IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 27, 2020, accepted July 19, 2020, date of publication July 27, 2020, date of current version August 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3011961

A Mask-Pooling Model With Local-Level Triplet Loss for Person Re-Identification

FUDAN ZHENG[1], TINGTING CAI[1], YING WANG[2], CHUFU DENG[1], ZHIGUANG CHEN[1,2], AND HUILING ZHU[3]

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[2]National Supercomputer Center in Guangzhou, Guangzhou 510006, China
[3]College of Information Science and Technology, Jinan University, Guangzhou 510632, China

Corresponding author: Zhiguang Chen (zhiguang.chen@nscc-gz.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0203904; in part by the National Natural Science Foundation of China under Grant 61872392; in part by the National Natural Science Foundation of China for Young Scientists under Grant 11701592; in part by the Joint Funds of the National Natural Science Foundation of China under Grant U1811263; in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2016ZT06D211; and in part by the Pearl River Science and Technology (S&T) Nova Program of Guangzhou under Grant 201906010008.

ABSTRACT Person Re-Identification (ReID) is an important yet challenging task in computer vision. Background clutter is one of the greatest challenges to overcome. In this paper, we propose a Mask-pooling model with local-level triplet loss (MPM-LTL) to tackle this problem and improve person ReID performance. Specifically, we present a novel pooling method, called mask pooling (MP), to gradually remove background features in feature maps through deep convolutional network. With mask pooling, the network can learn the most crucial person features. Moreover, we raise a novel local-level triplet loss for discriminative feature training. Furthermore, we propose a new hard triplets selection algorithm named Mask-guided TriHard. The method is based on human outline information, which is, to our best knowledge, employed for the first time for hard triplets selection. We achieve the state-of-the-art results on three benchmark person datasets Market-1501 [1], CUHK03 [2] and DukeMTMC-reID [3], [4].

INDEX TERMS Person re-identification, mask-pooling, hard triplets selection, local-level triplet loss.

I. INTRODUCTION

Person re-identification (ReID), which aims at identifying the same person among different cameras, has drawn increasing attention in computer vision since it plays a critical role in pedestrian retrieval, public security and criminal investigation [5]. It is a challenging problem due to large variations in person pose, illumination and viewpoint of cameras, occlusion, low image resolution and cluttered backgrounds, as illustrated in Fig. 1.

In the past few years, a variety of approaches have been proposed to address these difficulties. When learning features, most of these methods make use of the advantages of various deep neural networks and obtain better results than traditional image processing methods [6]–[11]. However, they tend to focus on the features of the entire image, which include not only the whole body of the person but also the cluttered background. Only until recently have local features, such as body regions and joints [12]–[14], been used and proved to be more discriminative than global features. However, these local features are obtained either by horizontally segmenting the entire image or by dividing the whole image according to the body parts of the person [15]–[19], which still contain large amounts of background information, as shown in Fig.2 (a)-(c). One of our intuitions is that, background information may, to some extent,

FIGURE 1. Challenges in Person ReID: (a) variant person poses, (b) variant illumination, (c) variant camera viewpoints, (d) occlusions, (e) low image resolutions, (f) cluttered background. All images are from dataset Market-1501. Best viewed in color.

The associate editor coordinating the review of this manuscript and approving it for publication was Paolo Remagnino.

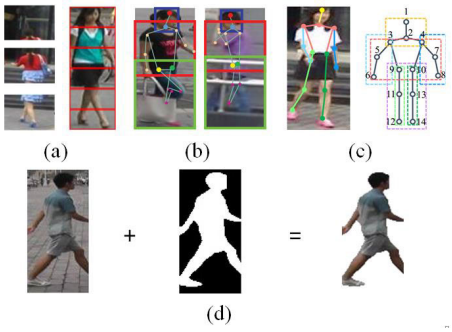


FIGURE 2. (a) - (c) are partition methods in part feature learning, which still contain large amounts of background information: (a) horizontally segmentation [16], [20], (b) key points and part regions [13], (c) body regions and joints [12]. (d) illustrates the motivation of our method: to remove the extra background using mask information. Best viewed in color.

be a barrier to human feature extraction. To verify this intuition, we use the mask information, which is transformed into a new channel (mask channel) of the images, to remove the extra background, as shown in Fig.2 (d). In our model, this mask channel, which reflects whether each pixel in the image belongs to background or non-background (0 for background while 1 for non-background), is taken together with the normal RGB channels as input. Background information is not removed at the very beginning, but gradually removed during the feature extraction process of the network. For this mask channel, we do not apply the convolution operation like the other three channels because the mask channel corresponds to the foreground/background information, which can be used to guide the network to better focus on the most important foreground information. In addition, other resize methods, such as linear interpolation, may blur the edges of the mask's contour. Therefore, we propose a new operation, called mask pooling, to highlight the role of mask in the network. Mask pooling differs from other traditional pooling methods in that, instead of simply downsampling based on mean or maximum value, it selectively downsamples according to background or non-background information, which can greatly retain meaningful information for feature extraction.

Various loss methods are applied to guide training. Since most work only focus on global features of the whole image, the loss methods they adopt are limited to global loss only, regardless of whether they use Contrastive loss [21], Triplet loss [9], [11], [15], [22]–[24], TriHard loss [8], [25] or Quadruplet loss [26]. Even in studies that learn part features, most of them still aggregate part features into a whole one and train it with a global identification loss. Unlike them, we divide the features extracted into several parts, design a classifier for each part, and introduce a local-level triplet loss to help training. It is worth mentioning that, while commonly used triplet loss is applied to global features, the triplet loss we propose is at local level.

In addition, since easy triplets contribute little to training, it is crucial and challenging to select hard triplets that are more instructive. With the aid of the extracted segmentation mask, we carefully select hard images to form triplets for local-level triplet loss, which improves the performance.

Specifically, we make the following three contributions:

- We propose a new pooling method called mask pooling. After moderate mask is extracted and transformed into a mask channel, mask pooling is employed to gradually remove background features in feature maps during the process of feature extraction and downsampling. With mask information and mask pooling, the network can learn the most crucial person features.
- We propose a local-level triplet loss for the training of each part of the extracted features. Different from usual triplet loss, which is always applied to the whole feature of an image, the proposed local-level triplet loss is applied to the corresponding parts of the triplet and proved to be better for discriminative feature training.
- We present a new hard triplets selection algorithm called Mask-guided TriHard for local-level triplet loss, which is better for network training. In addition to feature similarity, human outline information in the mask are taken into account when selecting hard triplets. The calculation of the mask intersection can guide the network to avoid selecting totally irrelevant body parts, which facilitates the alignment of the partitioned images and thus is beneficial to local-level triplet training.

II. RELATED WORK

A. DEEP LEARNING METHODS

Deep learning methods have been widely used in Person ReID since 2014 [2], [27]. These methods have shown to be more effective than traditional image processing methods. A recent trend is to design a deep neural network to learn features and metrics simultaneously. In addition to the most frequently used CNN [5]–[12], [16], [28], RNN and its variants (LSTM, GRU, etc.) [20], [29] are also used to extract temporal features, especially in video-based Person ReID.

B. PART FEATURE LEARNING

Among deeply learned features, part features draw increasing attention and have been proved to be more discriminative [18], [30], [32], [33]. For instance, [14], [32], [34], [35] employ pose information to help part feature learning, and [13], [15]–[20] partition pedestrians into several parts (horizontal stripes, rectangle blocks, etc.) to extract part features. The most challenging problem of the above pose-driven methods and partition methods lies in body part misalignment [36]. Body joints [12], pose boxes [34], keypoints [13] and semantic features of different body regions [12], [14] are effective means to address this problem. Inspired by these methods, we propose a novel model that divides pedestrian into several parts. Instead of aggregating the feature of each part into a whole feature and training it with a whole loss, for each part, we design a classifier and train it by a multi-classification loss.

