MHSA-Net: Multihead Self-Attention Network for Occluded Person Re-Identification
Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li, Senior Member, IEEE

Abstract—This article presents a novel person reidentification model, named multihead self-attention network (MHSA-Net), to prune unimportant information and capture key local information from person images. MHSA-Net contains two main novel components: multihead self-attention branch (MHSAB) and attention competition mechanism (ACM). The MHSAB adaptively captures key local person information and then produces effective diversity embeddings of an image for the person matching. The ACM further helps filter out attention noise and nonkey information. Through extensive ablation studies, we verified that the MHSAB and ACM both contribute to the performance improvement of the MHSA-Net. Our MHSA-Net achieves competitive performance in the standard and occluded person Re-ID tasks.

Index Terms—Attention competition mechanism (ACM), feature fusion, multihead self-attention, occluded person re-identification (Re-ID).

I. INTRODUCTION

PERSON re-identification (Re-ID) is a fundamental task in distributed multicamera surveillance. It identifies the same person in different (nonoverlapping) camera views. Re-ID has important applications in video surveillance and criminal investigation. With the surge of interest in deep representation learning, the person Re-ID task has achieved great progress in recent years [61]. Although recently many methods [12], [52], [58], [71], [72], [77], [81] have boosted the performance of the standard person Re-ID task, they did not consider the situation that the person is occluded by various obstructions like cars, trees, or other people. The occlusion in person images is still a key challenging issue that hinders Re-ID performance. Thus, this article aims to develop a Re-ID algorithm that can better handle occlusions in images.

In the occluded person Re-ID task, occluded regions often contain a lot of noise that results in mismatching. So a key issue in occluded Re-ID is to build discriminative features from unoccluded regions. Some part-based methods [39], [41], [55] manually crop the occluded target person in probe images and then use the unoccluded parts as the new query. However, these manual operations are inefficient in practice. Another type of approach is to use human model to help build person features. More recently, [17], [18], [29] applied pose estimators to obtain the person’s key points to locate effective regions of the person. However, the difference between training datasets of pose estimation and that of person retrieval often exist, making pose estimation-based feature extraction sometimes unstable. It is desirable to design an effective mechanism to adaptively capture the key features from nonocclusion regions without relying on human models.

We are inspired by the recent multihead self-attention mechanism (MHSMAM) [3], [25], [34], [67], which flexibly captures spatially different local salience from the whole image, and generates multiple attention maps, from different aspects, for a single image. With MHSMAM, noisy/unimportant regions can be pruned and key local feature information can be highlighted. Therefore, we believe the idea of MHSMAM can help a Re-ID model to better locate key features from occluded images. As shown in Fig. 1, compared with the Baseline, two outstanding attention Re-ID model RGA-SC [66] and SCSN (three-stage) [6], the MHSAM can help the person Re-ID model better capture key information of the target person from the unoccluded regions and avoid information from occluded regions. The baseline may undesirably pay attention to clutter regions (left example) or other persons (right example), while our MHSAM model handles such occlusions much better.

However, developing effective MHSMAM for the task of Re-ID is nontrivial and needs careful design. We propose a novel MHSAM module for the person Re-ID task with a set of new strategies to help select the key subregions in the image.

Fig. 1. Occluded person images’ attention maps are produced by our baseline, RGA-SC [66], SCSN (three-stage) [6] and the person Re-ID model equipped with MHSMAM [3], [25]. The red dot is the target person.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
We call our new attention module a multihead self-attention branch (MHSAB).

Furthermore, attention noise from the occluded or nonkey regions often exists, and it affects the performance of the Re-ID model, we propose to design an attention competition mechanism (ACM) to further help MHSAB suppress or filter out such attention noise from nonkey subregions. Our main contributions are as follows.

1) We proposed a new attention module, MHSAB, that can more effectively extract person features in occluded person Re-ID.
2) We proposed a new attention competition module (ACM) to better prune attention noise from unimportant regions.
3) By integrating MHSAB and ACM modules, our final multithread self-attention network (MHSAS-Net) framework demonstrates better performance over most state-of-the-art methods when processing occluded images, i.e., on four occlusion datasets: Occluded-DukeMTMC [29], P-DukeMTMC-reID [30], Partial-ReID [55], and Partial-iLIDS [39]. On standard generic person Re-ID datasets, e.g., Market-1501 [38], DukeMTMC-reID [46], [74], and CUHK03 [53], our MHSAS-Net produces similar results with these state-of-the-art algorithms.

II. RELATED WORK

A. Attention Mechanism in Person Re-ID

Attention mechanisms have been widely exploited in computer vision and natural language processing (NLP), for instance in Text-to-Image Synthesis [22], Object Tracking [2], Image/Video Captioning [28], Visual Question Answering [82], Neural Machine Translation [15], and some Video Tasks [37], [68], [69] It can effectively capture task-relevant information and reduce interference from less important ones. Recently, many person Re-ID approaches [5], [8], [10], [32], [33], [51], [60], [62] also introduced various attention mechanisms into deep models to enhance identification performance.

Chi et al. [8], Chunfeng et al. [10], Kalayeh et al. [33], and Xuelin et al. [61] applied a human part detector or a human parsing model to capture features of body parts. Dong et al. [14] explored both the human part masks and human poses to enhance human body feature extraction. Jiaxu et al. [29] and Jing et al. [32] exploited the connectivity of the key points to generate human part masks and focuses on the human’s representation. However, the success of such approaches heavily relies on the accuracy of the human parsing models or pose estimators.

Other methods typically focus on extracting the person appearance or gait information, from the 3-D space or depth images, to reduce the interference of background or occlusion. For example, Zheng and Yang [76] tried to project the 2-D person image into the 3-D space, and conduct the person matching in the 3-D space. Munaro et al. [42] proposed point cloud matching (PCM) strategy to compute the distances of multiview point cloud sets, so as to distinguish between different persons. Haque et al. [20] adopted 3-D LSTM to build motion dynamics of 3-D person point clouds for person matching. Rao et al. [45] proposed a self-supervised gait encoding approach that can leverage unlabeled 3-D skeleton data to learn gait representations for person Re-ID. Sivapalan et al. [50] extended the gait energy image (GEI) [11] to 3-D domain and proposed gait energy volume (GEV) strategy based on depth images to perform gait-based person Re-ID. In [35], convolutional neural network-long short-term memory (CNN-LSTM) with reinforced temporal attention (RTA) was proposed for person matching based on a split-rate RGB-depth transfer method.

Besides, many methods [5], [9], [51], [54], [63] tried to exploit a different type of attention mechanism that does not need to use human models to capture human body features. Jianlou et al. [27] proposed a dual attention matching network based on an interclass and an intraclass attention module to capture context information of video sequences for person Re-ID. ABD-Net [51] combined spatial and channel attention to directly learn human’s information from the data and context. Yifan et al. [63] calculated the similarity of the local features to enhance local part information. Chiat-Pin et al. [9] applied an attribute classification to gain local attention information. However, it does not consider how to filter out information from the occlusion regions in the image. Therefore, with its fixed and parameter-free attention patterns, information from the occlusion region will be inevitably included.

Similar to the one mentioned in [9], [27], [51], [63], our attention module also does not rely on an external human model. But different from these methods, our attention mechanism can adaptively enhance/suppress attention weights of local features through a multiparameter learning strategy. The attention information of occluded and unoccluded regions in our attention mechanism is adaptively adjusted according to the targeting task. For the person Re-ID task, our attention module can flexibly capture the key local features and prune out information from occlusion regions.

B. Occluded Person Re-ID

Occlusion is a key challenging issue in person Re-ID. Recent studies [17], [18], [29], [30], [39]–[41], [55] on this topic can be divided into two categories: 1) partial person Re-ID methods [39], [41], [55] and 2) occluded person Re-ID methods [17], [18], [29], [30], [40].

The former category aims to match a partial probe image to a gallery holistic image. For example, Zheng et al. [55] adopted a global-to-local matching mechanism to capture the key information from the spatial channel of the feature maps. DSR [39], [41] proposed a spatial feature reconstruction strategy to align the partial person image with holistic images. However, these methods need a manual crop of the occluded target person in the probe image, before the cropped unoccluded part can be used to retrieve the target person.

The latter category aims to directly capture key features from the whole occluded person image to perform the person matching. AFPB [30] combined the occluded/unoccluded classification task and person ID classification task to improve the performance of deep model on capturing key information. FPR [40] reconstructed the feature map of unoccluded
regions in occluded person image and further improved it by a foreground-background mask to avoid the influence of background clutter. Guan’an et al. [18], Jiaxu et al. [29], and Shang et al. [49] proposed pose guided feature alignment methods to match the local patches of query and the gallery images based on human key-points. Our MHSA-Net also belongs to this type of method. However, different from these methods, the MHSA-Net does not require any additional model. Also, the MHSA-Net can more effectively capture unoccluded local information.

