The Visual Computer (2024) 40:3515–3527
https://doi.org/10.1007/s00371-023-03049-9

ORIGINAL ARTICLE

An efficient multi-scale channel attention network for person re-identification

Qian Luo[1,2] · Jie Shao[1] · Wanli Dang[2] · Long Geng[2] · Huaiyu Zheng[2] · Chang Liu[2]

Accepted: 27 July 2023 / Published online: 23 August 2023

Abstract
At present, occlusion and similar appearance pose serious challenges to the task of person re-identification. In this work, we propose an efficient multi-scale channel attention network (EMCA) to learn robust and more discriminative features to solve these problems. Specifically, we designed a novel cross-channel attention module (CCAM) in EMCA and placed it after different layers in the backbone. The CCAM includes local cross-channel interaction (LCI) and channel weight integration (CWI). LCI focuses on both the maximum pooling features and the average pooling features to generate channel weights through convolutional layers, respectively. CWI combines the two channel weights to generate richer and more discriminant channel weights. Experiments on four popular person Re-ID datasets (Market-1501, DukeMTMC-ReID, CUHK-03 (detected) and MSMT17) show that the performance of our EMCA is consistently significantly superior to the existing state-of-the-art methods.

Keywords Person re-identification · Convolutional neural network · Attention mechanism

1 Introduction

Person re-identification (Re-ID) has been extensively studied as a person search problem across non-overlapping cameras [1, 2]. Person Re-ID is an effective way to realize cross-site-specific pedestrian tracking and has very broad application prospects, such as intelligent video surveillance, intelligent security and intelligent traceability systems, which has become a hot topic in both research and industry. With the development of deep convolutional neural networks, learning with discriminant features by stacking convolutional and pooling layers has achieved advanced results in person Re-ID. However, person Re-ID can be affected by factors such as occlusion and similar appearance in real-world scenarios. As shown in Fig. 1a–c, targets are difficult to distinguish due to similar appearances. In Fig. 1d–f, the targets are obscured by other people or objects while walking. Figure 1 shows that BagTricks [3] are not able to solve these problems well. To effectively address these challenges, we designed a more discriminant person Re-ID model.

Before that, substantial efforts have been made to solve different challenges. Among them, combined with body part information [6–10] has empirically proved to be effective in enhancing the feature robustness against body misalignment, incomplete parts and occlusions. Inspired by this observation, attention mechanisms were introduced to enhance features to effectively capture the discriminative appearance of the human body. Since then, attention-based models [11–17] have greatly improved the person Re-ID performance.

In recent years, the Re-ID models [18–21] extract discriminative features by embedding the attention module to solve the problem of occlusion and similar appearance. Wu et al. [18] proposed an attention deep architecture with multi-scale deep supervision for person re-identification. This

✉ Qian Luo
luoqian@caacetc.com

Jie Shao
shaojie@stu.xhu.edu.cn

Wanli Dang
dangwanli@caacsri.com

Long Geng
genglong@caacsri.com

Huaiyu Zheng
zhenghuaiyu@caacsri.com

Chang Liu
liuchang@caacsri.com

[1] Xihua University, Chengdu 610039, China

[2] The Second Research Institute of the Civil Aviation Administration of China, Chengdu 610041, China

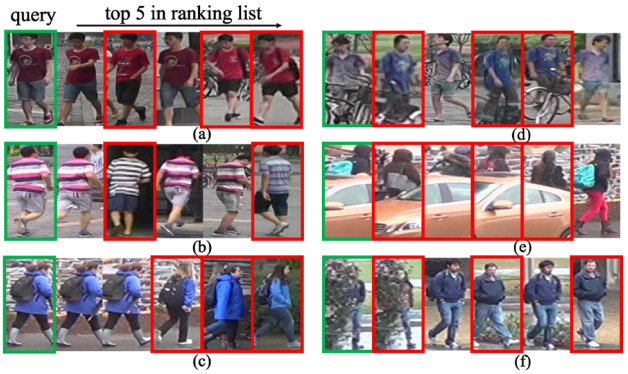

Fig. 1 Ranking result of the BagTricks [3] on the Market-1501 [4] and DukeMTMC-ReID [5]. The green boxes represent the query images, and the red boxes indicate the images with the wrong match

model adds attention modules at different stages of the backbone network to achieve more efficient multi-scale feature extraction. Zhao et al. [19] proposed a novel deep network composed of query-guided attention blocks, which enhanced the feature learning process of the target image in the gallery under the guidance of the query. Chen et al. [20] proposed a feature pyramid network (APNet) that adopt the "Squeeze-and-Excitation" block (SE) [22] as the channel-wise attention helps the model to focus on the more salient feature by assigning larger weight to channels that show a higher response. Gong et al. [21] proposed global–local attention to learn the semantic context in the channel and spatial dimensions by combining CBAM [23]. Although these networks perform well by embedding attention modules, there still exists two important and challenging problems. The first is how to effectively capture and exploit the information of feature map with different scales to enrich features. The second is that channel attention can only effectively capture the local information, while reducing the dimension through fully connected layers, and cause some key information to be lost.

Based on the above observations, we see it is necessary to develop an effective attention module. In this paper, we propose a novel lightweight attention module named CCAM, which can process the channel information of input features at different scales and can effectively capture more discriminative features without dimensionality reduction. As shown in Fig. 2, we see that CCAM can focus on more discriminant areas of the human body.

In summary, the contributions of this paper are summarized as follows:

- We propose a novel cross-channel attention module (CCAM) which consists of LCI and CWI. CCAM can process the channel information of input features at different scales and can effectively capture more discriminative features without dimensionality reduction.

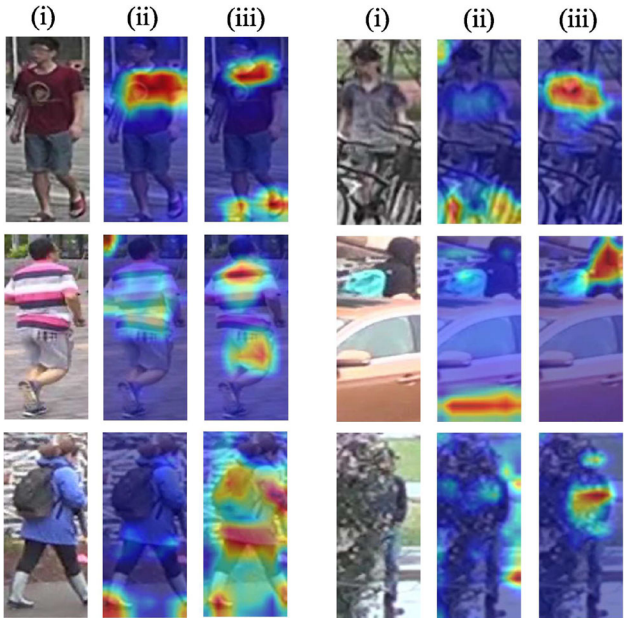

Fig. 2 Visualization of attention maps on Market-1501 [4] and DukeMTMC-ReID [5]. **i** Original images; **ii** heat map of BagTricks [3]; **iii** heat map of EMCA. BagTricks [3] is able to extract features effectively. However, it cannot capture some highly distinguishable features, but the EMCA can do it

- Based on Bagtricks [3], we design a network named EMCA by embedding CCAM into different layers of the network, which can capture more discriminative features to solve problems such as occlusion and similar appearance.
- We also analyze properties of four major public datasets, including Market-1501 [4], DukeMTMC-ReID [5], CUHK-03 (detected) [24] and MSMT17 [25]. By using the practical designs, training tricks and analyzed results, the proposed method achieves new state-of-the-art performance on all the four public datasets.