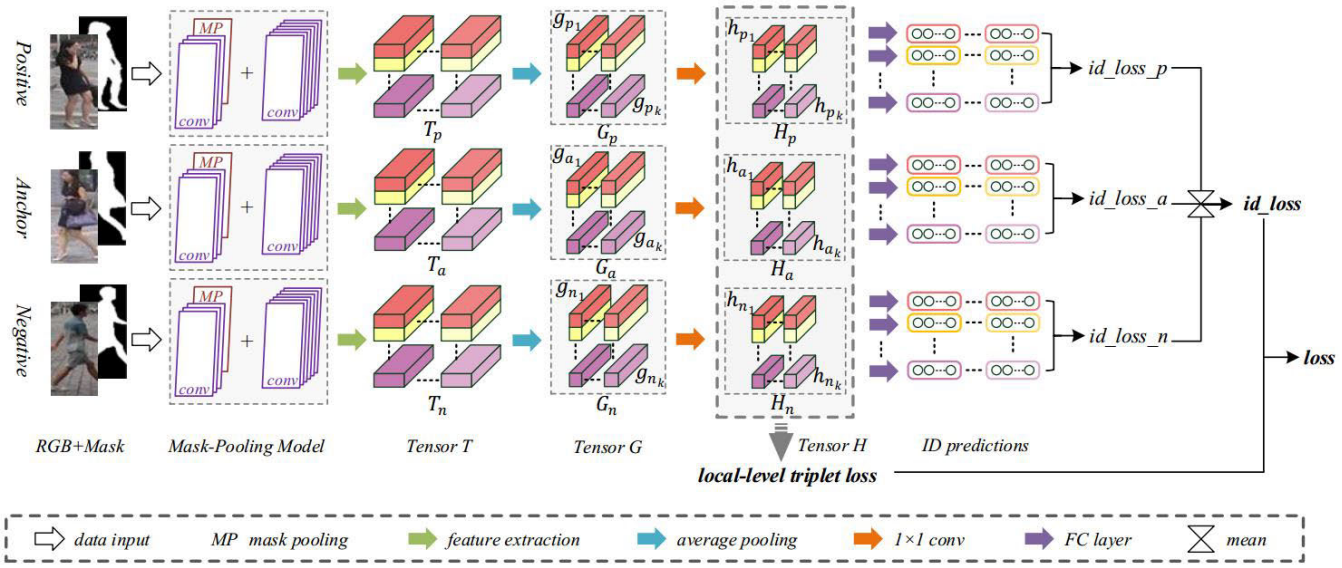


FIGURE 3. The proposed MPM-LTL model. The network takes synthetic images which contain mask information as inputs. After going through a Mask-pooling model, extracted part features are trained by local-level triplet loss and identification loss. Best viewed in color.

C. MASK-BASED FEATURE LEARNING

With the rapid development of deep learning based image segmentation methods including FCN [37] and Mask R-CNN [38], segmentation mask of the objects in an image can be well extracted and utilized in feature learning. DyeNet [39] contributes an mask-based approach, which is robust to distractors not belonging to the target segment, to perform person tracking in videos. CNN+MGTS [40] generates segmentation mask to emphasize foreground information, with the motivation that foreground information are more vital to re-identify a person. NWAPI [41] proposes an end-to-end noise weakened person ReID system by first suppress the background noise using Mask R-CNN. MGCAM [42] also designs a mask-guided contrastive attention model to learn features separately from the body and background regions. Different from NWAPI which combines panoramic features and foreground features, and MGCAM which directly adds the extracted foreground information into the images and carries out the ordinary convolution operation, we propose mask pooling, a special pooling operation only for the mask channel, to make full use of the mask to learn the most discriminate non-background features, and achieve much better results than NWAPI and MGCAM.

D. LOSS METHODS

The commonly used loss methods for metric learning include Contrastive loss [21], Triplet loss [9], [11], [15], [22]–[24], TriHard loss [8], [25], Quadruplet loss [26], and Margin sample mining loss [43]. Besides these global-level loss methods, part loss [30] is also adopted for training part features. The using of triplet loss is not novel in Person ReID task. However, all the triplet loss methods mentioned above are in global level, which compute triplet loss based on the characteristics of the entire feature maps. Unlike them, the triplet loss our model adopts is in local level, which pays more attention to local information aligned in the feature maps. More specifically, we use a local-level triplet loss to pull close the distance between parts in positive pairs of pedestrians and push away the distance between parts in negative pairs of pedestrians.

III. OUR APPROACH

A. OVERALL ARCHITECTURE

As shown in Fig.3, in our MPM-LTL model, a triplet of input images (marked as Positive, Anchor and Negative respectively) are sent into the branch network, called Mask-pooling model (MPM). These inputs contain not only the original images but also the corresponding masks, which are good for extracting non-background information. The network is deliberately designed to make full use of the masks (to be detailed in Section III-C). Extracted by MPM, each 3D feature tensor, T_p, T_a and T_n, is divided into k parts of the same size. An average pooling is performed to every part of the feature map of each channel, turning T_p, T_a and T_n into G_p, G_a and G_n with k vectors respectively. Then, k 1×1 convolutional layers are employed for dimension reducing. So finally we get $H_p = \{h_{p_1}, h_{p_2}, \ldots, h_{p_k}\}$, $H_a = \{h_{a_1}, h_{a_2}, \ldots, h_{a_k}\}$ and $H_n = \{h_{n_1}, h_{n_2}, \ldots, h_{n_k}\}$ as the final feature descriptors. Since part feature training has been proven to be very effective in [18], we associate each h_{p_i}, h_{a_i} and h_{n_i} $(1 \leq i \leq k)$ with a classifier. During training, each classifier predicts the identity of the corresponding part and is supervised by Cross-Entropy loss. Different from [18], corresponding to the triplet of inputs, we design a triplet of identification loss, i.e., *id_loss_p*, *id_loss_a* and *id_loss_n*, and define the final identification loss, *id_loss*, as the average of the three.

In order to fully utilize local features, we design a local-level triplet loss named *local_level_triplet_loss* to train the similarity between every part in the feature map of the triplet of images. In other words, each triplet of h_{p_i}, h_{a_i} and h_{n_i} $(1 \leq i \leq k)$, is trained by *local_level_triplet_loss*.

Therefore, the total loss consists of two parts: while *id_loss* is responsible for the training of each part of the image itself, supervised by the label of the image, *local_level_triplet_loss* is in charge of the learning of the corresponding part in anchor, positive and negative images, which aims to pull close the distance between the positive image pairs and push away the distance between the negative image pairs.

Details of *id_loss* and *local_level_triplet_loss* will be discussed in Section III-F and Section III-D respectively.

B. MASK EXTRACTION

Although Mask R-CNN [38] is effective for mask extraction on ordinary image datasets, it does not quite apply to Person ReID datasets, since images in Person ReID datasets are already cut by bounding boxes generated by hand labeling or detectors.

To better extract person mask from Person ReID images, we deal with the original datasets as follows:

- Firstly, since Person ReID images are all with low resolutions, we expand every original image by padding 0 around it (with the original image in the center). The expanded image has 3 times the width and height of the original image.
- Then, we perform mask extraction using a modified Mask R-CNN, with anchor areas of $\{32^2, 64^2\}$ and aspect ratios of $\{1{:}2, 1{:}1, 2{:}1\}$.
- Finally, in order to avoid loss of foreground edge due to low image resolution, we extend the mask contour by 3 pixels.

C. MASK-POOLING MODEL

Our model can take any deep convolutional neural network as backbone network. In this paper, we employ ResNet50 [44] with the consideration that it has competitive performance and relatively concise architecture.

Since the extracted mask is binary information, it can well reflect whether each pixel in the image belongs to background or non-background (0 for background while 1 for non-background). We transform this binary information into a mask channel. Unlike normal networks, which receive 3-channel images (RGB) as input, our network takes 4-channel images as input (RGB + Mask).