III. MHSA-NET OVERVIEW

Our MHSA-Net contains three modules: the global feature branch (GFB), the MHSAB, and the ACM, as illustrated in Fig. 2.

The GFB computes a basic large feature tensor $Q(x)$ and a global feature $q^*(x)$ for MHSAB and person matching cross entropy loss (CE loss). We use the widely adopted Backbone Network ResNet-50 [21] to compute feature tensor $Q(x)$, then down-sample it to $q^*(x)$ for MHSAB and person matching.

The MHSAB, the core component in MHSA-Net, captures the key local information and outputs the fusion feature $p^*(x)$ for the person matching. The MHSAB contains four submodules: 1) MHSAM; 2) feature regularization mechanism (FRM); 3) self-attention feature fusion module (SAFFM); and 4) residual learning module (RLM). MHSAB outputs attention weights $a(x)$, and fusion features $z(x)$ and $p^*(x)$ that capture key local information. These $a(x)$, $z(x)$ and $p^*(x)$, will be refined in the ACM.

The ACM is composed of a series of loss functions and a regularization item; and it updates attention weights $a(x)$, and fusion features $z(x)$ and $p^*(x)$ to enhance key person information and suppress nonkey person information.

In the Testing Stage: For the standard person Re-ID task, we concatenate the feature vector $p^*(x) \in \mathbb{R}^{512}$ and $q^*(x) \in \mathbb{R}^{512}$ to find the best matching person in the gallery by comparing the squared distance, i.e., $d(a, b) = \|a - b\|^2_2$. For the occlusion person task, we only use the feature vector $p^*(x) \in \mathbb{R}^{512}$ to find the best matching person in the gallery by comparing the squared distance.

IV. GFB (BASELINE)

Following recent state-of-the-art methods [5], [19], [47], [56], [63], [77], [83], we adopted ResNet-50 (pre-trained on ImageNet [26]) as the backbone network to encode a person image $x$. We modify the backbone ResNet-50 slightly to extract richer information via larger-sized high-level feature maps. The down-sampling operation at the beginning of stage 4 is not employed, then the output of the Backbone Network is $Q(x) \in \mathbb{R}^{24 \times 8 \times 2048}$. Following [19], [63], [83], we also append a series of downsampling operations to the large feature map $Q(x)$. As shown in Fig. 2, first we employ a global average pooling operation on the output feature $Q(x)$, the $24 \times 8 \times 2048$ tensor from the stage 4 of ResNet-50,
to get a feature vector \( q(x) \in \mathbb{R}^{2048} \). Then, \( q(x) \) is further reduced to a 512-dimensional feature vector \( q(x)^* \) through a \( 1 \times 1 \) convolution layer, a batch normalization layer, and a rectified linear unit (ReLU) layer. Finally, the feature vector \( q(x)^* \) is fed into the loss function, which is CE loss \( \mathcal{L}_{CE} \) in this baseline model. The baseline model of our MHSA-Net is composed of the GFB and the Backbone Network.

Unlike existing methods [5], [47], [56], [73], [77], we do not introduce the triplet loss into the GFB. In our experiments, we observed that incorporating triplet loss in the GFB negatively impacts the performance of MHSA-Net on generic person Re-ID with occlusions. It seems that this is more strict constraint on global features affects the local feature capturing in some degree. So, in our MHSA-Net, the loss function in the baseline model only contains \( \mathcal{L}_{CE} \).

V. MULTIHEAD SELF-ATTENTION BRANCH

We introduce the MHSA [3], [25], [67] into the person Re-ID pipeline, to help the network capture key local information from occluded images. However, there are two issues need to be solved for this MHSAM in the occluded person Re-ID task.

1) MHSAM [3], [25] can capture key local information using multiple embeddings; but these existing methods directly concatenate these embeddings, which result in a huge dimensional feature space, making search and training expensive and difficult. For our Re-ID task which is more complicated than the NLP task in [3] and 79], we need a more effective design on MHSAM to output a low dimensional and efficient person descriptor for the person matching task.

2) MHSAM produces multiple attention maps and feature embeddings for an image to encode rich information, which enhances the robustness of the deep model in representation learning [3]. But this design itself often makes different embeddings to redundantly encode similar or same personal information. Thus, it is desirable to make the generated embeddings diverse, namely, they capture various features of the person from different aspects.

Based on these observations, we propose a novel MHSAB to tackle the above issues. MHSAB contains three components: the MHSAM, FRM, and SAFFM. The MHSAM computes multiple attention maps for key subregions and multiple embeddings for each person image. The FRM contains a feature diversity regularization term (FDRT) and an improved hard triplet loss (IHTL) function. The FDRT enhances the diversity of the multiple embeddings in MHSAM, and the IHTL refines each individual embedding to better capture key information. The SAFFM adaptively combines multiple embeddings to produce a fused low-dimension feature vector.

A. Multihead Self-Attention Mechanism

As described in Section I, it is desirable to adaptively capture key local features in unoccluded regions and avoid information from occluded regions. To achieve this, we adopt a MHSAM [3], [25], [34]. The architecture of MHSAM is shown in Fig. 3. Here, we build \( K \)-head for this MHSAM in two steps.

1) First, given a person image, we learn its \( K \) attention weights \( \alpha(x) \in \mathbb{R}^{J \times K} \) (where \( J = [24 \times 8] \)) on each pixel \( j \in J \) of feature maps \( Q(x) \in \mathbb{R}^{J \times 2048} \).

2) Second, we compute the \( K \) attention-weighted embeddings of \( Q(x) \) for this person image. Specifically. 

**Step 1:** Compute \( \alpha(x) \) by the attention weight calculation module (Fig. 3)

\[
\alpha(x) = \text{softmax}(\omega_2 \text{ReLU}(\omega_1 Q(x)^T))
\]

where \( Q(x) \in \mathbb{R}^{J \times 2048} \) is reshaped to a matrix in \( \mathbb{R}^{192 \times 2048} \), \( \alpha(x) \in \mathbb{R}^{K \times 192} \) is reshaped to a tensor in \( \mathbb{R}^{24 \times 8 \times K} \), \( \omega_2 \in \mathbb{R}^{K \times 512} \) and \( \omega_1 \in \mathbb{R}^{512 \times 2048} \) are two parameter weight matrices to learn, and the softmax is applied pixel-wise so that on each pixel the \( K \) attention weights sum up to one.

**Step 2:** Multiply the attention weight \( \alpha(x) \) with feature maps \( Q(x) \), and further apply a nonlinear transformation, to get \( K \) attention-weighted embeddings \( P \in \mathbb{R}^{K \times 512} \) (Fig. 3)

\[
P(x) = \text{AvgPool}(\alpha(x) \odot Q(x)) \omega_1 + b_1
\]
where $\omega_3 \in \mathbb{R}^{2048 \times 512}$ is the parameter weight matrix, AvgPool($\cdot$) is the average pooling operation, and $b_3 \in \mathbb{R}^{512}$ is the bias to learn for the fully connection layer “FC” in Fig. 3. Since this, we can obtain $K$ feature branch heads, and the number of the heads is $K$. The $\odot$ is the element-wise product operation.

The attention weights $\alpha(x)$ in (1) are adaptively learned toward the objective of person matching in Re-ID. Greater $\alpha$ values indicate bigger importance of pixels/local regions and vice versa. As some examples shown in Figs. 1 and 7, key information from unoccluded regions can be captured by MHSAM, while occluded regions can be suppressed. Here, the hyperparameter $K$ is discussed in Section VII-E.1.

B. Feature Regularization Mechanism

FRM contains a FDRT and an IHTL. The FDRT encourages the multiple embeddings $P(x)$ to cover more key local information from various respects. The IHTL refines the MHSAM, while occluded regions can be suppressed. Here, key information from unoccluded regions can be captured by and vice versa. As some examples shown in Figs. 1 and 7, an embedding $x_p$ directly produced by MHSAM tend to capture sim-

perspectives, which enhances the model robustness. More diverse and can capture key information from different

1) Feature Diversity Regularization Term: The $K$ embeddings directly produced by MHSAM tend to capture similar/same person information redundantly. To avoid this, following [3], we also introduce the FDRT into MHSAM, to regularize the $K$ representations and enforce their diversity.