2 Related work

2.1 Brief overview of person re-identification

In recent years, with the rapid development of deep learning, the prevailing success of deep neural networks in computer vision has made human Re-ID no exception. A person Re-ID model typically consists of two components: a feature extractor and a similarity metric. A line of works focused on improving either the feature extractor, the similarity metric or both of them. For a better feature extractor, many methods [10, 26–28] concentrate on designing local-based models that divide the body into different parts from which features are extracted and fused to obtain a robust representation. To mine more clues as the prior knowledge, some methods uti-

lize human pose information [9, 29–31] for accurate part detection or person normalization. For a better similarity metric, some methods [3, 11, 20, 32] combined triplet loss with identification loss and jointly learned a metric in the model. Zhou et al. [7] trained the model with focal triplet loss, which imposes a constraint on the intra-class and an adaptively weight adjustment mechanism to handle the hard sample problem. Liao et al. [33] proposed an efficient mini-batch sampling method for large-scale deep metric learning.

2.2 Attention mechanisms in person Re-ID

Recently, attention mechanisms [22, 23, 34] have been widely used for a variety of visual tasks. Now, the attention mechanism is also successfully applied in Re-ID tasks. Chen et al. [12] proposed the holistic attention branch (HAB) to make the feature maps obtained by backbone could focus on persons so as to alleviate the influence of background, and partial attention branch (PAB) is proposed to make the extracted features can be decoupled into several groups that are separately responsible for different body parts, thus increasing the robustness to pose variation and partial occlusion. Wang et al. [13] proposed a batch coherence-guided channel attention (BCCA) module that highlights the relevant channels for each respective part from the output of a deep backbone model. Sun et al. [14] proposed a multi-level attention embedding and multilayer feature fusion (MEMF) model, in which the multi-level attention block can highlight representative features and assist global feature expression and multilayer feature fusion can increase the fine-grained feature expression. Zhong et al. [15] proposed a progressive feature enhancement (PFE) algorithm that filters pixel-wise and channel-wise noises on the intermediate feature maps through a two-stage attention module (TSAM) to further facilitate the layer-specific feature generation. Rao et al. [16] presented a counterfactual attention learning method to learn more effective attention based on causal inference. They proposed to learn the attention with counterfactual causality, which provides a tool to measure the attention quality and a powerful supervisory signal to guide the learning process. Motivated by the tremendous success of self-attention [35–39] in re-identification tasks, Yan et al. [17] proposed a cross-attention layer that associates different images of the same identity to learn attention maps that are effective across all these images. This reduces the discrepancy of the attention across different images of the same identity.

Our proposed attention mechanism differs from previous methods in several aspects. First, previous methods cannot effectively acquire and utilize channel information for feature maps at different scales, resulting in the loss of fine-grained feature. In contrast, CCAM aim at processing the global information of multi-scale input features and learning attention weights without dimensionality reduction to obtain the more detailed and discriminative feature representation. Second, although previous methods achieve high performance, their multiple branches and multiple tasks make the model structure too complex. Our attention maps are directly learned from the data and context, without relying on manually defined parts, pose estimation, nor part region proposals. CCAM is a plug-and-play lightweight attention module that is embedded within a single backbone, making our model more lightweight than the multitask learning alternatives [12, 14, 15].

2.3 Review the BagTricks

We briefly introduce our baseline architecture: BagTricks [3]. As illustrated in Fig. 3, BagTricks [3] is a strong baseline that collects some effective training tricks and adopts ResNet-50 as the backbone. The stride parameter of conv5_x is set to 1 instead of 2 to preserve more details. The global average pooling layer converts the output feature maps of conv5_x into feature vectors, which are fed to triplet loss for metric learning at the training stage. The feature vectors are linearly scaled to final feature vectors by a batch normalization (BN) layer, which are fed to ID loss (usually cross-entropy loss) at the training stage. The final feature vectors are directly used to compute the distance matrix at the inference stage.

BagTricks [3] serves as a strong baseline and has been proved to boast good performance. Although BagTricks [3] can extract features efficiently, it cannot capture highly differentiated features. We embedded CCAM into the backbone and achieved impressive performance. In Fig. 2, we see that EMCA can focus on the discriminative human body regions compared to BagTricks [3]. We perform extensive experiments on Market-1501 [4], DukeMTMC-ReID [5], CUHK-03 (detected) [24] and MSMT17 [25] and show that EMCA significantly outperforms BagTricks [3].

3 Proposed method

In this section, we first briefly describe the overall architecture of EMCA. Then, we will introduce LCI and CWI in CCAM in detail, respectively.

3.1 The structure of EMCA

The overall architecture of the proposed EMCA is shown in Fig. 4, which is modified from the baseline depicted in Fig. 3. The input part, backbone and loss functions are inherited. EMCA has three important components: ResNet-50, CCAM and the loss functions.

Major update of the architecture is the insertion of channel attention modules between convolutional layers to focus on features from different scales. In Fig. 4, we add the designed

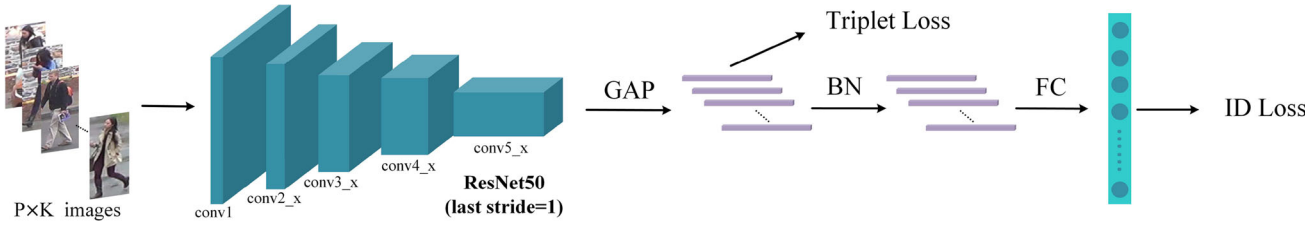

Fig. 3 Model architecture of BagTricks [3]. The backbone of the network is ResNet-50. The convolution stride of conv5_x is changed from 2 to 1 to retain more image details. Then, the features of conv5_x go through a global average pooling (GAP) and become a vector feature, which is fed to triplet loss at the training stage