To highlight the role of mask in the network, we make the following modifications to ResNet50, as can be seen in Fig. 4:

- For conv1 layer in ResNet50, while the first 3 channels (RGB) perform normal convolution (7×7 convolution, stride 2, 63 kernels), BatchNorm normalization and ReLU activation, the 4^{th} channel (mask channel) performs a 7×7 pooling. It is not an ordinary average pooling or maximum pooling, but a special pooling (we name it mask pooling) based on whether the pixel values of the current pooling region (7×7) represent background or non-background. In this mask pooling, the proportion of non-background pixels (that is, the proportion of ones in the mask channel) in each 7×7 pooling region is calculated. We believe that if this proportion exceeds a certain threshold, it is likely that this pooling region represents non-background information and should be pooled as non-background. Our intuition is that, too large threshold can lead to loss of pedestrian information, while too small threshold may bring about too much background information, which may cause interference. Multiple experiments suggest that 0.4-0.7 (40%-70%) are moderate threshold, which is consistent with our intuition. (Details will be discussed in Section IV-B.)
- For the pooling part in conv2_x layer in ResNet50, while the first 63 channels perform a normal 3×3 max pooling, the 64^{th} channel performs a 3×3 mask pooling.

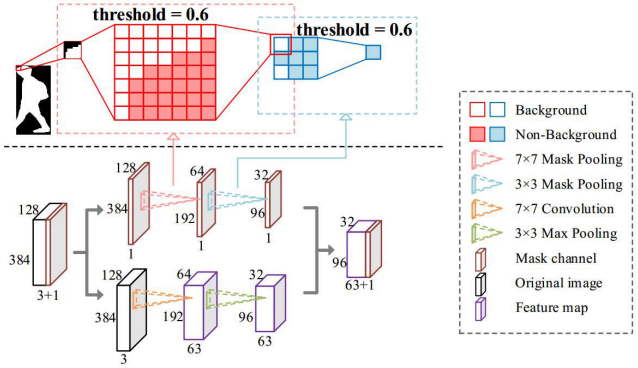


FIGURE 4. Illustration of mask pooling. If the proportion of non-background pixels in each pooling region exceeds a certain threshold (take threshold = 0.6 for example), then the pooling region will be pooled as non-background. Note that the mask pooling is performed in a single channel, while the 7×7 convolution and the 3×3 max pooling are performed with 63 kernels, which results in 63 channels, and they add up to 64 channels. Best viewed in color.

Fig.5 illustrates the feature maps extracted after mask pooling in the modified ResNet50. It is obvious that after mask pooling in conv2_x layer, the contour features of pedestrians are already well preserved, and most of the cluttered background is effectively removed. Therefore, there is no need to perform the mask pooling operation for the subsequent layers. So the 64 channels stack up again and are fed into the rest of the ResNet50 network.

D. LOCAL-LEVEL TRIPLET LOSS

Triplet loss [22], [23] is a widely used loss method for metric learning. In Person ReID, it is common that images with different IDs are very similar, while images with the same ID differ greatly due to variant person poses, illumination and viewpoints of cameras. Triplet loss takes into account both the distance between positive image pairs and between

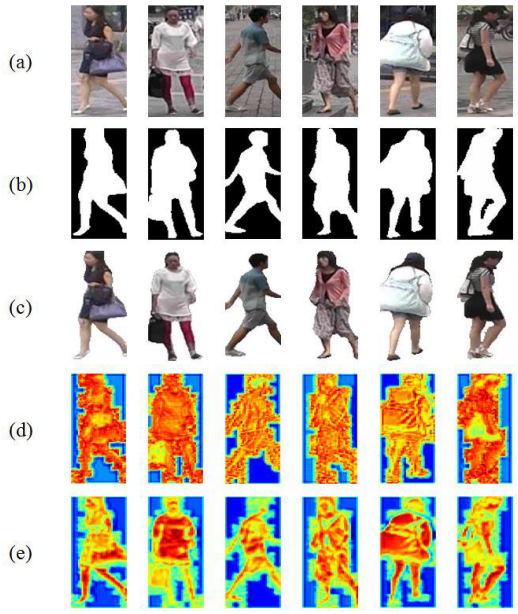


FIGURE 5. Illustration of the feature maps extracted after mask pooling. Line (a) are the original images in Market-1501. Line (b) are their extracted masks. Line (c) are images obtained by incorporating the masks into the fourth channel of the original images. Line (d) are the feature maps extracted after mask pooling in conv1 layer in the modified ResNet5. Line (e) are the feature maps extracted after mask pooling in conv2_x layer in the modified ResNet50. It can be seen that the information extracted is almost the information of pedestrians. Best viewed in color.

negative image pairs, so it is better to solve the problem. Also, constructing triplets is a way of data enhancement, which can effectively alleviate overfitting. Triplet loss requires three input images, including a pair of positive images and a pair of negative images. The three images are respectively named Anchor (a), Positive (p) and Negative (n), where image a and image p form a pair of positive images, and image a and image n form a pair of negative images. Triplet loss can be defined as

$$L_{triplet} = (d_{a,p} - d_{a,n} + \alpha)_+ \tag{1}$$

where $d_{a,p}$ and $d_{a,n}$ are the distance between positive image pairs and the distance between negative image pairs respectively, α is a distance margin between different identities. Training by triplet loss, the network will learn to pull close the distance between positive image pairs and push away the distance between negative image pairs.

In the proposed network, after dividing the feature map into k parts, we do not directly calculate the triplet loss of the whole feature map. Instead, we calculate the triplet loss of each part, and take the average to represent the whole triplet loss. In other words, we define local-level triplet loss *local_level_triplet_loss* as:

$$L_{local\_level\_triplet\_loss} = \frac{1}{k}\sum_{i=1}^{k}(d^i_{a,p} - d^i_{a,n} + \alpha)_+ \tag{2}$$

where k is the number of parts, $d^i_{a,p}$ is the distance between the i^{th} part of the positive image pair and $d^i_{a,n}$ is the distance between the i^{th} part of the negative image pair, respectively.

It is worth mentioning that some triplets, for example, triplets with similar positive/negative image pairs, are not contributing to training and will result in slower convergence. Therefore, it is crucial to select hard triplets that are helpful for training. The following subsection discusses the approach we use for hard triplets selection.

E. HARD TRIPLETS SELECTION FOR TRIPLET LOSS

To construct promising triplets, for each image in training set (anchor), we raise an algorithm called Mask-guided TriHard, which selects the corresponding positive and negative images depending on not only the distance between feature maps, but also their mask information.

Take positive image pairs for example. Generally speaking, less close positive image pairs are more valuable for training. However, if the two images in a pair present completely different body parts, the training is of little significance. This is also true for negative image pairs. So we use Euclidean distance and mask intersection to help sieving images. On the one hand, positive image pairs with larger Euclidean distance are obviously more helpful for training. On the other hand, training two completely different body parts as if they were the same would interfere with the network's ability to learn. Therefore, to avoid this situation, when selecting hard triplets, we use the mask information to exclude pairs of images with too little mask intersection.

The steps of the algorithm are as follows. For an anchor image, we first construct candidate positive and negative image set. Positive image set include images that have the same identity of a person with the anchor image but are in different cameras, while negative image set include images that have different identity with the anchor image but are in the same camera. Then, the top 10 images with the furthest and nearest European distance to the feature map of anchor image are left in the positive image set and negative image set respectively, while the others are removed. Further, mask intersection are calculated and images with the top 5 largest mask intersection with the anchor image are left for both the positive image set and negative image set. Finally, the positive and negative image which make the distance between the positive and negative image pairs closest are selected to form a triplet with the anchor image.

Details of Mask-guided TriHard are shown in Algorithm 1.