The $K$ embeddings in MHSAM are not overcomplete [13], [23]. So, we can restrict the Gram matrix of $K$ embeddings to be close to an identity matrix under Frobenius norm. First, we create a Gram matrix $G(x)$ of $P(x)$ by $G(x) = P(x)^T P(x)$. Each element in $G(x)$ denotes the correlation between $P(x)$. Here, $P(x)$ is normalized so that they are on an $L_2$ ball. Second, to enhance the diversity of the $K$ feature vectors in $P(x)$, we minimize the deviation of $G(x)$ from the identity matrix. Therefore, we define the FDRT as

$$\mathcal{L}_{FDRT} = \frac{1}{K^2} \|G(x) - I\|_1 \quad (3)$$

where $G(x)$ is the gram matrices of $P(x)$, and $I \in \mathbb{R}^{K \times K}$ is an identity matrix. With FDRT, the $K$ embeddings $P(x)$ are more diverse and can capture key information from different perspectives, which enhances the model robustness.

2) Improved Hard Triplet Loss: MHSAM produces $K$ embeddings $P(x) \in \mathbb{R}^{K \times 512}$ for each person image. To further filter out non-key information, we design a new loss function to help train the network so that each individual embedding can be used separately for person matching. We are inspired by the hard triple loss [1], which uses a hard sample mining strategy to achieve desirable performance. Hence, we propose an IHTL by revising the hard triple loss [1].

Before defining the IHTL, we first organize the training samples into a set of triplet feature units, $S = (s(x^a), s(x^p), s(x^n))$, or simply $S = (s^a, s^p, s^n)$ in the following. The raw person image triplet units is $X = (x^a, x^p, x^n)$. Here, $(s^a, s^p)$ represents a positive pair of features $y^a = y^p$, and $(s^a, s^n)$ indicates a negative pair of features with $y^a \neq y^n$. Here, $y \in Y$ is the person ID.

In the hard triple loss [1], a hard-sample mining strategy is introduced: a positive sample pair with the largest distance is defined as the hard positive sample pair; the negative sample pair with the smallest distance is defined as the hard negative sample pair. The hard triple loss function can then be defined using hard sample pairs

$$\mathcal{T}_{HardTriplet} = \ln (1 + \exp (\max \{d(s^a, s^p) - \min \{d(s^a, s^n)\}\))). \quad (4)$$

Based on the hard triplet loss function, we define an IHTL. We define the improved hard positive sample pair and improved hard negative sample pair in two steps.

1) Between each sample image pair, $K \times K$ distances can be computed, because each person image has $K$ embeddings $P(x) \in \mathbb{R}^{K \times 512}$ in MHSAM. We use the largest distance from these distances to measure the embeddings of the positive sample pairs, and use the smallest distance from these distances for the negative sample pairs.

2) We further use the hard samples mining strategy [1] to define the hard sample pairs. The improved hard positive sample pair is $\max \{d(P(x^a), P(x^p))\}$; the improved hard negative sample pair is $\min \{d(P(x^a), P(x^n))\}$. The IHTL is defined as

$$\mathcal{T}_{IHTL} = \ln (1 + \exp (\max \{d(P(x^a), P(x^p))\} - \min \{d(P(x^a), P(x^n))\}))) \quad (5)$$

where $i, j \in \{1, 2, \ldots, K\}, d(a, b) = \|a - b\|^2$ denotes the squared distance in feature space. Here, During training, the IHTL refines embeddings so that they individually can perform better person matching. This encourages the embeddings to focus on important information.

C. Self-Attention Feature Fusion Module

The output of FDRT, the $K$ embeddings $P^s(x) \in \mathbb{R}^{K \times 512}$ covers various properties of a person image. But directly using $P^s(x)$ by concatenation will lead to dimension explosion in person matching. Thus, we design a SAFFM to first learn $K$ attentional weights by a series of neural networks, then fuse the embeddings of the positive sample pairs, and use the smallest distance from these distances for the negative sample pairs.

Step-1: compute the attentional weight $\beta(x) \in \mathbb{R}^{512 \times 512}$ (Figs. 2 and 4). The matrix $P^s(x) \in \mathbb{R}^{512 \times 512}$ is transposed to $P^s(x) \in \mathbb{R}^{512 \times K}$, then compute $\beta$ by

$$\beta(x) = \text{softmax}(\omega_9 \text{ReLU}(\omega_8 P^s(x))) \quad (6)$$

where $\omega_9 \in \mathbb{R}^{512 \times 1024}$ and $\omega_8 \in \mathbb{R}^{1024 \times 512}$ are two parameter weight matrices to learn, and the softmax is applied pixel-wise so that each pixel on the each attention vector of the $\beta(x)$ sum up to one.

Step-2: compute the self-attention weighted feature vector $p(x)^* \in \mathbb{R}^{512}$, by

$$p(x)^* = \sum_{i=1}^{K}[\beta(x) \odot P^s]_i. \quad (7)$$
Here, \( p(x)^* \in \mathbb{R}^{512} \), \( \odot \) is the element-wise product operation.

SAFFM reduces the dimension of the multiple embeddings \( P^i(x) \) for both training and testing. In the training stage, \( p(x)^* \) is also fed to the CE loss and the hard triplet loss function

\[
\mathcal{L}_{\text{SAFFM}} = \mathcal{L}^*_\text{CE} + \mathcal{L}^*_\text{HardTriplet}.
\]

Here, the input to both \( \mathcal{L}^*_\text{CE} \) and \( \mathcal{L}^*_\text{HardTriplet} \) is \( p(x)^* \).

**D. Residual Learning Module**

As shown in the RLM module in Fig. 2, with MHSAM, \( P^i(x) \) aims to capture key information in unoccluded regions from local perspective; while \( q(x)^* \) captures global information of the whole person image. To prevent \( P^i(x) \) from being redundant with \( q(x)^* \), we cast their feature fusion as a residual learning task. Specifically,

1. To match with the dimension \( K \times 512 \) of \( P^i(x) \), we copy \( q(x)^* \) for \( K \) times to obtain \( Q(x)^* \in \mathbb{R}^{K \times 512} \).
2. The input to the residual block includes global feature \( Q(x)^* \) and local feature \( P^i(x) \). The parameters \((\omega_1, \omega_2, \omega_3, b_3)\) of \( P^i(x) \) will be optimized.
3. We define the residual learning embedding as

\[
Z(x) = \text{Norm}(Q(x)^* + P^i(x))
\]

where \( \text{Norm}() \) denotes the layer normalization [31]. This RLM encourages \( P^i(x) \) to only capture important local information.

In the training stage, \( Z(x) \in \mathbb{R}^{K \times 512} \) is simply summed along the first dimension to obtain \( z(x) \in \mathbb{R}^{512} \). And \( z(x) \) is also fed into the CE loss and the hard triplet loss function, that is

\[
\mathcal{L}_{\text{ReN}} = \mathcal{L}^{**}_\text{CE} + \mathcal{L}^{**}_\text{HardTriplet}.
\]

Here, the input of \( \mathcal{L}^{**}_\text{CE} \) and \( \mathcal{L}^{**}_\text{HardTriplet} \) is \( z(x) \). And \( z(x) \) does not participate in person matching in the testing stage.

Finally, the loss functions in MHSAB are summarized as

\[
\mathcal{L}_{\text{MHSAB}} = \mathcal{L}_{\text{SAFFM}} + \lambda_1 \mathcal{L}_{\text{FDRT}} + \mathcal{L}_{\text{ReN}} + \lambda_2 \mathcal{T}_{\text{HITL}}
\]

where \( \lambda_1 \) and \( \lambda_2 \) are the balance parameters (see detail in Sections VII-E2 and VII-E3).

**VI. ATTENTION COMPETITION MECHANISM**

MHSAB enhances attention on key subregions, but the extracted attention maps still contain some nonkey information. We propose an ACM to further refine the attention weights.

In [22], an attention competition strategy was proposed to filter out attention information of the nonkey words in the Text-to-Image generation task. This idea was composed of an attention regularization term and a series of cross-modal matching loss functions. This has been shown effective in the Text-to-Image generation task. In the image generation: an attention regularization term can effectively filter out the attention information of nonkey words; the cross-modal matching loss functions can effectively enhance or preserve the attention information of the keywords according to the objective. Similarly, we believe it can also help the person Re-ID model filter out the attention information of the nonkey subregions from the person images. Therefore, we also design a similar strategy in this Person Re-ID pipeline. To our knowledge, this is the first time a competition strategy was designed for Re-ID task. Through a series of experiments, we observe that this mechanism is promising.

Specifically, we use an attention regularization term to suppress nonkey information, and use the aforementioned person Re-ID loss function \( \mathcal{L}_{\text{MHSAB}} \) to enhance attention on important regions. The attention regularization term [22] is defined as

\[
\mathcal{L}_C = \sum_{i,j} (\min(a_{i,j}, \gamma))^2
\]

where the subscript “C” stands for “Competition,” and \( \gamma > 0 \) is a threshold. Fig. 5 shows a schematic of the ACM. The gray columns illustrate attention weights on nonkey subregions, and the green columns are for weights on key regions. In the initial state of training, as shown in Fig. 5(a), all attention weights in \( \alpha \) are small. In ACM, the attention regularization term \( \mathcal{L}_C \) sets a threshold and pushes the attention weights lower than this threshold toward zero; while \( \mathcal{L}_{\text{MHSAB}} \) increases attention weights of subregions if they benefit person matching. An illustration of this procedure is given in Fig. 5(b).
The Total Loss Functions in MHSA-Net: The total loss function $L_{\text{total}}$ is

$$L_{\text{total}} = L_{\text{MHSA}} + L_{\text{GFB}} + \lambda_3 L_{\text{C}}$$

(13)

where $\lambda_3$ is the balance parameter (see its discussion in Section VII-F).