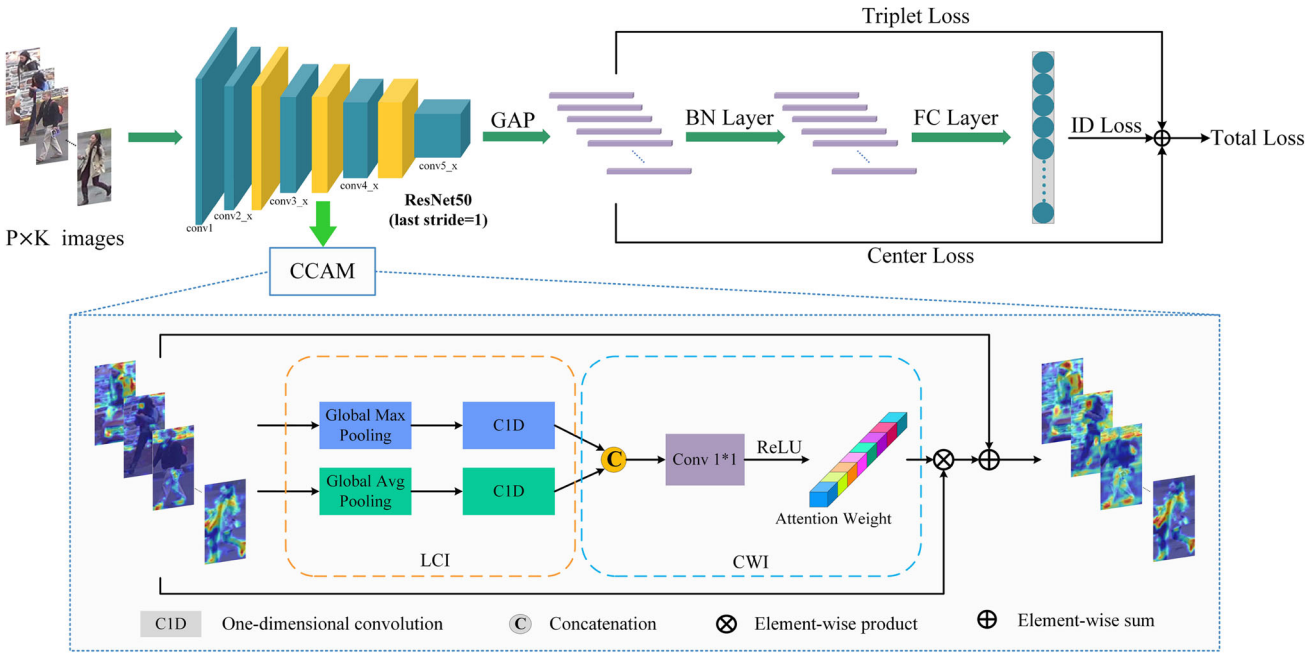

Fig. 4 EMCA architecture, which is modified from Fig. 2. The backbone of the network is ResNet-50 and place CCAM after conv2_x, conv3_x and conv4_x. In the training stage, ID loss, triplet loss and center loss are used to optimize model parameters

CCAM after conv2_x, conv3_x and conv4_x layer, which can effectively extract different levels of discriminative features. Given an intermediate feature map, the LCI and CWI in CCAM calculate the attention weights of the features to obtain the attention feature map.

In order to obtain a more robust Re-ID model, EMCA is trained under the loss function L_{total} consisting of ID loss (cross-entropy loss) L_{ID}, triplet loss L_{Tri} and center loss L_{C}:

$$L_{\text{total}} = L_{\text{ID}} + L_{\text{Tri}} + \beta L_{\text{C}} \tag{1}$$

where β is the balanced weight of center loss. ID loss, triplet loss and center loss play their respective roles in the model. ID loss can distinguish between positive and negative samples to some extent, but it is difficult to handle some difficult samples. Triplet loss can shorten the distance to positive samples while training models, while maximizing distance from negative samples. Center loss learns a center for the deep features of each class, which makes up for the absolute distance that triplet loss ignores them. Specifically, we add the triplet loss and center loss after the global average pooling (GAP) layer and place the ID loss after the fully connected (FC) layer.

3.2 The cross-channel attention module

The overall architecture of our proposed channel attention module is shown in Fig. 5. Given the input feature $X^l \in \mathbb{R}^{C\times H\times W}$ form lth layer, we get $W_f^l \in \mathbb{R}^{C\times 1\times 1}$ after LCI and CWI. Multiply the W_f^l with X^l to get the weighted attention feature $A^l \in \mathbb{R}^{C\times H\times W}$, then add A^l and X^l to obtain the final discriminant features $X^l \in \mathbb{R}^{C\times H\times W}$, as shown below:

$$\widetilde{X^l} = A^l + X^l = (X^l \otimes W_f^l) + X^l \tag{2}$$

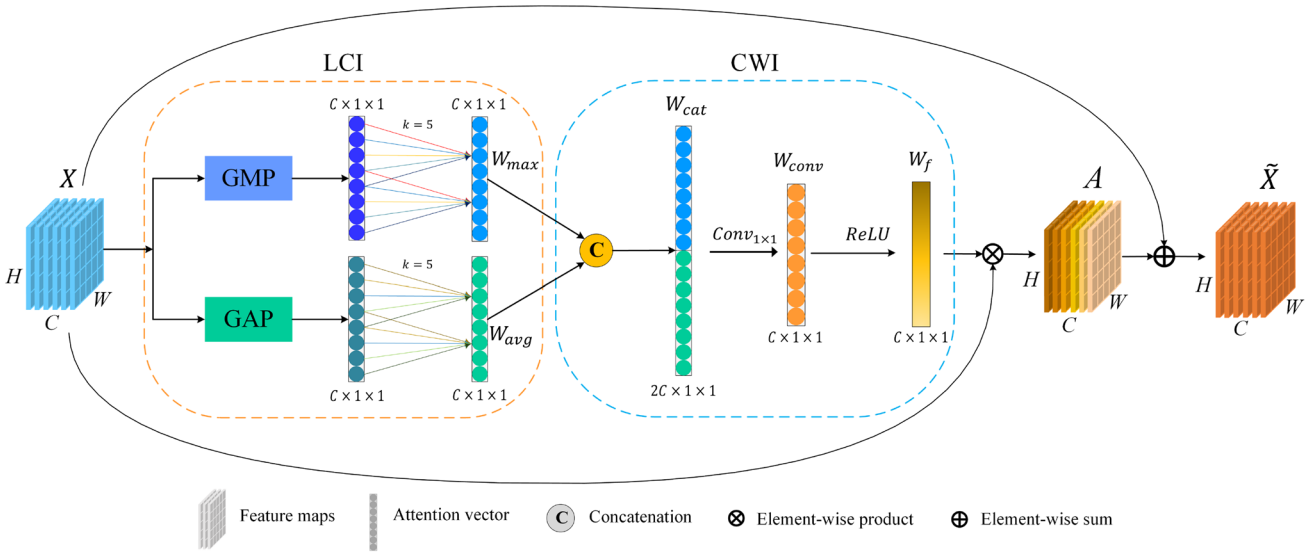

Fig. 5 Structure of cross-channel Attention Module. CCAM includes LCI and CWI. LCI focuses on both the maximum pooling feature and the average pooling feature, and the two types of features interact locally across channels to generate corresponding channel weights. CWI integrate channel weights from the two channel weights

The residual connection $(+X^l)$ allows us to insert a new block in any pretrained model without breaking its initial performance.

3.2.1 Local cross-channel interaction

It is well known that different feature channels represent different feature information. [22, 23] generated attention weights through dimensionality reduction of fully connected layers. Wang et al. [34] believed that the dimensionality reduction operation in this method will have side effects on channel attention prediction and proposes a local cross-channel interaction strategy without dimensionality reduction. Inspired by this view, we construct LCI to perform local cross-channel operations to aggregate local channel information. Many person Re-ID models typically use average pooling features to aggregate spatial information, which results in feature data that is more sensitive to background information. We think that the maximum pooling collects another important clue about the object feature. Unlike the previous approach, we not only use the average pooling features, but also use both the average pooling features and the maximum pooling features. As shown in Fig. 5, the working principle of LCI is mainly divided into two steps. The first step is to give an input feature map $X^l \in \mathbb{R}^{C\times H\times W}$, where C is the total number of channels, $H \times W$ is the size of the feature map and l is a layer of the network. Then, the maximum pooling feature $F^l_{max} \in \mathbb{R}^{C\times 1\times 1}$ and the average pooling feature $F^l_{avg} \in \mathbb{R}^{C\times 1\times 1}$ were extracted from layer l, respectively, as shown below:

$$F^l_{\max} = \mathrm{GMP}(X^l) = \underset{\substack{i=1,\cdots,H\\ j=1,\cdots,W}}{\mathrm{Max}} \{X^l_{ij}\} \tag{3}$$

$$F^l_{\mathrm{avg}} = \mathrm{GAP}(X^l) = \frac{1}{\mathrm{WH}} \sum_{i=1}^{H}\sum_{j=1}^{W} X^l_{ij} \tag{4}$$

where GMP is the global maximum pooling and GAP is the global average pooling. In the second step, we perform fast 1D convolution of size k on the resulting $F^l_{\max}$ and F^l_{avg} to generate maximum pooling channel weights $W^l_{\max} \in \mathbb{R}^{C\times 1\times 1}$ and average pooling channel weights $W^l_{\mathrm{avg}} \in \mathbb{R}^{C\times 1\times 1}$, which are obtained by:

$$W^l_{\max} = \mathrm{C1D}_k(\mathrm{GMP}(X^l)) = \mathrm{C1D}_k(F^l_{\max}) \tag{5}$$

$$W^l_{\mathrm{avg}} = \mathrm{C1D}_k(\mathrm{GAP}(X^l)) = \mathrm{C1D}_k(F^l_{\mathrm{avg}}) \tag{6}$$

where $\mathrm{C1D}_k$ indicates 1D convolution, the convolution kernel size is k. Here, Eqs. (5) and (6) guarantee efficiency by appropriately capturing local cross-channel interactions through one-dimensional convolution.

3.2.2 Channel weight integration

Previously in the LCI component, we obtained the channel weights $W^l_{\max}$ and W^l_{avg}. We design CWI to integrate channel weights from the two channel weights which have more comprehensive and discriminative information. As shown in Fig. 5, CWI works in two main steps. First, CWI connects the $W^l_{\max}$ and the W^l_{avg} in the 0th dimension to obtain the

connected channel weights $W^l_{\text{cat}} \in \mathbb{R}^{2C\times1\times1}$. It can be represented by the following equation:

$$W^l_{\text{cat}} = \text{Concat}_{0d}(W^l_{\max}, W^l_{\text{avg}}) \tag{7}$$

where Concat_{0d} means that two tensors are connected along the zeroth dimension. Immediately after, the W^l_{cat} through convolutional layers of size 1×1 to extract richer channel weights $W^l_{\text{conv}} \in \mathbb{R}^{C\times1\times1}$. Finally, we use the ReLU activation function to get the final channel weight $W^l_f \in \mathbb{R}^{C\times1\times1}$, as shown below:

$$W^l_f = \text{ReLU}(W^l_{\text{conv}}) = \text{ReLU}(\text{Conv}_{1\times1}(W^l_{\text{cat}})) \tag{8}$$

where $\text{Conv}_{1\times1}$ represents a convolutional block with a convolutional kernel size of 1×1.

3.3 Discussion

In this subsection, we discuss the proposed attention module with other attention modules structures and explain why our CCAM is effective and efficient. As shown in Fig. 6, we compare three different attention structures including Squeeze-and-Excitation module (SE) [22] and the channel attention module of CBAM (CBAM-C) [23], ECA [34] with our CCAM.

In SE [22] module, they use spatially global average-pooled features to compute channel-wise attention, by using two fully connected (FC) layers with the nonlinearity. CBAM-C [23] is similar to SE [22] module, but it additionally uses global maximum pooling features, and finally adds the global average pooling features to the global maximum pooling features to obtain channel weights with comprehensive feature information. In ECA [34] module, they found that avoiding dimensionality reduction is important for learning channel attention. They also use the global average pooling features to compute channel-wise attention, but unlike the structure of SE [22], they use one-dimensional convolution to obtain channel weights instead of fully connected layers to avoid dimensionality reduction.

Compared to these attention modules mentioned above, the proposed CCAM concentrates the advantages of these attention modules and is more efficient than them. In general, the advantages of CCAM over these attention modules are: (1) The global average pooling feature and the global maximum pooling feature are used simultaneously to obtain richer feature information; (2) one-dimensional convolution is used to replace the fully connected layer to avoid the loss of feature information caused by dimensionality reduction (we perform comparative experiments on fully connected layers and one-dimensional convolution in Table 7 of Sect. 4.4); (3) the two types of weights are concatenated instead of added, and the final channel weights are integrated through the convolutional layer. Finally, we make CCAM with these attention modules for comparison experiments, and the results are shown in Table 6 of Sect. 4.4.

4 Experiment

To evaluate our model, we conducted experiments on four popular person re-identification datasets: Market-1501 [4], DukeMTMC-ReID [5], CUHK-03 (detected) [24] and MSMT17 [25]. First, we compare the performance of our model with existing comparative advanced methods on four datasets. Second, we reported a set of ablation experiments to verify the effectiveness of each component. Finally, we

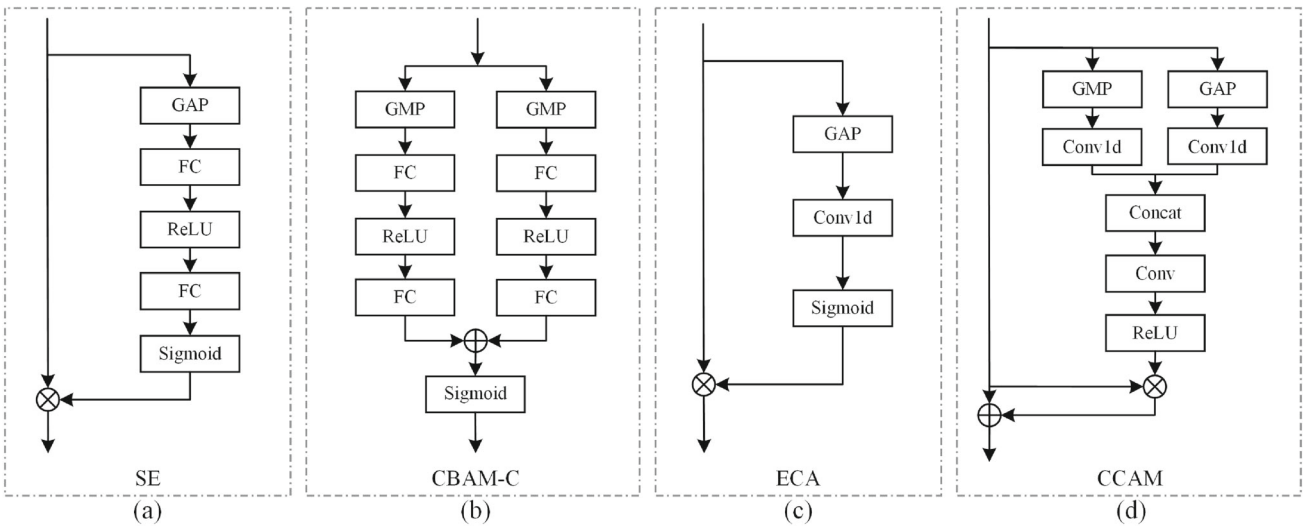

Fig. 6 Schematic comparison between **a** SE [22], **b** channel attention module of CBAM (CBAM-C) [23], **c** ECA [34] and **d** CCAM

Table 1 Statistics of used datasets

Dataset	#total ID	#training ID	#gallery ID	#image	#gallery image	#camera
Market-1501 [4]	1501	751	752	32,668	19,732	6
DukeMTMC-ReID [5]	1404	702	1110	36,411	17,661	8
CUHK-03 (detected) [24]	1467	767	700	13,161	5332	2
MSMT17 [25]	4101	1041	3060	126,441	82,161	15

provide more visual analysis to illustrate the effectiveness of our model.