F. IDENTIFICATION LOSS

We design an identification loss for the training of every anchor, positive and negative image in each triplet based on Tensor H. The identification loss is a Cross-Entropy loss. Take identification loss for anchor image *id_loss_a* for example, it can be calculated as:

$$L_{id\_loss\_a} = -\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_a} l(a_j) log p(a^i_j) \tag{3}$$

where k is the number of parts, n_a is the number of anchor images, a^i_j is the i^{th} part of the anchor image a_j, $l(\cdot)$ is the

Algorithm 1 Mask-Guided TriHard: A Hard Triplets Selection Algorithm for Triplet Loss

Input: The training set S_{train}
Output: Hard triplet set T

begin
$P = []$ //candidate positive image set
$N = []$ //candidate negative image set
for $a \in S_{train}$ **do**
// subscript *id* stands for the identity of a person
// subscript *cam* stands for the camera of an image
$P = \{p | p_{id} = a_{id}, p_{cam} \neq a_{cam}\}$
$N = \{n | n_{id} \neq a_{id}, n_{cam} = a_{cam}\}$
// calculate the Euclidean distance between the feature map of a and p, a and n:
$D_{a,P} = \left\{d_{a,p} | p \in P\right\}$
$D_{a,N} = \left\{d_{a,n} | n \in N\right\}$
update $P = P \cap \left\{p | Top\_10\_largest(D_{a,P})\right\}$
update $N = N \cap \left\{n | Top\_10\_smallest(D_{a,N})\right\}$
// calculate mask intersection of a and p, a and n:
$m_{a,p} = \sum_{i,j}(a \circ p)_{ij}$
$m_{a,n} = \sum_{i,j}(a \circ n)_{ij}$
$M_{a,P} = \left\{m_{a,p} | p \in P\right\}$
$M_{a,N} = \left\{m_{a,n} | n \in N\right\}$
update $P = P \cap \left\{p | Top\_5\_largest(M_{a,P})\right\}$
update $N = N \cap \left\{n | Top\_5\_largest(M_{a,N})\right\}$
// calculate the absolute value of $(d_{a,n} - d_{a,p})$
$ABS_{a,p,n} = \left\{abs(d_{a,n} - d_{a,p}) | p \in P, n \in N\right\}$
$(p, n) = \arg\min_{(p,n)}(ABS_{a,p,n})$
triplet $t_a = (a, p, n)$
end for
return $T = \{t_a | a \in S_{train}\}$
end

label of the image, and $p(\cdot)$ is the predicted value of the part of the image.

The final identification loss L_{id_loss} is the average of $L_{id\_loss\_a}$, $L_{id\_loss\_p}$ and $L_{id\_loss\_n}$:

$$L_{id\_loss} = -\frac{1}{3k}\sum_{i=1}^{k}\Big(\sum_{j=1}^{n_a} l(a_j) logp(a_j^i) + \sum_{j=1}^{n_p} l(p_j) logp(p_j^i) + \sum_{j=1}^{n_n} l(n_j) logp(n_j^i)\Big) \quad (4)$$

And the total loss is the sum of *local_level_triplet_loss* and *id_loss*:

$$L = \frac{1}{k}\sum_{i=1}^{k}((d_{a,p}^i - d_{a,n}^i + \alpha)_+ - \frac{1}{3}\Big(\sum_{j=1}^{n_a} l(a_j) logp(a_j^i) + \sum_{j=1}^{n_p} l(p_j) logp(p_j^i) + \sum_{j=1}^{n_n} l(n_j) logp(n_j^i)\Big)) \quad (5)$$

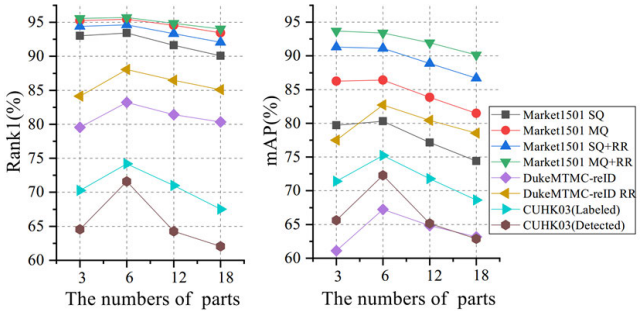


FIGURE 6. Illustration of the numbers of parts to be divided after feature extraction by MPM. Both 3 and 6 parts mean to divide the feature tensor into horizontal stripes. 12 and 18 parts mean to divide the feature tensor into 6 parts horizontally, and 2 and 3 parts vertically, respectively. Best viewed in color.

IV. EXPERIMENTS

A. DATASETS AND SETTINGS

We evaluate our proposed method on three public person ReID datasets: Market-1501 [1], CUHK03 [2] and DukeMTMC-reID [3], [4].

1) MARKET-1501

Market-1501 is collected in front of a supermarket in Tsinghua University by 6 cameras. It contains 1, 501 identities and 32, 668 bounding boxes generated by a DPM detector [45]. Within all these 1, 501 identities, 751 are used for training and the rest 750 are used for testing. It also provides false alarm detection results for training. We use the evaluation packages provided by [1].

2) CUHK03

CUHK03 contains 1, 467 identities and 14, 096 person images which are also captured by six surveillance cameras. Misalignment, occlusions, body part missing, illumination changes are quite common in this dataset. It offers both hand-labeled and DPM-detected bounding boxes. We adopt both the original training/testing protocol (20 random train/test splits) [2] and the new protocol (767 identities for training and the rest 700 for testing) [46].

3) DukeMTMC-reID

DukeMTMC-reID is a subset of DukeMTMC [3] captured by 8 cameras for multi-camera tracking. There are 1,404 identities appearing in more than two cameras. Hand-drawn pedestrian bounding boxes are available. We adopt the training/testing protocol in [4], which randomly select 702 identities as the training set and the remaining 702 identities as the testing set.

4) EVALUATION METRICS

We use the Cumulative Matching Characteristic (CMC) [47] curve and Mean Average Precision (mAP) [1] to evaluate the performance of the proposed methods.

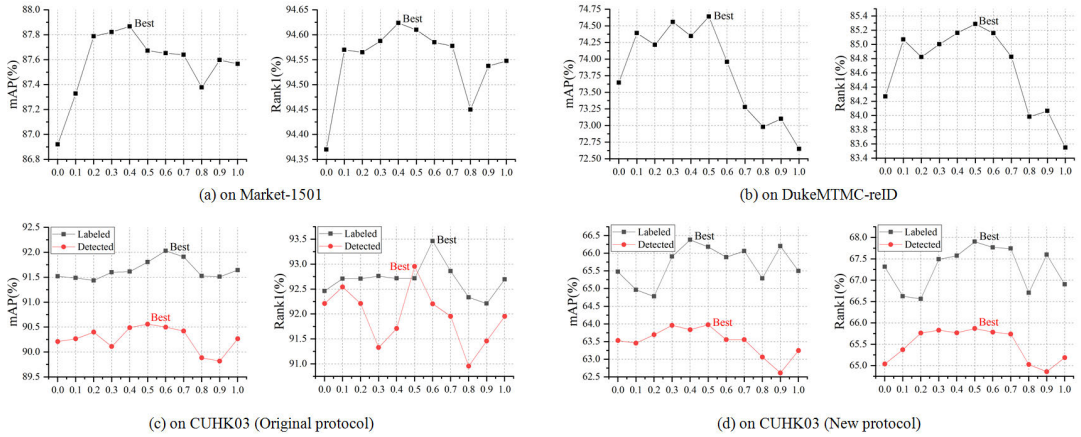


FIGURE 7. Illustration of the mAP and Rank-1 performances achieved by different thresholds on Market-1501, DukeMTMC-reID and CUHK03 (with both original protocol and new protocol). Best viewed in color.

B. IMPLEMENTATION DETAILS

1) TRAINING

We train the model in 3 stages:

- **Stage 1.** The Mask-pooling model (base network) is pre-trained on ImageNet. After that, we set batch size to 64 and train the whole model with *id_loss* for 60 epochs. The base learning rate is initialized at 0.01 for the base network, and 0.1 for the rest convolutional layers and the fully connection layer, and decays to 0.1× of the base learning rate after 40 epochs.
- **Stage 2.** Fix the weights of the whole network except the fully connection layer and train the fully connection layer for 20 epochs. The learning rate is set to 0.001 without decay.
- **Stage 3.** Train the whole model for 20 epochs with learning rate set to be 0.001 for the base network and the fully connection layer, and 0.01 for the rest of the model.