VII. Experiment

To evaluate the MHSA-Net, we conduct extensive experiments on three widely used generic person Re-ID benchmarks, i.e., Market-1501 [38], DukeMTMC-reID [46], [74] and CUHK03 [53] datasets, and four occluded person Re-ID benchmarks, i.e., Occluded-DukeMTMC [29], P-DukeMTMC-reID [30], Partial-REID [55] and Partial-iLIDS [39]. First, we compare the performance of MHSA-Net with state-of-the-art methods on these datasets. Second, we perform ablation studies to validate the effectiveness of each component.

A. Datasets and Evaluation

We follow almost all person Re-ID approaches [4], [29], [30], [39], [46], [55], [63], [70], [74], [77], [81], [83] to set the following seven person Re-ID datasets.

Market-1501 [38] has 32,668 labeled images of 1501 identities collected from six camera views. The dataset is partitioned into two nonoverlapping parts: the training set with 12,936 images from 751 identities, and the test set with 19,732 images from 751 identities. In the testing stage, 3,368 query images from 750 identities are used to retrieve the persons from the rest of the test set, i.e., the gallery set.

DukeMTMC-reID [46], [74] is another large-scale person Re-ID dataset. It has 36,411 labeled images of 1404 identities collected from eight camera views. The training set consists of 16,522 images from 702 identities; We used 2228 query images from the other 702 identities, and 17,661 gallery images.

CUHK03 [53] is a challenging Re-ID benchmark. It has 14,096 images of 14,674 identities captured from six cameras. It contains two datasets. CUHK03-Labeled: the bounding boxes of person images are from manual labeling. CUHK03-Detected: the bounding boxes of person images are detected from deformable part models (DPMs), which is more challenging due to severe bounding box misalignment and background cluttering. Following [63], [77], [81], [83], we used the 767/700 split [53] of the detected images.

Occluded-DukeMTMC [29] has 15,618 training images, 17,661 gallery images, and 2210 occluded query images. We use this dataset to evaluate our MHSA-Net in Occluded Person Re-ID task.

P-DukeMTMC-reID [30] is a modified version based on DukeMTMC-reID [46], [74]. There are 2652 images (665 identities) in the training set, 2163 images (634 identities) in the query set and 9053 images in the gallery set.

Partial-REID [55] is a specially designed partial person Re-ID benchmark that has 600 images from 60 people. Each person has five partial images in query set and five full-body images in gallery set. These images are collected at a university campus under different viewpoints, backgrounds, and occlusions.

Partial-iLIDS [39] is a simulated partial person Re-ID dataset based on the iLIDS dataset. It has a total of 476 images of 119 people.

Evaluation Protocol: We employed two standard metrics adopted in most person Re-ID approaches, namely, the cumulative matching curve (CMC) that generates ranking accuracy, and the mean average precision (mAP). The CMC curve shows the probability that a query identity appears in different-sized candidate lists. This evaluation measurement is valid only if there is only one ground truth match for a given query. In this article, we report the Rank-1 accuracy. The mAP calculates the area under the Precision-Recall curve, which is known as average precision (AP). Then, the mean value of APs of all queries, i.e., mAP, is calculated, which considers both precision and recall of an algorithm, thus providing a more comprehensive evaluation.

B. Implementation Details

Following many recent approaches [9], [29], [63], [77], [83], the input images are resized to $384 \times 128$ and then augmented by random horizontal flip and normalization in the training stage. In the testing stage, the images are also resized to $384 \times 128$ and augmented only by normalization. Using the ImageNet pretrained ResNet-50 as the backbone, our network is end-to-end in the whole training stage. Our network is trained using 2 GTX 2080Ti GPUs with a batch size of 128. Each batch contains 32 identities, with four samples per identity. We use Adam optimizer [36] with 400 epochs. The base learning rate is initialized to $10^{-3}$ with a linear warm-up [44] in first 50 epochs, then decayed to $10^{-4}$ after 200 epochs, and further decayed to $10^{-5}$ after 300 epochs. The whole training procedure has 400 epochs and takes approximately 2 h. Our MHSA-Net achieves the satisfactory performance in the general person Re-ID and occluded person Re-ID tasks, when $\lambda_1 = 1e - 4, \lambda_2 = 1.0, \lambda_3 = 1e - 3, \gamma = 1e - 3$ and $K = 8$.

C. Comparison With State-of-the-Art Methods

In this section, we compared MHSA-Net with a series of state-of-the-art approaches on seven person Re-ID datasets. Here, MHSA-Net concatenates the local feature $p(x)^*$ and global feature $q(x)^*$ to conduct the person matching task. Compared with the proposed MHSA-Net, MHSA-Net$^3$ indicates that we drop the $L_{CE}$ in the training process. The MHSA-Net$^3$ only uses the local feature $p(x)^*$ to conduct the person Re-ID task.

1) Person Re-ID on General Datasets: First, we compared MHSA-Net with the state-of-the-art generic person Re-ID approaches on Market-1501, DukeMTMC-Re-ID, CUHK03-Labeled, and CUHK03-Detected datasets, and reported the results in Table I. We randomly set $K = 5$ and $K = 8$ in these experiments (Through experiments, we observed that the $K$ value does not affect the result much. Some discussions on different $K$ values are given in Section VII-E1).
Our MHSA-Net gets Rank-1 = 94.6, 87.3, 73.4, 75.8 and mAP = 84.0, 73.1, 70.2, 73.0 for Market-1501, DukeMTMC-reID, CUHK03-Detected and CUHK03-Labeled, respectively. If we introduce the Reranking [79] into the MHSA-Net, i.e., MHSA-Net + Reranking (K = 8), the accuracy further increases to Rank-1 = 95.5, 90.7, 80.2, 82.6 and mAP = 93.0, 87.2, 80.9, 84.2 for Market-1501, DukeMTMC-reID, CUHK03-Detected and CUHK03-Labeled, respectively. Recently, state-of-the-art performance on Market-1501 and DukeMTMC-RE-ID has been saturated. Yet the MHSA-Net still gains effective improvement over the baseline model and outperforms most existing methods. The CUHK03 is the most challenging dataset among the three. Following the data setting in [64], [78], [82], and [84], MHSA-Net also outperforms the most state-of-the-art methods on both CUHK03-Labeled and CUHK03-Detected datasets.

2) Occluded Person Re-ID: A feature of MHSA-Net is that it handles Re-ID of occluded persons well. So, we also compared MHSA-Net with a series of occluded person Re-ID methods on the Occluded-DukeMTMC dataset, P-DukeMTMC-reID dataset, Partial-REID dataset, and Partial-iLIDS dataset.

3) Occluded-DukeMTMC and P-DukeMTMC-reID: MHSA-Net$^*$ only uses the local features $p(x)^*$ for the occluded Re-ID task. As shown in Table II, on Occluded-DukeMTMC, our MHSA-Net$^*$ achieves 59.7 Rank-1 accuracy and 44.8 mAP, which outperforms most previous methods. Compared with the baseline model, the MHSA-Net$^*$ gains 20.8 Rank-1 and 19.6 mAP improvement. As shown in Table III, on P-DukeMTMC-reID, our MHSA-Net$^*$ achieves 70.7 Rank-1 accuracy and 41.1 mAP, which outperforms all the previous methods. Compared with the baseline model, the MHSA-Net$^*$ gains 9.7 Rank-1 and 14.1 mAP improvement. The MHSA-Net combines both global feature $q(x)^*$ and local feature $p(x)^*$ to do the occluded person Re-ID task. The global feature $q(x)^*$ captures global information from the whole person image. Hence, it inevitably encodes contents of scene regions that occlude the person, and this leads to decreased performance. This can be remedied by reducing the constraints on the GFB. Specifically, if we drop the $L_{CE}$ in the GFB, the extraction of global feature becomes a simple downsampling from local features, making this global feature

---

**TABLE I**

Comparison with the Many State-of-the-Art Generic Person Re-ID Methods on Market-1501, DukeMTMC-reID, and CUHK03 Datasets. MHSA-Net$^†$ indicates that we drop the $L_{CE}$ in the training process. MHSA-Net$^∗$ indicates that we only use the key local feature $p(x)^*$ from MHSAb+ ACM to conduct the person matching task.