4.1 Datasets

Understanding the datasets is the most important step in person Re-ID. The person Re-ID datasets consists of training set, gallery set and query set. The training set is used to train the model, the query set and gallery set are used to test the model. We evaluate our method on four popular person re-identification datasets, which are described in detail in Table 1.

Market-1501 [4] contains 32,668 labeled images of 1501 identities captured by six cameras, which are almost equally divided into a training set and a test set (gallery + query). There are 12,936 images of 751 identities in the training set, while the rest are used for testing. There are 750 identities in the test set, including 3,368 query images and 19,732 gallery images.

DukeMTMC-ReID [5] provided a large dataset recorded by eight cameras, which included 36,411 labeled images of 1404 identities. The 1404 identities are randomly divided, with 702 identities for training and the others for testing. Among them, 16,522 images of 702 identities are used for training, and 2228 query images and 17,661 gallery images of 1110 identities for testing.

CUHK-03 (detected) [24] consists of two separate datasets: Detected and Labeled. The difference is how labels are generated. Labels of CUHK-03 (detected) are without manual correction, which is easy to obtain, close to the industry but more difficult to cope with. The dataset contains 13,161 images with 1467 person IDs split into training and test sets without overlap. These images are captured by only two cameras.

MSMT17 [25] is the current largest publicly available person Re-ID dataset. It has 126,441 images of 4101 identities captured by a 15-camera network, which are also split into training and test sets without overlap. Note the gallery set contains 82,161 images of 3060 person IDs. The numbers are much greater than those in the training set. MSMT17 is significantly more challenging than the other three, due to its massive scale, more complex and dynamic scenes.

4.2 Implementation details and evaluation

In the experiment, we used an RTX 3090 GPU with 24GB RAM for training. Bagtricks [3] introduced tricks including warm-up learning rate and adjusted the last stride of conv5_x layer as 1 and data argumentation with random erasing, which we confirm to be effective. Label smoothing is not harmful, which we keep. The backbone of the network is ResNet-50 initialized with pretrained parameters on ImageNet [40], and the dimension of the FC layer in the network was changed to the number of identities N in the dataset. During training, we randomly select K images of P identities to form a training batch. In this work, we set $P = 16$ and $K = 4$. The images in the batch are re-sized to 256×128 pixels, and then, each image is flipped, normalized and randomly erased at a probability level of 0.5. The Adam method is used to optimize the model. We set up the model for a total of 240 training epochs. In practice, we spent 10 epochs linearly increasing the learning rate from 3.5×10^{-5} to 3.5×10^{-4}. Then, the learning rate is decayed to 3.5×10^{-5} and 3.5×10^{-6} at 50th epoch and 90th epoch respectively. In equation (1), the weight size of Center Loss β set to 0.0005. For the size of the one-dimensional convolution kernel k in Eqs. (5) and (6), we refer to the parameter setting of [34], which is set to 5.

We adopt the cumulative match characteristic (CMC) [41] and the mean average precision (mAP) [4] as evaluation indicators for our model. The CMC curve shows the recognition accuracy of Rank-n, which can effectively evaluate the performance of the model. The abscissa of CMC curve is Rank-n, where $n = 1, 5, 10$, etc. The ordinate is the recognition accuracy. The mAP represents the average accuracy of correctly retrieving the specified identities in the database. Combining CMC curve and mAP can comprehensively measure the performance of the model.

4.3 Comparison to state-of-the-art methods

We compare EMCA with other methods on Market-1501 [4], DukeMTMC-ReID [5], CUHK-03 (detected) [24] and MSMT17 [25], as shown in Table 2, 3, 4 and 5, respectively. For fair comparisons, no post-processing such as re-ranking strategies or multi-query fusion was used for our methods.

Table 2 Comparison to state-of-the-art methods on Market-1501

Methods	Publications	mAP	Rank-1
BagTricks [3]	CVPR 2019	85.9	94.5
PISNet [19]	ECCV 2020	87.1	95.6
AGW [42]	arXiv 2020	87.8	95.1
GASM [43]	ECCV 2020	84.7	95.3
PAT [35]	CVPR 2021	88.0	95.4
PFE [15]	TIP 2021	87.5	95.2
FA-Net [44]	TIP 2021	84.6	95.0
ADC-2OIB [45]	CVPR 2021	87.7	94.8
APNet-S [20]	TIP 2021	89.0	96.1
TransReID [36]	ICCV 2021	88.9	95.2
BINet [46]	TIP 2021	88.7	95.3
OSNet [28]	TPAMI 2021	84.7	93.1
CAL [47]	CVPR 2022	87.5	94.7
FED [48]	CVPR 2022	86.3	95.0
EMCA (Ours)	–	89.1	95.6

Table 3 Comparison to state-of-the-art methods on CUHK-03 (detected)

Methods	Publications	mAP	Rank-1
BagTricks [3]	CVPR 2019	56.6	58.8
LSTS-NET [49]	IJCV 2020	67.9	70.11
AGW [42]	arXiv 2020	62.0	63.6
DPD [50]	TIP 2020	68.5	70.2
BINet [46]	TIP 2021	69.8	72.3
OSNet [28]	TPAMI 2021	67.8	72.3
LReID [51]	CVPR 2021	50.8	56.16
EMCA (Ours)	–	69.8	73.6

Table 4 Comparison to state-of-the-art methods on DukeMTMC-ReID

Methods	Publications	mAP	Rank-1
BagTricks [3]	CVPR 2019	76.4	86.4
PISNet [19]	ECCV 2020	78.8	88.8
GASM [43]	ECCV 2020	74.4	83.3
AGW [42]	arXiv 2020	79.6	89.0
PAT [35]	CVPR 2021	53.6	64.5
ADC-2OIB [45]	CVPR 2021	74.9	87.4
PFE [15]	TIP 2021	77.1	89.2
FA-Net [44]	TIP 2021	77.0	88.7
APNet-S [20]	TIP 2021	78.8	89.3
TransReID [36]	ICCV 2021	80.6	89.6
BV-Person [17]	ICCV 2021	80.6	90.5
OSNet [28]	TPAMI 2021	76.6	88.7
FED [48]	CVPR 2022	78.0	89.4
EMCA (Ours)	–	80.8	90.6

Table 5 Comparison to state-of-the-art methods on MSMT17

Methods	Publications	mAP	Rank-1
BagTricks [3]	CVPR 2019	49.0	73.0
AGW [42]	arXiv 2020	49.3	68.3
GASM [43]	ECCV 2020	52.5	79.5
PFE [15]	TIP 2021	56.2	80.1
FA-Net [44]	TIP 2021	51.0	76.8
BINet [46]	TIP 2021	52.8	76.1
OSNet [28]	TPAMI 2021	55.1	79.1
LReID [51]	CVPR 2021	27.9	54.1
CAL [47]	CVPR 2022	57.3	79.7
EMCA (Ours)	–	56.8	80.1

Table 2 shows that our EMCA achieved 89.1% mAP and 95.6% Rank-1 on the Market-1501 [4]. Its Rank-1 accuracy is slightly 0.5% lower than APNet [20] and the same as Rank-1 accuracy in PISNet [19], yet EMCA clearly surpasses all methods in terms of mAP. Table 3 shows that EMCA achieved 69.8% mAP and 73.6% Rank-1 on CUHK-03 (detected) [24]. Its accuracy in terms of mAP is equal to BINet [46], but its Rank-1 is higher than all methods. Table 4 shows that our EMCA achieves 80.8% mAP and 90.6% Rank-1 on DukeMTMC-ReID [5], which is significantly better than all methods. Table 5 shows that EMCA achieved 56.8% mAP and 80.1% Rank-1 in MSMT17 [25]. Its mAP accuracy is slightly 0.5% lower than CAL [47] and its Rank-1 accuracy is equal to PFE [15], yet EMCA is clearly higher than all methods in terms of mAP and Rank-1.