2) PARAMETERS SETTINGS

Crucial parameters of MPM-LTL are set as follows:

- As mentioned in Section III-A, extracted by MPM, each 3D feature tensor, T_p, T_a and T_n, is divided into k parts of the same size. Lots of experiments on the three datasets show that it is best to horizontally divide the feature tensors into 6 parts, as can be seen from Fig. 6. Since the images collected for Person ReID have been generated through detector or manual annotation, in most cases, the 6 parts divided can roughly represent different parts of the body, such as the head, neck, feet and so on.
- As mentioned in Section III-B, since images in Person ReID datasets are already cut by bounding boxes, it is not suitable to directly use Mask R-CNN to extract the mask. Thus, we expand the images. Considering the pixel size of the original image, we expand the images by 3 times and extend 3 pixels around the mask contour after extracting the mask.
- The size of the input image is set 384 × 128.
- As described in Section III-C and Fig.4, during mask pooling, the proportion of non-background pixels (that is, the proportion of ones in the mask channel) in each 7 × 7 or 3 × 3 pooling region is calculated. We believe that if this proportion exceeds a certain threshold, it is likely that this pooling region represents non-background information and should be pooled as non-background. Our intuition is that, too large threshold can lead to loss of pedestrian information, while too small threshold may bring about too much background information, which may cause interference. We have done multiple experiments to verify this intuition. Fig.7 illustrates the performances achieved by different thresholds on Market-1501, DukeMTMC-reID and CUHK03 (with both the original and new protocol). For simplicity, the mAP and Rank-1 values here are the average of the different query patterns (Single-Query, Multi-Query, Single-Query + Re-ranking, Multi-Query + Re-ranking). As Fig.7 shows, performances achieve best in threshold = 0.4 on Market-1501, in threshold = 0.5 on DukeMTMC-reID, in threshold = 0.5/0.6 on CUHK03 (with the original protocol) and in threshold = 0.4/0.5 on CUHK03 (with the new protocol). Besides, Fig.7 suggests that 0.4-0.7 are moderate thresholds, which is consistent with our intuition. To further summarize the results from a macro perspective, we divide the thresholds into three groups, i.e., threshold = 0.0 to 0.3, threshold = 0.4 to 0.7, and threshold = 0.8 to 1.0, and investigate the performances under various query patterns on the three datasets. mAP and Rank-1 values are averaged according to the three groups. Experimental results in Fig.8 shows that for all query patterns, threshold = 0.4 to 0.7 achieves the best performance on all datasets. In future studies, we will try to make the network learn the threshold automatically.

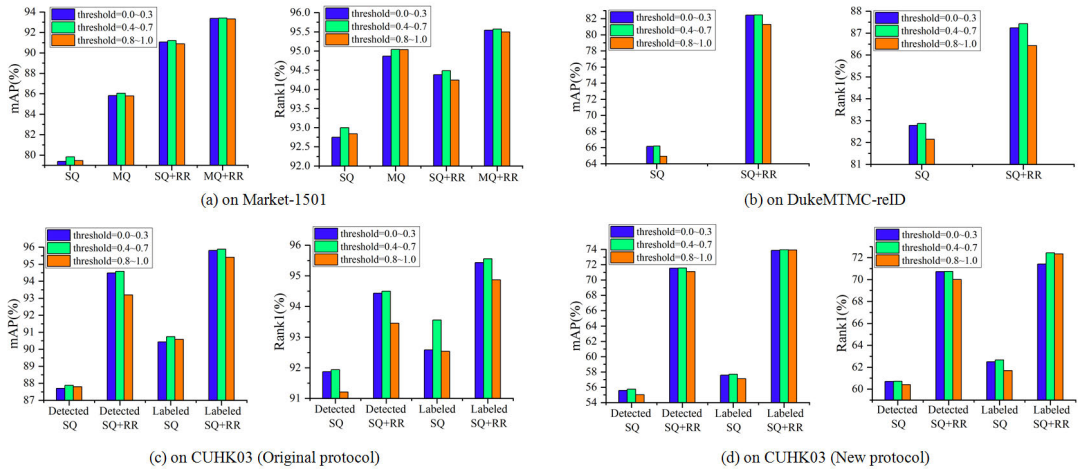


FIGURE 8. Illustration of the mAP and Rank-1 performances achieved by different groups of thresholds on Market-1501, DukeMTMC-reID and CUHK03 (with both original protocol and new protocol). Best viewed in color.

TABLE 1. Ablation study of MPM-LTL on Market-1501. The best results are bolded. MP: Mask Pooling. GTL: Global-level Triplet Loss. LTL: Local-level Triplet Loss. SQ: Single-Query. MQ: Multi-Query. RR: Re-ranking.

Datasets	Market-1501							
Models	SQ		MQ		SQ+RR		MQ+RR	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
(1) Baseline	86.55	69.07	91.48	77.57	88.39	82.50	92.93	88.57
(2) Baseline + Mask (convolution)	88.45	69.87	91.86	77.70	90.29	84.56	92.84	88.29
(3) Baseline + Mask (bilinear interpolation)	88.66	70.86	93.20	78.57	90.56	84.88	93.29	88.84
(4) Baseline + Mask + MP	89.82	73.67	93.29	80.31	92.10	86.30	94.21	89.90
(5) Baseline + Mask + MP + GTL	90.59	75.35	94.03	81.96	92.73	87.60	94.83	91.16
(6) Baseline + Mask + MP + LTL	92.87	79.53	94.83	85.27	94.30	90.94	**95.72**	93.31
(7) Baseline + Mask + MP + LTL + TriHard (Ours)	**93.38**	**80.45**	**95.40**	**86.47**	**94.39**	**91.39**	**95.72**	**93.76**

- After the mask pooling and convolution operations in Mask-pooling model, the size of Tensor T is $24 \times 8 \times 2048$. With part number fixed at 6, after an average pooling, the size of Tensor G is $6 \times 1 \times 2048$. Then, a 1×1 convolution turns the size of Tensor H to $6 \times 1 \times 256$.
- When using triplet loss, the distance margin between different identities α is 0.5.

3) EXPERIMENT ENVIRONMENT

All experiments are conducted on 4 GPUs with 16G memory each.

C. ABLATION STUDY AND PERFORMANCE EVALUATION

We investigate the effectiveness of each component in our proposed model by conducting a series of experiments on Market-1501, CUHK03 and DukeMTMC-reID.

(1) Baseline. In the baseline model, we employ the original ResNet50 as the backbone network, which takes the original RGB images in the datasets as inputs, without any mask information.

(2) Baseline + Mask (convolution). To assess the benefit of mask, the baseline model described above takes masks as additional inputs, besides the original RGB images. All the four input channels are down sampled using convolution.

(3) Baseline + Mask (bilinear interpolation). Different from (2), the mask channel is down sampled using bilinear interpolation.

TABLE 2. Ablation study of MPM-LTL on DukeMTMC-reID. The best results are bolded. MP: Mask Pooling. GTL: Global-level Triplet Loss. LTL: Local-level Triplet Loss. SQ: Single-Query. MQ: Multi-Query. RR: Re-ranking.

Datasets	DukeMTMC-reID			
Models	SQ		SQ+RR	
	Rank1	mAP	Rank1	mAP
(1) Baseline	77.78	58.92	81.33	74.70
(2) Baseline + Mask (convolution)	78.82	60.66	83.89	76.87
(3) Baseline + Mask (bilinear interpolation)	78.90	60.20	83.03	75.45
(4) Baseline + Mask + MP	81.46	64.52	86.09	80.89
(5) Baseline + Mask + MP + GTL	82.09	64.20	87.07	81.73
(6) Baseline + Mask + MP + LTL	83.03	66.39	86.89	82.45
(7) Baseline + Mask + MP + LTL + TriHard (Ours)	**83.44**	**67.16**	**87.12**	**82.69**

(4) Baseline + Mask + Mask Pooling. To evaluate the effectiveness of mask pooling, mask-pooling model is used, as described in Section III-C and Fig.4.