| Method     | Market-1501 | DukeMTMC-reID | CUHK03-Detected | CUHK03-Labeled |
|------------|-------------|---------------|----------------|----------------|
|            | Rank-1      | mAP           | Rank-1         | mAP            | Rank-1      | mAP           |
| MLEN [60]  | 90.0        | 74.3          | 81.0           | 62.8           | 52.8        | 47.8          | 54.7          | 49.2          |
| HA-CNN [55] | 91.2        | 75.7          | 80.5           | 63.8           | 41.7        | 38.6          | 44.4          | 41.0          |
| PCb+RPF [64]| 93.8        | 81.6          | 83.3           | 69.2           | 62.8        | 56.7          | -             | -             |
| MInc [7]   | 93.1        | 82.3          | 84.9           | 71.8           | 65.5        | 60.5          | -             | -             |
| PAN [76]   | 82.8        | 63.4          | 71.6           | 51.5           | 36.3        | 34.0          | 36.9          | 35.0          |
| PAN [48]   | 90.3        | 76.1          | -              | -              | 69.3        | 67.2          | -             | -             |
| VCF [16]   | 90.9        | 86.7          | -              | -              | 70.4        | 70.4          | -             | -             |
| PGFA [29]  | 91.2        | 76.8          | 82.6           | 65.5           | -           | -             | -             | -             |
| IIACNN+DHANET [53] | 91.3 | 76.0          | 81.3           | 64.1           | -           | -             | -             | -             |
| IANet [47] | 94.4        | 83.1          | 87.1           | 73.4           | -           | -             | -             | -             |
| BDB [84]  | 94.2        | 84.3          | 86.8           | 72.1           | 72.8        | 69.3          | 73.6          | 71.7          |
| AA-Net [9] | 93.9        | 83.4          | 87.7           | 74.3           | -           | -             | -             | -             |
| CAMA [57]  | 94.7        | 84.5          | 85.8           | 72.9           | 66.6        | 64.2          | -             | -             |
| OSNet [58] | 94.8        | 84.9          | 88.6           | 73.5           | 72.3        | 67.8          | -             | -             |
| RANGEv2 [58] | 94.7    | 86.8          | 87.0           | 78.2           | 64.6        | 61.6          | 67.4          | 64.3          |
| JWSA [45]  | 94.8        | 83.2          | 88.3           | 75.6           | 72.3        | 67.8          | -             | -             |
| HOReID [18] | 94.2        | 84.9          | 86.9           | 75.6           | -           | -             | -             | -             |
| SCNN (stage 4) [6] | 92.4 | 88.3          | 91.0           | 79.0           | 84.7        | 81.0          | 86.8          | 84.0          |
| SCNN (stage 3) [6] | 95.7 | 88.5          | 90.1           | 79.0           | 84.1        | 80.2          | 86.3          | 83.3          |
| RGA-SC [67] | 95.8        | 88.1          | 86.1           | 74.9           | 77.3        | 73.3          | 80.4          | 76.4          |
| Baseline   | 92.0        | 78.8          | 81.0           | 62.8           | 56.3        | 53.0          | 58.6          | 55.2          |
| MHSA-Net (K=5) | 94.3        | 83.5          | 87.1           | 73.0           | 73.4        | 70.2          | 75.8          | 73.0          |
| MHSA-Net (K=8) | 94.6        | 84.0          | 87.5           | 73.1           | 72.8        | 69.3          | 75.6          | 72.7          |
| MHSA-Net* (K=8) | 94.0        | 82.9          | 86.3           | 72.5           | 72.4        | 69.7          | 74.4          | 72.0          |
| MHSA-Net* (K=8) | 94.3        | 82.5          | 87.0           | 72.6           | 72.7        | 69.9          | 75.2          | 72.3          |
| MHSA-Net* Re-ranking [80] (K=8) | 95.5        | 93.0          | 90.7           | 87.2           | 80.2        | 80.9          | 82.6          | 84.2          |
TABLE II
COMPARISON WITH THE OTHER OCCLUDED PERSON RE-ID METHODS ON OCCLUDED-DukeMTMC DATASET. MHSA-Net indicates that we drop the LCE in the training process. MHSA-Net* indicates that we only use the key local feature p*(x) from MHSAB + ACM to conduct the person matching task. The first, second, and third highest scores are shown in red, green, and blue, respectively.

| Method                  | Occluded-DukeMTMC |
|-------------------------|-------------------|
|                         | Rank-1 | Rank-5 | Rank-10 | mAP  |
| Random Erasing [81]     | 40.5   | 59.6   | 66.8    | 30.0 |
| HA-CNN [55]             | 34.4   | 51.9   | 59.4    | 26.0 |
| Adver Occluded [24]     | 44.5   | -      | -       | 32.2 |
| PCB [64]                | 42.6   | 57.1   | 62.9    | 33.7 |
| Part Bilinear [66]      | 36.9   | -      | -       | -    |
| FD-GAN [65]             | 40.8   | -      | -       | -    |
| DSR [39]                | 40.8   | 58.2   | 65.2    | 30.4 |
| SFR [41]                | 42.3   | 60.3   | 67.3    | 32.0 |
| PPR [40]                | 51.4   | 68.6   | 74.9    | 37.3 |
| HOREd [18]              | 55.1   | -      | -       | -    |
| PVPM-Aug [17]           | 57.3   | 72.6   | 77.2    | 45.7 |
| Baseline                | 38.9   | 53.5   | 60.1    | 25.6 |
| MHSA-Net* (K=8)         | 59.7   | 74.3   | 79.5    | 44.8 |
| MHSA-Net (K=8)          | 55.4   | 70.2   | 76.4    | 42.4 |
| MHSA-Net† (K=8)         | 58.2   | 73.2   | 78.4    | 43.1 |

TABLE III
COMPARISON WITH THE OTHER OCCLUDED PERSON RE-ID METHODS ON P-DukeMTMC-reID DATASET. MHSA-Net† indicates that we drop the LCE in the training process. MHSA-Net* indicates that we only use the key local feature p*(x) from MHSAB + ACM to conduct the person matching task. The first, second, and third highest scores are shown in red, green, and blue, respectively.

| Method                  | P-DukeMTMC-reID |
|-------------------------|-----------------|
|                         | Rank-1 | Rank-5 | Rank-10 | mAP  |
| OSNet[78]               | 33.7   | 46.5   | 54.0    | 20.1 |
| PCB+IPPE [64]           | 40.4   | 54.6   | 61.1    | 23.4 |
| PCB [64]                | 43.6   | 57.1   | 63.3    | 24.7 |
| PGFA [29]               | 44.2   | 56.7   | 63.0    | 23.1 |
| PVPM-Aug [17]           | 51.5   | 64.4   | 69.6    | 29.2 |
| Baseline                | 61.0   | 72.5   | 78.4    | 27.0 |
| MHSA-Net* (K=8)         | 70.7   | 81.0   | 84.6    | 41.1 |
| MHSA-Net (K=8)          | 67.9   | 79.7   | 83.7    | 37.6 |
| MHSA-Net† (K=8)         | 69.6   | 81.4   | 85.0    | 37.5 |

MHSA-Net* and MHSA-Net† also achieve the best performance on both datasets. In both of these two data settings, compared with the baseline model, MHSA-Net* and MHSA-Net† gain a large improvement on both datasets. Like in Occluded-DukeMTMC and P-DukeMTMC-reID, MHSA-Net* and MHSA-Net† have better performance than MHSA-Net in Partial-REID and Partial-iLIDS datasets.

D. Ablation Study of MHSA-Net

We conducted ablation studies to show effectiveness of each component in the MHSA-Net. We show the ablation experiments results in Tables V and VI. By a series of discussions of hyperparameters in Section VII-E, we found the best hyperparameters setting for the proposed MHSA-Net: \( \lambda_1 = 1 - 4 < 1 \); \( \lambda_2 = 1 < 1 \); \( \gamma = 1 < 2 \); \( K = 8 \). In these experiments, we set \( K = 8 \) in MHSAM, as we observed it produces stable and effective person matching results.

Through ablation studies, we have the following observations.

1) Each individual component effectively improves the performance of the baseline model, as shown in Table V. Compared with the baseline model, the entire MHSA-Net (\( K = 8 \)) achieves: 2.6 Rank-1 and 7.2 mAP improvement on Market-1501; 6.3 Rank-1 and 10.3 mAP improvement on DukeMTMC-Re-ID; 16.5 Rank-1 and 16.3 mAP on CUHK03-Detected; 17.0 Rank-1 and 17.5 mAP on CUHK03-Labeled. When \( K = 5 \) or 6, the change in performance is minor.