On Market-1501 [4], DukeMTMC-ReID [5], CUHK-03 (detected) [24] and MSMT17 [25], our EMCA clearly outperformed the baseline BagTricks [3], which is 3.2%, 4.4%, 13.2% and 7.8% higher than the baseline on mAP and which is 1.1%, 4.2%, 14.8% and 7.1% higher than the baseline on Rank-1, respectively. In addition, Compared with AGW [42], which are also designed based on BagTricks [3], our EMCA performs better than them. The above experiments can prove that EMCA is a powerful person Re-ID model.

4.4 Ablation study

The Effect of CCAM To study the effectiveness of the CCAM, we add CCAM to the baseline network. We choose some advanced attention modules for comparison, including SE [22], the channel attention of CBAM (CBAM-C) [23] and ECA [34]. For the fairness of comparison, we re-implement these attention modules on top of the baseline. Table 6 indicates the experimental results of these attention modules. We found that CCAM was significantly superior to other attention modules on four datasets. Experiments have shown that our CCAM has a significant impact on baseline performance.

Table 6 mAP and the Rank-1 of CMC are used to evaluate the effectiveness of CCAM on four datasets

Methods	Market-1501		DukeMTMC-ReID		CUHK-03 (detected)		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	85.9	94.5	76.4	86.4	56.6	58.8	49.0	73.0
+SE	86.8	94.2	76.9	87.5	67.4	69.9	52.2	76.5
+CBAM-C	87.2	94.8	77.8	87.7	67.4	69.9	53.1	76.7
+ECA	87.3	95.1	77.7	87.7	66.4	69.1	52.4	76.2
EMCA (Ours)	89.1	95.6	80.8	90.6	69.8	73.6	56.8	80.1

Table 7 Evaluate the effectiveness of feature selection on four datasets

Methods	Market-1501		DukeMTMC-ReID		CUHK-03 (detected)		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	85.9	94.5	76.4	86.4	56.6	58.8	49.0	73.0
+GMP	82.3	92.8	73.8	86.2	53.9	55.7	47.2	71.5
+GAP	86.8	94.9	77.5	88.5	65.9	68.3	54.8	78.0
+FC Layer	88.1	94.9	78.8	88.8	67.6	70.3	55.4	79.0
EMCA (Ours)	89.1	95.6	80.8	90.6	69.8	73.6	56.8	80.1

Effective Feature Selection. In order to verify the effectiveness of using both the maximum pooling features and the average pooling features, we experiment on Market-1501 [4], DukeMTMC-ReID [5], CUHK-03 (detected) [24] and MSMT17 [25]. We use the GMP and GAP separately on a baseline basis. In Table 7, we found that using only GMP in CCAM performed less than baseline, while using only GAP performed better than baseline. Obviously, when EMCA uses both GMP and GAP, its mAP and Rank-1 accuracy are higher than using one of features alone and certainly higher than baseline. In addition, we experimented with the effect of CCAM with FC layer and found that its accuracy is not as high as CCAM with one-dimensional convolution. Figure 7 shows the CMC curve of the EMCA on four datasets. We noticed that EMCA achieved the best performance. This indicates the correct selection of features and confirms that EMCA performs optimally when using both the maximum pooling features and the average pooling features with one-dimensional convolution.

Efficient Position to Place Attention Module. Figures 8 and 9 compare the performance of the CCAM after the different layers of ResNet-50, where layer2 means placing the attention module after conv2_x, and so on. We verify the performance of placing the CCAM between conv2_x, conv3_x and conv4_x, respectively. As shown in Figs. 8 and 9, mAP and Rank-1 of CMC have the highest accuracy when CCAM is placed after conv2_x, conv3_x and conv4_x. Obviously, EMCA works best by placing CCAM after conv2_x, conv3_x and conv4_x at the same time.

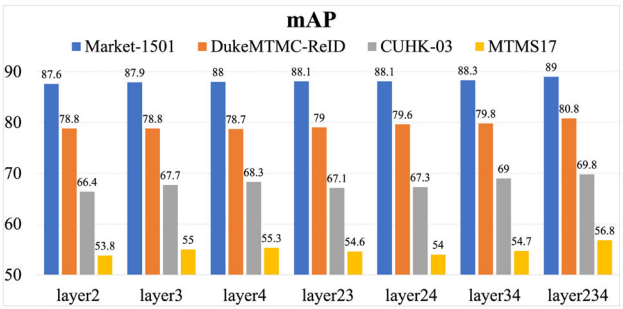

Fig. 8 Evaluate the mAP of effective placement of attention module on four datasets

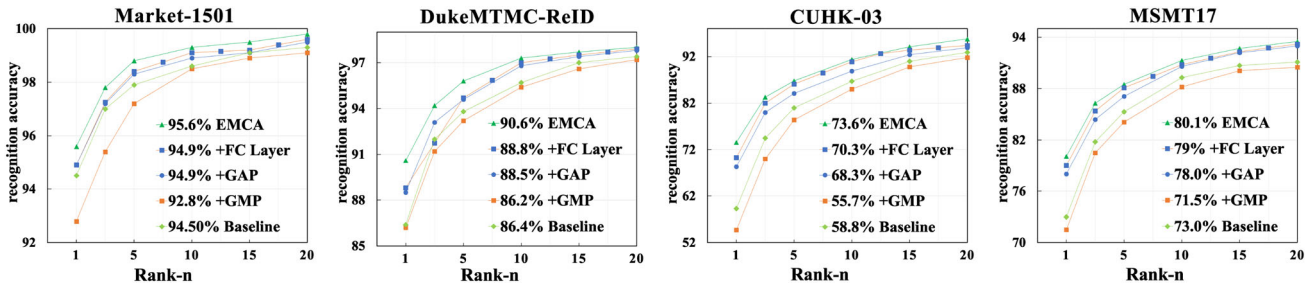

Fig. 7 CMC curve of Baseline, Baseline+GMP, Baseline+GAP, EMCA with FC Layer and EMCA on the four datasets

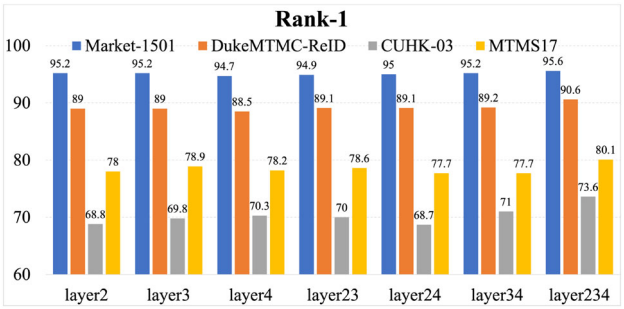

Fig. 9 Evaluate the Rank-1 of effective placement of attention module on four datasets

4.5 Visualizations

Attention Pattern Visualization. We use the Grad-CAM [52] tool to analyze the attention map of Baseline, Baseline+GMP, Baseline+GAP and EMCA. The Grad-CAM tool marks areas that the model considers important. The redder the marked area is, the more important it is. We conduct a set of attention visualizations on final output feature maps of the Baseline, Baseline+GMP, Baseline+GAP and EMCA, as shown in Fig. 10. We notice that the feature maps from the baseline show little attentiveness. The attention maps of Baseline+GMP and Baseline+GAP pay little attention to human features. In contrast, we see that the attention maps of EMCA are focus on more parts of the human body, which can more effectively capture the discriminative features of person and pay little attention to irrelevant information around person.