(5) Baseline + Mask + Mask Pooling + Global-level Triplet Loss. The four models above are models without triplet architecture. When training with triplet loss, a triplet architecture is adopted. Global-level means that the whole feature maps are directly used for training, without being divided.

(6) Baseline + Mask + Mask Pooling + Local-level Triplet Loss. Feature maps are divided into several parts for training and triplet loss is calculated with the corresponding parts.

(7) Baseline + Mask + Mask Pooling + Local-level Triplet Loss + TriHard (Ours). Hard triplets are selected

TABLE 3. Ablation study of MPM-LTL on CUHK03. The best results are bolded. MP: Mask Pooling. GTL: Global-level Triplet Loss. LTL: Local-level Triplet Loss. SQ: Single-Query. MQ: Multi-Query. RR: Re-ranking.

Datasets	CUHK03 (Original protocol)							
Models	Labeled				Detected			
	SQ		SQ+RR		SQ		SQ+RR	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
(1) Baseline	84.00	80.01	90.00	90.73	79.00	76.37	89.50	88.50
(2) Baseline + Mask (convolution)	88.44	85.89	92.46	93.08	86.43	82.85	89.95	90.80
(3) Baseline + Mask (bilinear interpolation)	87.44	84.68	92.46	93.04	83.92	80.64	88.44	88.33
(4) Baseline + Mask + MP	89.95	88.50	94.47	94.78	88.50	85.59	92.00	92.62
(5) Baseline + Mask + MP + GTL	90.50	87.25	94.50	94.62	89.45	86.30	95.48	94.37
(6) Baseline + Mask + MP + LTL	93.00	91.14	95.25	95.00	93.00	89.24	96.00	95.01
(7) Baseline + Mask + MP + LTL + TriHard (Ours)	**95.00**	**92.52**	**97.50**	**97.36**	**93.50**	**89.51**	**96.98**	**95.67**
Datasets	CUHK03 (New protocol)							
Models	Labeled				Detected			
	SQ		SQ+RR		SQ		SQ+RR	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
(1) Baseline	38.43	35.71	48.14	49.91	39.43	35.39	47.43	49.56
(2) Baseline + Mask (convolution)	50.00	45.48	62.00	62.40	47.21	41.03	56.29	55.82
(3) Baseline + Mask (bilinear interpolation)	49.29	44.76	61.21	62.08	48.36	42.24	58.29	57.56
(4) Baseline + Mask + MP	55.29	50.13	67.50	67.87	55.00	49.64	66.43	67.83
(5) Baseline + Mask + MP + GTL	57.36	50.78	67.21	67.45	58.29	54.01	71.50	71.51
(6) Baseline + Mask + MP + LTL	64.36	59.04	75.07	75.83	61.79	55.99	71.64	**72.40**
(7) Baseline + Mask + MP + LTL + TriHard (Ours)	**66.29**	**59.89**	**75.64**	**76.30**	**62.71**	**56.61**	**71.79**	72.32

TABLE 4. Comparison with other methods on Market-1501. The best results are bolded.

Methods	SQ		MQ	
	Rank1	mAP	Rank1	mAP
MSCF [50] (IEEE Access 2019)	82.90	-	-	-
CMFE [29] (j.neucom 2019)	84.70	65.80	-	-
TriNet [25] (arXiv 2017)	84.92	69.14	90.53	76.42
JLML [17] (IJCAI 2017)	85.10	65.50	89.70	74.50
JLDE [9] (j.neucom 2018)	85.21	67.69	90.73	76.17
PESR [19] (IEEE Access 2020)	85.60	79.20	-	-
AACN [48] (CVPR 2018)	85.90	66.87	89.78	75.10
AOS [49] (CVPR 2018)	86.49	70.43	91.32	78.33
IC-TL [11] (j.neucom 2018)	86.60	70.10	-	-
PSE [35] (CVPR 2018)	87.70	69.00	-	-
STN [8] (IEEE Access 2019)	87.82	71.93	-	-
NWAPI [41] (IEEE Access 2019)	89.78	71.69	-	-
GLAD [13] (ACM MM 2017)	89.90	73.90	-	-
MLFN [51] (CVPR 2018)	90.00	74.30	92.30	82.40
JA-ReID [31] (IEEE Access 2019)	90.40	76.10	-	-
PSE+ECN [35] (CVPR 2018)	90.40	80.50	-	-
FD-GAN [52] (NIPS 2018)	90.50	77.70	-	-
HA-CNN [53] (CVPR 2018)	91.20	75.70	93.80	82.80
DuATM [24] (CVPR 2018)	91.42	76.62	-	-
PABR [33] (ECCV 2018)	91.70	79.60	94.00	85.20
PCB [18] (ECCV 2018)	92.30	77.40	-	-
PAAN [10] (IEEE Access 2019)	92.40	77.60	**95.53**	84.26
KPM+RSA+HG [54] (CVPR 2018)	92.70	**82.50**	-	-
GCSL [28] (CVPR 2018)	93.50	81.60	-	-
PCB+RPP [18] (ECCV 2018)	**93.80**	81.60	-	-
MPM-LTL (OURS)	93.38	80.45	95.40	**86.47**
Re-ranking [46] (CVPR 2017)	77.11	63.63	-	-
ECN(RR) [35] (CVPR 2018)	82.30	71.10	-	-
MGCAM(RR) [42] (CVPR 2018)	83.79	74.33	-	-
MSCF(RR) [50] (IEEE Access 2019)	85.70	-	-	-
CMFE(RR) [29] (j.neucom 2019)	85.90	67.20	-	-
TriNet(RR) [25] (arXiv 2017)	86.67	81.07	91.75	87.18
cTransNet(RR) [6] (IEEE Access 2020)	88.10	71.20	-	-
PAN(RR) [36] (arViv 2017)	88.57	81.53	-	-
AOS(RR) [49] (CVPR 2018)	88.66	83.30	92.51	88.60
AACN(RR) [48] (CVPR 2018)	88.69	82.96	92.16	87.32
PSE+ECN(RR) [35] (CVPR 2018)	90.30	84.00	-	-
PABR(RR) [33] (ECCV 2018)	93.40	89.90	95.40	93.10
MPM-LTL (OURS, RR)	**94.39**	**91.39**	**95.72**	**93.76**

TABLE 5. Comparison with other methods on CUHK03. The best results are bolded. The first two lines use the original training/testing protocol [2].The last two lines use the new protocol [46].

Methods	Labeled		Detected	
	Rank1	mAP	Rank1	mAP
MSCAN [16] (CVPR 2017)	74.21	-	67.99	-
Quadruplet [26] (CVPR 2017)	-	-	75.53	-
JLML [17] (IJCAI 2017)	83.20	-	80.60	-
PartNet [32] (ICCV 2017)	85.40	-	81.60	-
GLAD [13] (ACM MM 2017)	85.00	-	82.20	-
JLDE [9] (j.neucom 2018)	86.60	-	85.30	-
PDC [14] (ICCV 2017)	88.70	-	78.29	-
IC-TL [11] (j.neucom 2018)	-	-	86.80	-
TriNet [25] (arXiv 2017)	89.63	-	87.58	-
CMFE [29] (j.neucom 2019)	-	-	88.30	-
GCSL [28] (CVPR 2018)	90.20	-	88.80	-
MSCF [50] (IEEE Access 2019)	91.20	-	-	-
AACN [48] (CVPR 2018)	91.39	-	89.51	-
PABR [33] (ECCV 2018)	91.50	-	88.00	-
MPM-LTL (OURS)	**95.00**	**92.52**	**93.50**	**89.51**
Re-ranking [46] (CVPR 2017)	69.90	70.89	69.67	72.45
CMFE(RR) [29] (j.neucom 2019)	-	-	89.20	-
MSCF(RR) [50] (IEEE Access 2019)	93.50	-	-	-
KPM+RSA+HG [54] (CVPR 2018)	94.90	94.00	-	-
MPM-LTL (OURS, RR)	**97.50**	**97.36**	**96.98**	**95.67**
PAN [36] (arXiv 2017)	36.90	35.00	36.30	34.00
HA-CNN [53] (CVPR 2018)	44.40	41.00	41.70	38.60
AOS [49] (CVPR 2018)	-	-	47.14	43.33
CMFE [29] (j.neucom 2019)	-	-	48.20	-
MLFN [51] (CVPR 2018)	54.70	49.20	52.80	47.80
JA-ReID [31] (IEEE Access 2019)	-	-	58.00	56.50
STN [8] (IEEE Access 2019)	61.20	54.80	60.20	54.60
PCB [18] (ECCV 2018)	-	-	61.30	54.20
PCB+RPP [18] (ECCV 2018)	-	-	**63.70**	**57.50**
MPM-LTL (OURS)	**66.29**	**59.89**	62.71	56.61
ECN(RR) [35] (CVPR 2018)	-	-	30.20	27.30
PAN(RR) [36] (arXiv 2017)	43.90	45.80	41.90	43.80
MGCAM(RR) [42] (CVPR 2018)	50.14	50.21	46.71	46.87
CMFE(RR) [29] (j.neucom 2019)	-	-	49.30	-
AOS(RR) [49] (CVPR 2018)	-	-	54.56	56.09
MPM-LTL (OURS, RR)	**75.64**	**76.30**	**71.79**	**72.32**