2) We show the effectiveness of each embedding in \( P_{\perp}(x) \) ∈ \( \mathbb{R}^{K \times 512} \), i.e., \( P_{\perp}(x), i = 1, 2, \ldots, K \), \( K = 8 \). We only use the \( P_{\perp}(x) \) to search the target person. Here, we conduct the ablation studies on Occluded-DukeMTMC and CUHK03-Detected datasets in Table VI. As we can see Table VI, compared with the baseline model, each feature \( P_{\perp}(x) \) achieves large

q(x)* less sensitive to occlusions. We denote the pipeline using such a design as MHSA-Net†. In Tables II and III, the performance of MHSA-Net† is clearly better than MHSA-Net, and only slightly worse than MHSA-Net*. The MHSA-Net† also achieves a competitive performance in Table I.

In summary, for general database, we can use MHSA-Net. For datasets with certain occlusions, we can use MHSA-Net†, which best balances the global and local features. For datasets with severe occlusions, we can use MHSA-Net*, where local features play more important roles. As shown in Fig. 6, for the manually drawn occlusion, the MHSAM can better avoid the feature extraction of the occlusion part, and better extract the key person information of the nonocclusion part.

4) Partial-REID and Partial-iLIDS: The comparison of Re-ID on these two datasets is shown in Table IV. We also trained our model using the Market-1501 training set. Our
TABLE V
RESULTS PRODUCED BY COMBINING DIFFERENT COMPONENTS OF THE MHSA-NET. MHSA-NET† INDICATES THAT WE DROP THE LFCE IN THE TRAINING PROCESS. MHSA-NET∗ = MHSAM + IHTL + FDRT + ACM DENOTES THAT WE ONLY USE THE LOCAL FEATURE p∗(x) CONDUCTS PERSON RE-ID TASK. FROM “BASELINE” TO “MHSA-NET†” IN THIS TABLE, THE PARAMETER K IS SET TO 8 IN IMPLEMENTATION.

| Method                  | Market-1501 | Dukemtmc-reID | CUHK03-Detected | CUHK03-Labeled |
|-------------------------|-------------|---------------|-----------------|----------------|
|                         | Rank-1      | mAP           | Rank-1          | mAP            | Rank-1          | mAP            | Rank-1          | mAP            |
| Baseline                | 92.0        | 78.8          | 81.0            | 62.8           | 56.3            | 53.0           | 58.6            | 55.2           |
| Baseline=MHSAM (K=8)    | 93.0        | 80.2          | 83.2            | 70.2           | 68.1            | 64.4           | 72.1            | 69.4           |
| Baseline=MHSAM+ACM (K=8)| 94.4        | 81.9          | 86.5            | 73.1           | 69.9            | 65.8           | 72.9            | 70.1           |
| Baseline=MHSAM+FDRT (K=8)| 93.6       | 82.1          | 86.3            | 72.6           | 70.0            | 65.7           | 73.4            | 69.5           |
| Baseline=MHSAM+IHTL (K=8)| 93.5       | 82.2          | 86.4            | 72.6           | 71.2            | 67.6           | 73.6            | 70.5           |
| Baseline=MHSAM+IHTL+ACM (K=8)| 94.1     | 83.3          | 86.8            | 72.7           | 72.2            | 69.9           | 75.9            | 73.4           |
| Baseline=MHSAM+FDRT+ACM (K=8)| 94.1    | 83.6          | 86.5            | 73.1           | 71.1            | 69.4           | 72.9            | 68.9           |
| Baseline=MHSAM+IHTL+FDRT (K=8)| 93.9      | 83.2          | 86.9            | 73.2           | 72.2            | 69.3           | 75.0            | 72.3           |
| MHSAM-Net (K=8)        | 94.0        | 82.9          | 86.3            | 72.5           | 72.4            | 69.7           | 74.4            | 72.0           |
| MHSAM-Net (K=8)        | 94.3        | 83.5          | 87.1            | 73.0           | 73.4            | 70.2           | 75.8            | 73.0           |
| MHSAM-Net (K=6)        | 94.2        | 84.1          | 87.0            | 73.0           | 73.4            | 70.1           | 75.2            | 72.8           |
| MHSAM-Net (K=8)        | 94.6        | 84.0          | 87.3            | 73.1           | 72.8            | 69.3           | 75.6            | 72.7           |

TABLE VI
RESULTS ON OCCLUDED-DUKEMTM C AND CUHK03-DETECTED DATASET FOR EACH EMBEDDING IN THE P⊥(x). THE BOLD IS THE BEST RESULT.

| Method                  | Occluded-Dukemtc | CUHK03-Detected |
|-------------------------|------------------|-----------------|
|                         | Rank-1 | mAP | Rank-1 | mAP |
| Baseline                | 38.9   | 25.6 | 56.3   | 53.0 |
| P⊥ 1(x)                | 55.2   | 40.4 | 68.7   | 65.0 |
| P⊥ 2(x)                | 55.7   | 40.4 | 68.1   | 64.4 |
| P⊥ 3(x)                | 53.3   | 39.6 | 69.4   | 66.4 |
| P⊥ 4(x)                | 54.9   | 40.1 | 68.6   | 64.7 |
| P⊥ 5(x)                | 54.5   | 39.9 | 69.1   | 66.3 |
| P⊥ 6(x)                | 54.3   | 40.2 | 69.4   | 66.4 |
| P⊥ 7(x)                | 53.2   | 38.7 | 70.0   | 67.7 |
| P⊥ 8(x)                | 53.8   | 39.4 | 68.9   | 65.7 |

improvement over the baseline model in these two datasets. So, it indicates that each feature in P⊥(x) can better carry on both the generic and occluded person Re-ID tasks.

In the visualization results in Fig. 7, compared with Baseline, MHSAB can effectively capture more key subregions. Based on the MHSAB, we introduce the ACM into MHSAB, i.e., MHSAB + ACM. Compared with MHSAB, some attention information is suppressed and some attention information is highlighted. And subjectively, we can see that the highlight attention areas conducted by MHSAB + ACM are more important than that of MHSAB. Besides, the left half of Fig. 7, compared with baseline model, our MHSAB, and MHSAB + ACM can better capture the key information from the unoccluded regions in the occlusion person images.

Besides, as shown in Fig. 8, we visualized the feature maps of the 8 feature branches (from K₁ to K₈) of MHSAB under different variants. The feature maps corresponding to Kᵢ, (i = 1, 2, ..., K) represents the person information captured by the Kᵢth attention head in MHSAB.

1) As shown in “Group A: MHSAB” of Fig. 8, the FDRT can help the MHSAB effectively capture diversity information for person matching. Since this, MHSAM can capture key information from different perspectives, which enhances the model robustness.

2) If we remove the FDRT from MHSAB, i.e., “Group B: MHSAB w/o FDRT,” the feature maps (from K₁ to K₈) are mixed with some redundant information, and some feature responses (in red box especially) are scattered and weak.

3) If we remove the FDRT and IHTL from MHSAB, i.e., “Group B: MHSAB w/o FDRT and IHTL,” the responses of key information in the feature maps become sparse and weak.

Without the constraint of IHTL and FDRT, it is difficult to ensure that every feature branch head in MHSAB can capture the key diversity information for person matching.

In all, results in Table V can sufficiently evidence that the effectiveness of each component.

E. Hyperparameters Discussion in MHSA-Net

1) Multihead Self-Attention Mechanism: In this module, we 1) try to find suitable K for MHSAM and 2) discuss the influence of different mechanisms in MHSAM.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
1) Table VII shows the main results in the Market-1501 and CUHK03-Detected datasets. We set $\lambda_1 = 0$ and $\lambda_2 = 0$ in (11), and $\lambda_3 = 0$ in (13). Here, we set $K = 1, 2, \ldots, 9$ in MHSAM. As the results show, MHSAM with $K > 0$ can effectively improve the baseline’s performance in both datasets. And we find $K = 7, 8$ are suitable parameters when MHSAM produces good performance in both datasets over both measures.

2) Table VIII shows the influence of different mechanisms in MHSAM on the Occluded-DukeMTMC dataset. First, the effects of different feature fusion operations to $P^\perp(x)$ $\in \mathbb{R}^{K \times 512}$ in MHSAM are compared: MHSA-Net* uses SAFFM to fuse the $P^\perp(x)$ and produces a 512-dimensional vector; MHSA-Net* (CONCAT) directly concatenates the $P^\perp(x)$ to one $(K + 1) \times 512$ vector; MHSA-Net* (SUM) simply sums up the $P^\perp(x)$ to one 512-D vector. The SAFFM results in the best performance in these three feature fusion operations. The output vector from SAFFM also has much lower dimension than that from feature concatenation. Second, the effect of the “RLM” is compared. “MHSA-Net* w/o RLM” drops the “RLM” from MHSA-Net*, and this leads to declined performance. Hence, “RLM” is beneficial for MHSA-Net* to capture useful local information.

