>
<td>47.6</td>
</tr>
<tr>
<td>MSCAN-body</td>
<td>CVPR2017</td>
<td>68.23</td>
<td>51.82</td>
</tr>
<tr>
<td>SFT[54]</td>
<td>CVPR2017</td>
<td>70.6</td>
<td>50.7</td>
</tr>
<tr>
<td>IDE+XQDA[52]</td>
<td>CVPR2017</td>
<td>70.51</td>
<td>55.12</td>
</tr>
<tr>
<td>MSCAN-Fusion</td>
<td>CVPR2017</td>
<td>71.77</td>
<td>56.05</td>
</tr>
<tr>
<td>IDE+XQDA+Rerank[52]</td>
<td>CVPR2017</td>
<td>73.94</td>
<td>68.45</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Distance Metric</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>Euclidean</td>
<td>74.29</td>
<td>59.59</td>
</tr>
<tr>
<td>KISSME</td>
<td>70.96</td>
<td>51.26</td>
<td></td>
</tr>
<tr>
<td>XQDA</td>
<td>74.19</td>
<td>59.13</td>
<td></td>
</tr>
<tr>
<td>Re-ranking</td>
<td>76.01</td>
<td>70.13</td>
<td></td>
</tr>
<tr>
<td>Ours-Siamese</td>
<td>Euclidean</td>
<td>75.66</td>
<td>61.29</td>
</tr>
<tr>
<td>KISSME</td>
<td>72.42</td>
<td>53.13</td>
<td></td>
</tr>
<tr>
<td>XQDA</td>
<td>75.35</td>
<td>60.34</td>
<td></td>
</tr>
<tr>
<td>Re-ranking</td>
<td>77.17</td>
<td>71.17</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4. Results on the MARS dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Ref</th>
<th>Rank-1</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN+XQDA[48]</td>
<td>ECCV2016</td>
<td>65.3</td>
<td>47.6</td>
</tr>
<tr>
<td>MSCAN-body[24]</td>
<td>CVPR2017</td>
<td>68.23</td>
<td>51.82</td>
</tr>
<tr>
<td>SFT[54]</td>
<td>CVPR2017</td>
<td>70.6</td>
<td>50.7</td>
</tr>
<tr>
<td>IDE+XQDA[52]</td>
<td>CVPR2017</td>
<td>70.51</td>
<td>55.12</td>
</tr>
<tr>
<td>MSCAN-Fusion</td>
<td>CVPR2017</td>
<td>71.77</td>
<td>56.05</td>
</tr>
<tr>
<td>IDE+XQDA+Rerank[52]</td>
<td>CVPR2017</td>
<td>73.94</td>
<td>68.45</td>
</tr>
<tr>
<td>Ours</td>
<td>76.01</td>
<td>70.13</td>
<td></td>
</tr>
<tr>
<td>Ours-Siamese</td>
<td>77.17</td>
<td>71.17</td>
<td></td>
</tr>
</tbody>
</table>

### Table 5. Results on the Market-1501 dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Ref</th>
<th>Labeled</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>PersonNet[37]</td>
<td>arXiv2016</td>
<td>14.8</td>
<td>13.6</td>
</tr>
<tr>
<td>SCSP[4]</td>
<td>CvPR2016</td>
<td>27.5</td>
<td>31.5</td>
</tr>
<tr>
<td>DNS[43]</td>
<td>CVPR2016</td>
<td>38.1</td>
<td>40.3</td>
</tr>
<tr>
<td>Gated[34]</td>
<td>ECCV2016</td>
<td>40.93</td>
<td>37.83</td>
</tr>
<tr>
<td>Point-to-Set[53]</td>
<td>CVPR2017</td>
<td>43.0</td>
<td>40.5</td>
</tr>
<tr>
<td>Ours</td>
<td>49.29</td>
<td>49.89</td>
<td>46.29</td>
</tr>
<tr>
<td>Ours-Siamese</td>
<td>50.14</td>
<td>50.21</td>
<td>46.71</td>
</tr>
</tbody>
</table>

### Table 6. Results on the CUHK03 dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Ref</th>
<th>Labeled</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>PersonNet[37]</td>
<td>arXiv2016</td>
<td>14.8</td>
<td>13.6</td>
</tr>
<tr>
<td>SCSP[4]</td>
<td>CvPR2016</td>
<td>27.5</td>
<td>31.5</td>
</tr>
<tr>
<td>DNS[43]</td>
<td>CVPR2016</td>
<td>38.1</td>
<td>40.3</td>
</tr>
<tr>
<td>Gated[34]</td>
<td>ECCV2016</td>
<td>40.93</td>
<td>37.83</td>
</tr>
<tr>
<td>Point-to-Set[53]</td>
<td>CVPR2017</td>
<td>43.0</td>
<td>40.5</td>
</tr>
<tr>
<td>Ours</td>
<td>49.29</td>
<td>49.89</td>
<td>46.29</td>
</tr>
<tr>
<td>Ours-Siamese</td>
<td>50.14</td>
<td>50.21</td>
<td>46.71</td>
</tr>
</tbody>
</table>
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.

\section*{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.

\textbf{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.

\textbf{Market-1501}: As this dataset 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.

\textbf{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.

\section*{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.

\section*{Acknowledgement}

This work is jointly supported by National Key Research and Development Program of China (2016YF-B1001000), National Natural Science Foundation of China (61525306, 61633021, 161721004, 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