for local-level triplet loss using the proposed Mask-guided TriHard algorithm. This is our proposed model.

As shown in Table 1, Table 2 and Table 3, inputs with mask information are better than the original datasets, which shows the effectiveness of the extracted mask. Models with mask pooling achieve much better results than models without it or models that use bilinear interpolation instead of mask pooling, which reflects the benefits of the mask pooling. Models with triplet architecture have better performance than models with single branch. Local-level triplet loss is much

better than global-level triplet loss. And finally, Mask-guided TriHard algorithm for hard triplets selection further improves the performance of the model. In general, the Mask Pooling and Local-level Triplet Loss significantly improve the performance of the model, while the TriHard only slightly improves the model, which is probably because the method fails to select hard samples according to the specific postures. Therefore, better hard triplets selection algorithm will be considered in future work, which will make more use of posture and pedestrian behaviors.

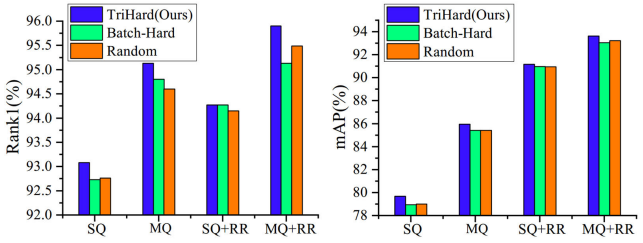


FIGURE 9. The comparison of the proposed Mask-guided TriHard algorithm, Batch-Hard [25] and random triplet selection method on Market-1501. Best viewed in color.

In addition to the random selection method and our TriHard method in the above ablation study (Table 1, 2, 3 (6) and (7)), we use the same backbone to test out another hard sample selection method, i.e., Batch Hard, and the comparison is shown in Fig. 9. As can be seen in Fig. 9, the hard triplet selection algorithm we design is more effective than Batch-Hard [25] and the random triplet selection method because it takes into account the feature distance between sample pairs and the intersection of their mask.

D. COMPARISON WITH THE STATE-OF-THE-ART METHODS

The above experiments have shown that each component of the proposed model MPM-LTL is effective. To verify the overall effect of our method, we compare it with the state-of-the-art methods.

Comparisons on Market-1501, CUHK03 and DukeMTMC-reID are exhibited in Table 4, Table 5 and Table 6 respectively. In this paper, we report Rank-1 = 95.72% and mAP = 93.76% for Market-1501, Rank-1 = 97.50% and mAP = 97.36% for labeled CUHK03, Rank-1 = 96.98% and mAP = 95.67% for detected CUHK03 using original protocol, Rank-1 = 75.64% and mAP = 76.30% for labeled CUHK03, Rank-1 = 71.79% and mAP = 72.32% for detected CUHK03 using new protocol, and Rank-1 = 87.12% and mAP = 82.69% for DukeMTMC-reID, setting new state-of-the-art on the three datasets.

What needs to be emphasized is that MGCAM [42] and NWAPI [41] also utilize mask to improve its performance. But with novel mask pooling, Mask-guided triplet selection and local-level triplet loss, our MPM-LTL **surpasses** MGCAM by **10.60%** in Rank-1 and **17.06%** in mAP on Market-1501, by **25.50%** in Rank-1 and **26.09%** in mAP on CUHK03 labeled dataset (using new protocol), and by **25.08%** in Rank-1 and **25.45%** in mAP on CUHK03 detected dataset (using new protocol), and surpasses NWAPI by 3.60% in Rank-1 and 8.76% in mAP on Market-1501, and by 1.57% in Rank-1 and 1.85% in mAP on DukeMTMC-reID.

TABLE 6. Comparison with other methods on DukeMTMC-reID. The best results are bolded.

Methods	Rank1	mAP
PAN [36] (arXiv 2017)	71.59	51.51
STN [8] (IEEE Access 2019)	76.55	61.02
CMFE [29] (j.neucom 2019)	76.80	60.20
AACN [48] (CVPR 2018)	76.84	59.25
AOS [49] (CVPR 2018)	79.17	62.10
PESR [19] (IEEE Access 2020)	79.40	55.20
PSE [35] (CVPR 2018)	79.80	62.00
FD-GAN [52] (NIPS 2018)	80.00	64.50
HA-CNN [53] (CVPR 2018)	80.50	63.80
KPM+RSA+HG [54] (CVPR 2018)	80.70	66.40
JA-ReID [31] (IEEE Access 2019)	80.90	65.70
MLFN [51] (CVPR 2018)	81.00	62.80
PCB [18] (ECCV 2018)	81.70	66.10
DuATM [24] (CVPR 2018)	81.82	64.58
NWAPI [41] (IEEE Access 2019)	81.87	65.31
PAAN [10] (IEEE Access 2019)	82.59	65.53
PCB+RPP [18] (ECCV 2018)	83.30	**69.20**
MPM-LTL (OURS)	**83.44**	67.16
PAN(RR) [36] (arXiv 2017)	75.94	66.74
CMFE(RR) [29] (j.neucom 2019)	78.20	61.30
cTransNet(RR) [6] (IEEE Access 2020)	81.10	62.80
AOS(RR) [49] (CVPR 2018)	84.11	78.19
PSE+ECN(RR) [35] (CVPR 2018)	85.20	79.80
MPM-LTL (OURS, RR)	**87.12**	**82.69**

V. CONCLUSION

In this paper, we propose a novel Mask-pooling Model with Local-level Triplet Loss for Person ReID. By applying mask pooling, we gradually remove the background features through deep convolutional process and acquire the most crucial person features. By employing local-level triplet loss, the model fully utilize the local features and capture the discriminative details. By exploiting the Mask-guided TriHard, we make use of outline information to select hard positive image pairs and negative image pairs. We also provide the experiments to show that the proposed model gets the state-of-the-art results on Market-1501, CUHK03 and DukeMTMC-reID using our MPM-LTL.

REFERENCES

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124, doi: 10.1109/ICCV.2015.133.

[2] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159, doi: 10.1109/CVPR.2014.27.

[3] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV*, Amsterdam, The Netherlands, Oct. 2016, pp. 17–35, doi: 10.1007/978-3-319-48881-3_2.

[4] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782, doi: 10.1109/ICCV.2017.405.

[5] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: http://arxiv.org/abs/1610.02984

[6] R. Sun, W. Lu, Y. Zhao, J. Zhang, and C. Kai, "A novel method for person re-identification: Conditional translated network based on GANs," *IEEE Access*, vol. 8, pp. 3677–3686, 2020, doi: 10.1109/ACCESS.2019.2962301.

[7] Y. Rao, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition and person re-identification," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 701–718, Jun. 2019, doi: 10.1007/s11263-018-1135-x.