2) Feature Diversity Regularization Term: This section discusses the suitable $\lambda_1$ for $L_{FDRT}$, and FDRT’s performance under different values of hyperparameter $K$ in MHSAM. Table IX and Fig. 9 shows the results. We set $\lambda_2 = 0$ [see (11)] and $\lambda_3 = 0$ [see (13)].

Table IX shows Rank-1 and mAP results under different $\lambda_1$, from $10^{-6}$ to $10^{-1}$. Both Rank-1 and mAP reaches the highest score when $\lambda_1 = 10^{-4}$.

Then, with this setting of $\lambda_1 = 10^{-4}$ [see (6)], we discuss performance of FDRT on different hyperparameter $K$ values in MHSAM. We conducted ablation studies in Market-1501 and CUHK03-Detected datasets. Compared with MHSAM (in Orange polylines and bars Fig. 9), adding FDRT in MHSAM (in Blue polylines and bars, respectively) improves the performance of MHSAM in person Re-ID.
TAN et al.: MHSA-Net FOR OCCLUDED PERSON Re-ID

Fig. 9. Results of MHSAM + FDRT and MHSAM in the Market-1501 and CUHK03-Detected testing data from different hyperparameters $K$ in MHSAM. Here, the hyperparameter $\lambda_1 = 10^{-4}$. (a) Influence of the parameter $K$ on the MHSAM + FDRT under Rank-1 score in Market-1501 dataset. (b) Influence of the parameter $K$ on the MHSAM + FDRT under mAP score in Market-1501 dataset. (c) Influence of the parameter $K$ on the MHSAM + FDRT under Rank-1 score in CUHK03-Detected dataset. (d) Influence of the parameter $K$ on the MHSAM + FDRT under mAP score in CUHK03-Detected dataset.

Fig. 10. Results of the MHSAM + IHT and the MHSAM in the Market-1501 and CUHK03-Detected testing data from different hyperparameters $K$ in MHSAM. Here, the hyperparameter $\lambda_2 = 1.0$. (a) Influence of the parameter $K$ on the MHSAM + IHT under Rank-1 score in Market-1501 dataset. (b) Influence of the parameter $K$ on the MHSAM + IHT under mAP score in Market-1501 dataset. (c) Influence of the parameter $K$ on the MHSAM + IHT under Rank-1 score in CUHK03-Detected dataset. (d) Influence of the parameter $K$ on the MHSAM + IHT under mAP score in CUHK03-Detected dataset.

3) Improved Hard Triplet Loss: For this module, we find the suitable weight $\lambda_2$ in IHTL, and discuss the performance of different $K$ values in MHSAM. Table X and Fig. 10 shows the main results. Here we set $\lambda_1 = 0$ [see (11)] and $\lambda_3 = 0$ [see (13)].

Table X shows that with the setting of $\lambda_2$ from 0.01 to 100, Rank-1 and mAP scores get better and then decline. For both $K = 7$ and 8, IHTL gets the best performance when $\lambda_2 = 1.0$.

Then, with $\lambda_2 = 1.0$ set, we compare the performance of different $K$ in MHSAM. We conducted ablation studies in Market-1501 and CUHK03-Detected datasets. Compared with the MHSAM (Orange polylines and bars in Fig. 10), introducing IHTL into MHSAM, i.e., MHSAM + IHTL (Blue polylines and bars in Fig. 10), improves the performance of MHSAM in person Re-ID.

Table X

| Method                  | CUHK03-Detected | Rank-1 | mAP |
|-------------------------|-----------------|--------|-----|
| $K = 7$ Baseline+MHSAM  | 68.9            | 65.1   |
| $K = 7$, $\lambda_2 = 10^2$ | 61.2            | 59.5   |
| $K = 7$, $\lambda_2 = 10^1$ | 67.8            | 65.5   |
| $K = 7$, $\lambda_2 = 1.0$  | 70.7            | 67.5   |
| $K = 7$, $\lambda_2 = 10^{-1}$ | 69.8            | 65.2   |
| $K = 7$, $\lambda_2 = 10^{-2}$ | 68.6            | 64.8   |
| $K = 8$ Baseline+MHSAM  | 68.1            | 64.4   |
| $K = 8$, $\lambda_2 = 10^2$ | 63.4            | 61.1   |
| $K = 8$, $\lambda_2 = 10^1$ | 69.9            | 67.0   |
| $K = 8$, $\lambda_2 = 1.0$  | 71.2            | 67.6   |
| $K = 8$, $\lambda_2 = 10^{-1}$ | 70.9            | 67.5   |
| $K = 8$, $\lambda_2 = 10^{-2}$ | 68.7            | 65.1   |

Table XI

| Method                  | CUHK03-Detected | Rank-1 | mAP |
|-------------------------|-----------------|--------|-----|
| Baseline+MHSAM          | 68.1            | 64.4   |
| $\lambda_3 = 10^{-3}$, $\gamma = 10^{-3}$ | 68.7            | 65.1   |
| $\lambda_3 = 10^{-4}$, $\gamma = 10^{-3}$ | 69.1            | 65.0   |
| $\lambda_3 = 10^{-3}$, $\gamma = 10^{-3}$ | 69.9            | 65.8   |
| $\lambda_3 = 10^{-2}$, $\gamma = 10^{-3}$ | 69.0            | 65.3   |
| $\lambda_3 = 10^{-1}$, $\gamma = 10^{-3}$ | 69.4            | 65.1   |
| $\lambda_3 = 1$, $\gamma = 10^{-3}$ | 58.4            | 55.6   |
| $\lambda_3 = 10^{-3}$, $\gamma = 10^{-4}$ | 65.8            | 65.9   |
| $\lambda_3 = 10^{-3}$, $\gamma = 10^{-2}$ | 69.5            | 65.4   |
| $\lambda_3 = 10^{-3}$, $\gamma = 10^{-1}$ | 67.9            | 64.5   |
| $\lambda_3 = 10^{-3}$, $\gamma = 5 \times 10^{-1}$ | 67.4            | 64.2   |
| $\lambda_3 = 10^{-3}$, $\gamma = 1$ | 66.9            | 63.8   |

F. Attention Competition Mechanism

In this module, we try to find suitable hyperparameters $\lambda_3$ [see (13)] and $\gamma$ [see (12)] in the ACM. Table XI shows the main results. We conducted the experiments on the CUHK03-Detected test dataset with $K = 8$ set in MHSAM.

The competition loss functions in ACM are $L_{MHSAB} = L_{SAFFM} + \lambda_1 L_{FDRT} + \lambda_2 L_{IHTL} + L_{ReN}$ and $L_C$. The $L_{FDRT}$ and $L_{IHTL}$ are two terms we proposed here. To demonstrate the effectiveness of ACM individually, we set $\lambda_1 = 0$ and $\lambda_2 = 0$ in $L_{MHSAB}$. (Table V in Section VII-D shows the effectiveness of ACM combined with other contributions.)

Table XI shows that when $\gamma \leq 10^{-2}$ and $\lambda_3 \leq 10^{-1}$, the ACM effectively improves the performance of MHSAM in the person Re-ID task. When $\gamma$ is too big, the performance declines. This is because each element in $\alpha$ belongs to $[0, 1]$. If we set a too-big $\gamma$, attention weights on most regions will be suppressed by $L_C$. Based on these observations, we set $\lambda_3 = 10^{-3}$ and $\gamma = 10^{-3}$ in our MHSA-Net. In all, based on the suitable parameter setting, ACM can effectively improve the performance of the MHSAB.

VIII. LIMITATION AND DISCUSSION

Although our proposed MHSA-Net achieved the competitive performance in the occluded person Re-ID task, some limitations and discussion must be taken into consideration.
First, our proposed MHSANet mainly considers the situation of objects occluding person. The situation of person occluding person is not considered in the proposed MHSANet. Therefore, it is necessary to continue to optimize MHSANet in future work, to improve the model performance on the situation of person occluding person.

Besides, the person search requires a combination of the person detection task and the person Re-ID task in actual scenarios. Only relying on the person Re-ID model cannot effectively search for the target person. Therefore, it is necessary to study how to combine MHSANet with person detection models to build an end-to-end person search framework.

IX. CONCLUSION

We proposed a MHSANet to improve the ability of the Re-ID model on capturing the key information from the occluded person image. Specifically, we introduced the MHSAB to adaptively capture key local person information, and produce multiple diversity embeddings of one person image to facilitate person matching. We also designed an ACM to further help MHSAB prune out nonimportant local information. Extensive experiments were conducted to validate the effectiveness of each component in MHSANet, and they showed that MHSANet achieves competitive performance on three standard person Re-ID datasets and four occlusion person Re-ID datasets.

ACKNOWLEDGMENT

No conflict of interest: Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li declare that they have no conflict of interest.