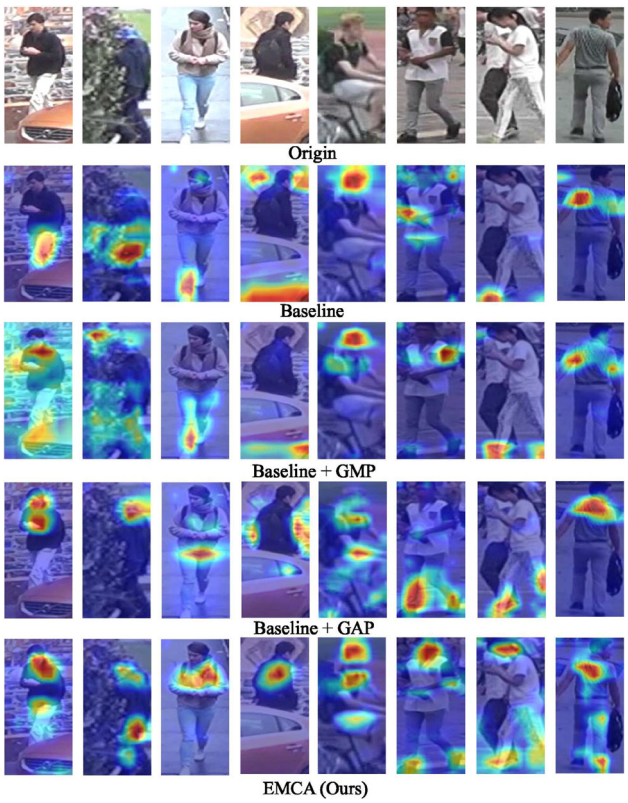

Fig. 10 Visualization of attention maps from Baseline, Baseline+GMP, Baseline+GAP and EMCA. As shown in row five, the attention map of EMCA can focus on more areas of the foreground, rather than focusing on the background and few person features

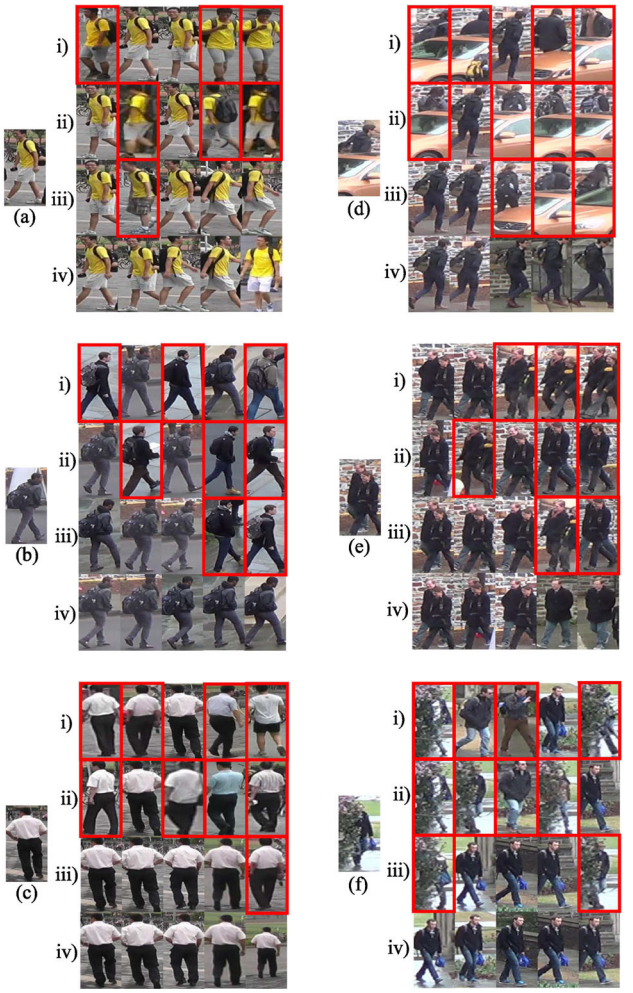

Fig. 11 Six Re-ID examples of Baseline, Baseline+GMP, Baseline+GAP and EMCA on Market-1501 and DukeMTMC-ReID. Left: query image. Right: **i** rank-5 results of Baseline. **ii** rank-5 results of Baseline+GMP. **iii** rank-5 results of Baseline+GAP. **iv** rank-5 results of EMCA. Images in red boxes are negative results. Among them, EMCA retrieval performance is the best

Re-ID Qualitative Visual Results. Figure 11 shows Re-ID visual example of Baseline, Baseline+GMP, Baseline+GAP and EMCA. We observed that EMCA had no matching errors compared to Baseline, Baseline+GMP and Baseline+GAP, which indicates that our method solved the problem of occlusion and appearance similarity to a certain extent.

5 Conclusion

In this work, we design a network named EMCA to learn more representative, robust and discriminative feature embeddings to solve occlusion and similar appearance problems in Re-ID tasks. In the model, we propose a novel attention module named CCAM, which can process the channel information of input features at different scales and can effectively capture more discriminative features without dimensionality reduction. CCAM consists of LCI and CWI. LCI focuses on both the maximum pooling features and the average pooling features to generate channel weights separately, and CWI combines the two channel weights to generate richer and more discriminant channel weights. EMCA demonstrated its state-of-the-art performance through extensive experiments on four datasets, where ablation studies and visualizations showed the effectiveness of the model structure and each added component.

However, image-based person re-identification is very difficult and limited to solve problems such as occlusion and similar appearance. With the excellent results of a large number of video-based person re-identification research works, we observed that the information in the video frames have greater research value for alleviating the problems of occlusion and appearance similarity. In the future, we can incorporate more effective attention modules into video-based person re-identification networks to address the challenges of occlusion and similar appearance.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Qian Luo, Jie Shao and Wanli Dang. The first draft of the manuscript was written by Jie Shao, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This document is the results of the research project funded by the NNSFC and CAAC (U2133211) and the Young Scientists Fund of the National Natural Science Foundation of China (62203452).