[8] D. Chen, P. Chen, X. Yu, M. Cao, and T. Jia, "Deeply-learned spatial alignment for person re-identification," *IEEE Access*, vol. 7, pp. 143684–143692, 2019, doi: 10.1109/ACCESS.2019.2945353.

[9] C. Yuan, J. Guo, P. Feng, Z. Zhao, C. Xu, T. Wang, G. Choe, and K. Duan, "A jointly learned deep embedding for person re-identification," *Neurocomputing*, vol. 330, pp. 127–137, Feb. 2019, doi: 10.1016/j.neucom.2018.11.010.

[10] Y. Zhang, X. Gu, J. Tang, K. Cheng, and S. Tan, "Part-based attribute-aware network for person re-identification," *IEEE Access*, vol. 7, pp. 53585–53595, 2019, doi: 10.1109/ACCESS.2019.2912844.

[11] D. Wu, S.-J. Zheng, W.-Z. Bao, X.-P. Zhang, C.-A. Yuan, and D.-S. Huang, "A novel deep model with multi-loss and efficient training for person re-identification," *Neurocomputing*, vol. 324, pp. 69–75, Jan. 2019, doi: 10.1016/j.neucom.2018.03.073.

[12] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 907–915, doi: 10.1109/CVPR.2017.103.

[13] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. MM*, Mountain View, CA, USA, 2017, pp. 420–428, doi: 10.1145/3123266.3123279.

[14] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 3980–3989, doi: 10.1109/ICCV.2017.427.

[15] 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 *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 1335–1344, doi: 10.1109/CVPR.2016.149.

[16] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. CVPR*, Honolulu, HI, USA, Sep. 2017, pp. 7398–7407, doi: 10.1109/CVPR.2017.782.

[17] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. IJCAI*, Melbourne, VIC, Australia, 2017, pp. 2194–2200, doi: 10.24963/ijcai.2017/305.

[18] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 501–518, doi: 10.1007/978-3-030-01225-0_30.

[19] Y. Ha, J. Tian, Q. Miao, Q. Yang, J. Guo, and R. Jiang, "Part-based enhanced super resolution network for low-resolution person re-identification," *IEEE Access*, vol. 8, pp. 57594–57605, 2020, doi: 10.1109/ACCESS.2020.2971612.

[20] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. ECCV*, Amsterdam, The Netherlands, Oct. 2016, pp. 135–153, doi: 10.1007/978-3-319-46478-7_9.

[21] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. ECCV*, Amsterdam, The Netherlands, Oct. 2016, pp. 791–808, doi: 10.1007/978-3-319-46484-8_48.

[22] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015, doi: 10.1016/j.patcog.2015.04.005.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[24] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 5363–5372, doi: 10.1109/CVPR.2018.00562.

[25] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[26] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 1320–1329, doi: 10.1109/CVPR.2017.145.

[27] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. ICPR*, Stockholm, Sweden, Aug. 2014, pp. 34–39, doi: 10.1109/ICPR.2014.16.

[28] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 8649–8658, doi: 10.1109/CVPR.2018.00902.

[29] W. Zhong, L. Jiang, T. Zhang, J. Ji, and H. Xiong, "Combining multilevel feature extraction and multi-loss learning for person re-identification," *Neurocomputing*, vol. 334, pp. 68–78, Mar. 2019, doi: 10.1016/j.neucom.2019.01.005.

[30] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019, doi: 10.1109/TIP.2019.2891888.

[31] S. Jiao, J. Wang, G. Hu, Z. Pan, L. Du, and J. Zhang, "Joint attention mechanism for person re-identification," *IEEE Access*, vol. 7, pp. 90497–90506, 2019, doi: 10.1109/ACCESS.2019.2927170.

[32] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 3239–3248, doi: 10.1109/ICCV.2017.349.

[33] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 418–437, doi: 10.1007/978-3-030-01264-9_25.

[34] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," 2017, *arXiv:1701.07732*. [Online]. Available: http://arxiv.org/abs/1701.07732

[35] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 420–429, doi: 10.1109/CVPR.2018.00051.

[36] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," 2017, *arXiv:1707.00408*. [Online]. Available: http://arxiv.org/abs/1707.00408

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[38] K. He, G. Gkioxari, P. DollÍáar, and R. B. Girshick, "Mask R-CNN," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[39] X. Li and C. C. Loy, "Video object segmentation with joint re-identification and attention-aware mask propagation," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 93–110, doi: 10.1007/978-3-030-01219-9_6.

[40] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a maskguided two-stream CNN model," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 764–781, doi: 10.1007/978-3-030-01234-2_45.

[41] X. Yang, Y. Tang, N. Wang, B. Song, and X. Gao, "An end-to-end noise-weakened person re-identification and tracking with adaptive partial information," *IEEE Access*, vol. 7, pp. 20984–20995, 2019, doi: 10.1109/ACCESS.2019.2899032.

[42] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 1179–1188, doi: 10.1109/CVPR.2018.00129.

[43] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," 2017, *arXiv:1710.00478*. [Online]. Available: http://arxiv.org/abs/1710.00478

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

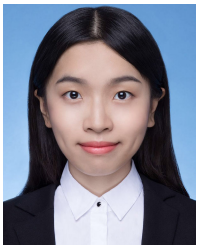
[45] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010, doi: 10.1109/TPAMI.2009.167.

[46] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 3652–3661, doi: 10.1109/CVPR.2017.389.

[47] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The relation between the ROC curve and the CMC," in *Proc. AutoID*, Buffalo, NY, USA, Oct. 2005, pp. 15–20, doi: 10.1109/AUTOID.2005.48.
[48] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 2119–2128, doi: 10.1109/CVPR.2018.00226.
[49] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 5098–5107, doi: 10.1109/CVPR.2018.00535.
[50] Z. Huang, Z. Yu, Y. Li, Y. Wang, S. Lin, D. Sun, Y. Zhong, H. Cao, and H. Gregersen, "Contribution-based multi-stream feature distance fusion method with *k*-distribution re-ranking for person re-identification," *IEEE Access*, vol. 7, pp. 35631–35644, 2019, doi: 10.1109/ACCESS.2019.2904278.
[51] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 2109–2118, doi: 10.1109/CVPR.2018.00225.
[52] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. NeurIPS*, Montrİäeal, QC, Canada, 2018, pp. 1230–1241.
[53] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person reidentification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 2285–2294, doi: 10.1109/CVPR.2018.00243.
[54] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 2265–2274, doi: 10.1109/CVPR.2018.00241.

FUDAN ZHENG received the B.S. and M.S. degrees from the School of Information Science and Technology, Sun Yat-sen University, China, in 2007 and 2009, respectively. She is currently pursuing the Ph.D. degree with the School of Data and Computer Science, Sun Yat-sen University. Her main research interests include computer vision and machine learning.

TINGTING CAI received the B.S. degree from the School of Software Engineering, Beijing Jiaotong University, China, in 2017. She is currently pursuing the master's degree with the School of Data and Computer Science, Sun Yat-sen University, China. Her main research interests include computer vision and machine learning.



YING WANG received the B.S. degree from the School of Mathematics, Sun Yat-sen University, China, in 2017. She is currently a Researcher with the National Supercomputer Center, Guangzhou, China. Her main research interests include deep learning and computer vision.

CHUFU DENG received the B.S. degree from the School of Computer, Guangdong University of Technology, China, in 2018. He is currently pursuing the master's degree with the School of Data and Computer Science, Sun Yat-sen University, China. His main research interests include deep learning and computer vision.

ZHIGUANG CHEN received the B.S. degree from the School of Computer Science and Technology, Harbin Institute of Technology, China, in 2007, and the M.S. and Ph.D. degrees from the School of Computer Science, National University of Defense Technology, China, in 2009 and 2013, respectively. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, China. His research interests include high-performance computing and large-scale storage systems.

HUILING ZHU received the B.S. degree from the Department of Mathematics, Tsinghua University, China, in 2006, and the Ph.D. degree from the Department of Mathematics, National University of Singapore, Singapore, in 2012. He is currently an Associate Professor with the College of Information Science and Technology, Jinan University, China. His research interests include machine learning and data mining.