REFERENCES

[1] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” 2017, arXiv:1703.07737.
[2] A. He, C. Luo, X. Tian, and W. Zeng, “A twofold Siamese network for real-time object tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 4834–4843.
[3] V. Ashish et al., “A structured self-attentive sentence embedding,” in Proc. Conf. Workshop Neural Inf. Process. Syst., 2017, pp. 1–15.
[4] S. Bai, X. Bai, and Q. Tian, “Scalable person re-identification on supervised smoothed manifold,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Los Alamitos, CA, USA, Jul. 2017, pp. 3356–3365.
[5] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, “Second-order non-local attention networks for person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3760–3769.
[6] X. Chen et al., “Salience-guided cascaded suppression network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3297–3307.
[7] W. Cheng, Z. Qian, H. Chang, L. Wenyu, and W. Xinggang, “Mancs: A multi-task attentional network with curriculum sampling for person re-identification,” in Proc. Eur. Conf. Comput. Vis. Pattern Recognit. (ECCV), Sep. 2018, pp. 365–381.
[8] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Pose-driven deep convolutional network for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3960–3969.
[9] C.-P. Tay, S. Roy, and K.-H. Yap, “AAANet: Attribute attention network for person re-identifications,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7134–7143.
[10] C. Song, Y. Huang, W. Ouyang, and L. Wang, “Mask-guided contrastive attention model for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 1179–1188.
[11] L. Chunli and W. Kejun, “A behavior classification based on enhanced gait energy image,” in Proc. Int. Conf. Netw. Digit. Soc., May 2010, pp. 589–592.
[12] Y. Dai, J. Liu, Y. Bai, Z. Tong, and L.-Y. Duan, “Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification,” IEEE Trans. Image Process., vol. 30, pp. 7815–7829, 2021.
[13] D. Xie, J. Xiong, and S. Pu, “All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6176–6185.
[14] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan, “Towards unified human parsing and pose estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 843–850.
[15] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, arXiv:1409.0473.
[16] F. Liu and L. Zhang, “View confusion feature learning for person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6639–6648.
[17] S. Gao, J. Wang, H. Lu, and Z. Liu, “Pose-guided visible part matching for occluded person ReID,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 1174–1183.
[18] G. Wang et al., “High-order information matters: Learning relation and topology for occluded person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6449–6458.
[19] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” 2018, arXiv:1804.01438.
[20] A. Haque, A. Alahi, and L. Fei-Fei, “Recurrence attention models for depth-based person identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1229–1238.
[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[22] H. Tan, X. Liu, X. Li, Y. Zhang, and B. Yin, “Semantics-enhanced adversarial nets for text-to-image synthesis,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 10501–10510.
[23] H. Xu, Z. Wang, H. Yang, D. Liu, and J. Liu, “Learning simple thresholded features with sparse reward support,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 4, pp. 970–982, Apr. 2020.
[24] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, “Adversarially occluded samples for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 5098–5107.
[25] C. Jean-Baptiste, L. Andreas, and J. Martin, “On the relationship between self-attention and convolutional layers,” in Proc. ICLR, 2020, pp. 1–18.
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.
[27] J. Si et al., “Dual attention matching network for context-aware feature sequence based person re-identification,” 2018, arXiv:1803.09937.
[28] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 375–383.
[29] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature alignment for occluded person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 542–551.
[30] J. Zhuo, Z. Chen, J. Lai, and G. Wang, “Occluded person re-identification,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2018, pp. 1–6.
[31] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, “Layer normalization,” 2016, arXiv:1607.06450.
[32] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, “Attention-aware compositional network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2119–2128.
[33] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, “Human semantic parsing for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1062–1071.
[34] M. E. Kalfaoglou, S. Kalkan, and A. A. Alatan, “Late temporal modeling in 3D CNN architectures with BERT for action recognition,” 2020, arXiv:2008.01232.
[35] N. Karianakis, Z. Liu, Y. Chen, and S. Soatto, “Person depth reid: Robust person re-identification with commodity depth sensors,” 2017, arXiv:1705.09882.
[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015, pp. 1–5.
[37] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, “Spatio-temporal attention networks for action recognition and detection,” IEEE Trans. Multimedia, vol. 22, no. 11, pp. 2990–3001, Nov. 2020.
[38] 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.

[39] L. He, J. Liang, H. Li, and Z. Sun, “Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, doi: 10.1109/CVPR.2018.00105.

[40] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, “Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification,” in Proc. ICCV, 2019, pp. 8450–8459.

[41] L. He, Z. Sun, Y. Zhu, and Y. Wang, “Recognizing partial biometric patterns,” 2018, arXiv:1810.07399.

[42] M. Domenico, A. Bemporad, E. Menegatti, and G. L. Van, “One-shot person re-identification with a consumer depth camera,” in Advances in Computer Vision and Pattern Recognition, 2014, pp. 161–181.

[43] X. Ning, K. Gong, W. Li, and L. Zhang, “JWSAA: Joint weak salience and attention feature for person re-identification,” Neurocomputing, vol. 453, pp. 801–811, Sep. 2021.

[44] P. Goyal et al., “Accurate, large minibatch SGD: Training ImageNet in 1 hour,” 2017, arXiv:1706.02677.

[45] H. Rao et al., “A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jun. 28, 2021, doi: 10.1109/TPAMI.2021.3092833.

[46] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 17–35.

[47] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, “Interaction-aware network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9317–9326.

[48] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng, “Discriminative feature learning with foreground attention for person re-identification,” IEEE Trans. Image Process., vol. 28, no. 9, pp. 4671–4684, Dec. 2019.

[49] S. Liao, A. K. Jain, and S. Z. Li, “Partial face recognition: Alignment-free approach,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 5, pp. 1193–1205, May 2013.

[50] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, “Gait energy volumes and frontal gait recognition using depth images,” in Proc. Int. Joint Conf. Biometrics (IJCB), Oct. 2011, pp. 1–6.

[51] T. Chen et al., “ABD-Net: Attentive but diverse person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8351–8361.

[52] Z. Wang, J. Jiang, Y. Wu, M. Ye, X. Bai, and S. Sato, “Learning sparse and identity-irrelevant hidden attributes for person re-identification,” IEEE Trans. Image Process., vol. 29, pp. 2013–2025, 2020.

[53] 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.

[54] W. Li, X. Zheng, S. Sun, and S. Gong, “Harmonious attention network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2285–2294.

[55] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, “Partial person re-identification,” in Proc. ICCV, 2015, pp. 4678–4686.

[56] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, “Towards rich feature discovery with class activation maps augmentation for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1389–1398.

[57] G. Wu, X. Zhu, and S. Gong, “Learning hybrid ranking representation for person re-identification,” Pattern Recognit., vol. 121, Jan. 2022, Art. no. 108239.

[58] W. Wu, D. Tao, H. Li, Z. Yang, and J. Cheng, “Deep features for person re-identification on metric learning,” Pattern Recognit., vol. 110, Feb. 2021, Art. no. 107424.

[59] X. Chang, T. M. Hospedales, and T. Xiang, “Multi-level factorisation net for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2109–2118.

[60] X. Qian et al., “Pose normalized image generation for person re-identification,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 650–667.

[61] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3069237.

[62] Y.-J. Cho and K.-J. Yoon, “Improving person re-identification via pose-aware multi-shot matching,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1354–1362.

---

Hongchen Tan received the Ph.D. degree in computational mathematics from the Dalian University of Technology, Dalian, China, in 2021. He is a Lecturer with the Artificial Intelligence Research Institute, Beijing University of Technology, Beijing, China. Various parts of his work have been published in top conferences and journals, such as the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE International Conference on Computer Vision, and Neurocomputing. His research interests are person re-identification, image synthesis, and object detection.
Xiuping Liu received the Ph.D. degree in computational mathematics from the Dalian University of Technology, Dalian, China, in 1999.
She is a Professor with the School of Mathematical Sciences, Dalian University of Technology. Her research interests include shape modeling and analyzing, and computer vision.

Baocai Yin received the M.S. and Ph.D. degrees in computational mathematics from the Dalian University of Technology, Dalian, China, in 1988 and 1993, respectively.
He is a Professor with the Artificial Intelligence Research Institute, Beijing University of Technology, Beijing, China. He is also a Researcher with the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing, and the Beijing Advanced Innovation Center for Future Internet Technology, Beijing. His research interests include multimedia, image processing, computer vision, and pattern recognition.

Xin Li (Senior Member, IEEE) received the B.E. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2003, and the M.S. and Ph.D. degrees in computer science from the State University of New York at Stony Brook, Stony Brook, NY, USA, in 2005 and 2008, respectively.
He is a Professor with the Division of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA, USA. His research interests are in geometric and visual data computing, processing, and understanding, computer vision, and virtual reality. For more detail, please see http://www.ece.lsu.edu/xinli.