Data availability The data and material used or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Y.-C., Zhu, X., Zheng, W.-S., Lai, J.-H.: Person re-identification by camera correlation aware feature augmentation. IEEE Trans. Pattern Anal. Mach. Intell. **40**(2), 392–408 (2017)
2. Zahra, A., Perwaiz, N., Shahzad, M., Fraz, M.M.: Person re-identification: a retrospective on domain specific open challenges and future trends. arXiv preprint arXiv:2202.13121 (2022)
3. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
4. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)
5. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, pp. 17–35. Springer (2016)
6. Yin, J., Wu, A., Zheng, W.-S.: Fine-grained person re-identification. Int. J. Comput. Vis. **128**, 1654–1672 (2020)
7. Zhou, Q., Zhong, B., Lan, X., Sun, G., Zhang, Y., Zhang, B., Ji, R.: Fine-grained spatial alignment model for person re-identification with focal triplet loss. IEEE Trans. Image Process. **29**, 7578–7589 (2020)
8. Gao, S., Wang, J., Lu, H., Liu, Z.: Pose-guided visible part matching for occluded person reid. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11744–11752 (2020)
9. Li, Z., Lv, J., Chen, Y., Yuan, J.: Person re-identification with part prediction alignment. Comput. Vis. Image Underst. **205**, 103172 (2021)
10. Wang, P., Zhao, Z., Su, F., Zu, X., Boulgouris, N.V.: Horeid: deep high-order mapping enhances pose alignment for person re-identification. IEEE Trans. Image Process. **30**, 2908–2922 (2021)
11. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8351–8361 (2019)
12. Chen, Y., Wang, H., Sun, X., Fan, B., Tang, C., Zeng, H.: Deep attention aware feature learning for person re-identification. Pattern Recognit. **126**, 108567 (2022)
13. Wang, K., Wang, P., Ding, C., Tao, D.: Batch coherence-driven network for part-aware person re-identification. IEEE Trans. Image Process. **30**, 3405–3418 (2021)
14. Sun, J., Li, Y., Chen, H., Zhang, B., Zhu, J.: Memf: multi-level-attention embedding and multi-layer-feature fusion model for person re-identification. Pattern Recognit. **116**, 107937 (2021)
15. Zhong, Y., Wang, Y., Zhang, S.: Progressive feature enhancement for person re-identification. IEEE Trans. Image Process. **30**, 8384–8395 (2021)
16. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1025–1034 (2021)
17. Yan, C., Pang, G., Wang, L., Jiao, J., Feng, X., Shen, C., Li, J.: Bv-person: a large-scale dataset for bird-view person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10943–10952 (2021)

18. Wu, D., Wang, C., Wu, Y., Wang, Q.-C., Huang, D.-S.: Attention deep model with multi-scale deep supervision for person re-identification. IEEE Trans. Emerg. Top. Comput. Intell. **5**(1), 70–78 (2021)
19. Zhao, S., Gao, C., Zhang, J., Cheng, H., Han, C., Jiang, X., Guo, X., Zheng, W.-S., Sang, N., Sun, X.: Do not disturb me: Person re-identification under the interference of other pedestrians. In: European Conference on Computer Vision, pp. 647–663. Springer (2020)
20. Chen, G., Gu, T., Lu, J., Bao, J.-A., Zhou, J.: Person re-identification via attention pyramid. IEEE Trans. Image Process. **30**, 7663–7676 (2021)
21. Gong, Y., Wang, L., Li, Y., Du, A.: A discriminative person re-identification model with global–local attention and adaptive weighted rank list loss. IEEE Access **8**, 203700–203711 (2020)
22. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
23. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
24. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
25. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)
26. Lian, S., Jiang, W., Hu, H.: Attention-aligned network for person re-identification. IEEE Trans. Circuits Syst. Video Technol. **31**(8), 3140–3153 (2020)
27. Zhang, Z., Xie, Y., Li, D., Zhang, W., Tian, Q.: Learning to align via Wasserstein for person re-identification. IEEE Trans. Image Process. **29**, 7104–7116 (2020)
28. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **44**(9), 5056–5069 (2021)
29. Yang, J., Zhang, J., Yu, F., Jiang, X., Zhang, M., Sun, X., Chen, Y.-C., Zheng, W.-S.: Learning to know where to see: a visibility-aware approach for occluded person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11885–11894 (2021)
30. Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1613–1623 (2023)
31. Zhang, Z., Zhang, H., Liu, S., Xie, Y., Durrani, T.S.: Part-guided graph convolution networks for person re-identification. Pattern Recognit. **120**, 108155 (2021)
32. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3186–3195 (2020)
33. Liao, S., Shao, L.: Graph sampling based deep metric learning for generalizable person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7359–7368 (2022)
34. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539 (2020). https://doi.org/10.1109/CVPR42600.2020.01155
35. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: occluded person re-identification with part-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2898–2907 (2021)
36. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15013–15022 (2021)
37. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
38. Lai, S., Chai, Z., Wei, X.: Transformer meets part model: adaptive part division for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4150–4157 (2021)
39. Pervaiz, N., Fraz, M., Shahzad, M.: Per-former: rethinking person re-identification using transformer augmented with self-attention and contextual mapping. Vis. Comput. 1–16 (2022)
40. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
41. Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: The relation between the roc curve and the cmc. In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), pp. 15–20. IEEE (2005)
42. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 2872–2893 (2021)
43. He, L., Liu, W.: Guided saliency feature learning for person re-identification in crowded scenes. In: European Conference on Computer Vision, pp. 357–373. Springer (2020)
44. Liu, Y., Zhou, W., Liu, J., Qi, G.-J., Tian, Q., Li, H.: An end-to-end foreground-aware network for person re-identification. IEEE Trans. Image Process. **30**, 2060–2071 (2021)
45. Zhang, A., Gao, Y., Niu, Y., Liu, W., Zhou, Y.: Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 598–607 (2021)
46. Chen, X., Zheng, X., Lu, X.: Bidirectional interaction network for person re-identification. IEEE Trans. Image Process. **30**, 1935–1948 (2021)
47. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1060–1069 (2022)
48. Wang, Z., Zhu, F., Tang, S., Zhao, R., He, L., Song, J.: Feature erasing and diffusion network for occluded person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4754–4763 (2022)
49. Li, S., Song, W., Fang, Z., Shi, J., Hao, A., Zhao, Q., Qin, H.: Long-short temporal–spatial clues excited network for robust person re-identification. Int. J. Comput. Vis. **128**, 2936–2961 (2020)
50. Martinel, N., Foresti, G.L., Micheloni, C.: Deep pyramidal pooling with attention for person re-identification. IEEE Trans. Image Process. **29**, 7306–7316 (2020)
51. Pu, N., Chen, W., Liu, Y., Bakker, E.M., Lew, M.S.: Lifelong person re-identification via adaptive knowledge accumulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7901–7910 (2021)

52. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

Qian Luo received the Ph.D. degree from Sichuan University, China, in 2012. He is a master's and Ph.D. supervisor. He joined Civil Aviation Chengdu Electronic Technology Co., Ltd., China, where he held the position of vice president. His research interests include big data technology, artificial intelligence, accurate perception of airport operation situation, accurate characterization of flight support network evolution mechanism, and fine collaborative decision-making of multiple business entities.

Jie Shao received the bachelor's degree from Chengdu College of University of Electronic Science and Technology of China, in 2020. He is currently pursuing the master's degree with the School of Computer and Software Engineering, Xihua University, China. His research areas are machine learning, person re-identification and image processing.

Wanli Dang is currently working toward the Ph.D. degree in electronics and information technology with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. She is currently an Engineer with The Second Research Institute, Civil Aviation Administration of China, Beijing, China. Her research interests include human action recognition for airport, pedestrian detection, and tracking.

Long Geng received the master's degree from Xihua University, China, in 2017. He joined in the Second Research Institute, Civil Aviation Administration of China, where he is an Engineer. His current research interests include computer vision, video image processing, object detection, target tracking and airport operation and control.

Huaiyu Zheng serves as an Assistant Researcher at The Second Research Institute of the Civil Aviation Administration of China (CAAC). He attained his Master's degree from the College of Engineering at Northeastern University. His scholarly pursuits are entrenched within the file of computer vision, encompassing a broad spectrum from the multiple object tracking to the domain of action prediction.

Chang Liu received the M.S. degree in Management science and engineering from SiChuan University, Chengdu, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Computer Science, Sichuan University. He joined in the Second Research Institute, Civil Aviation Administration of China, where he is an Engineer. His major research interests include machine learning and airport operation.