

Unsupervised person re-identification based on removal of camera bias and dynamic updating of the memory bank

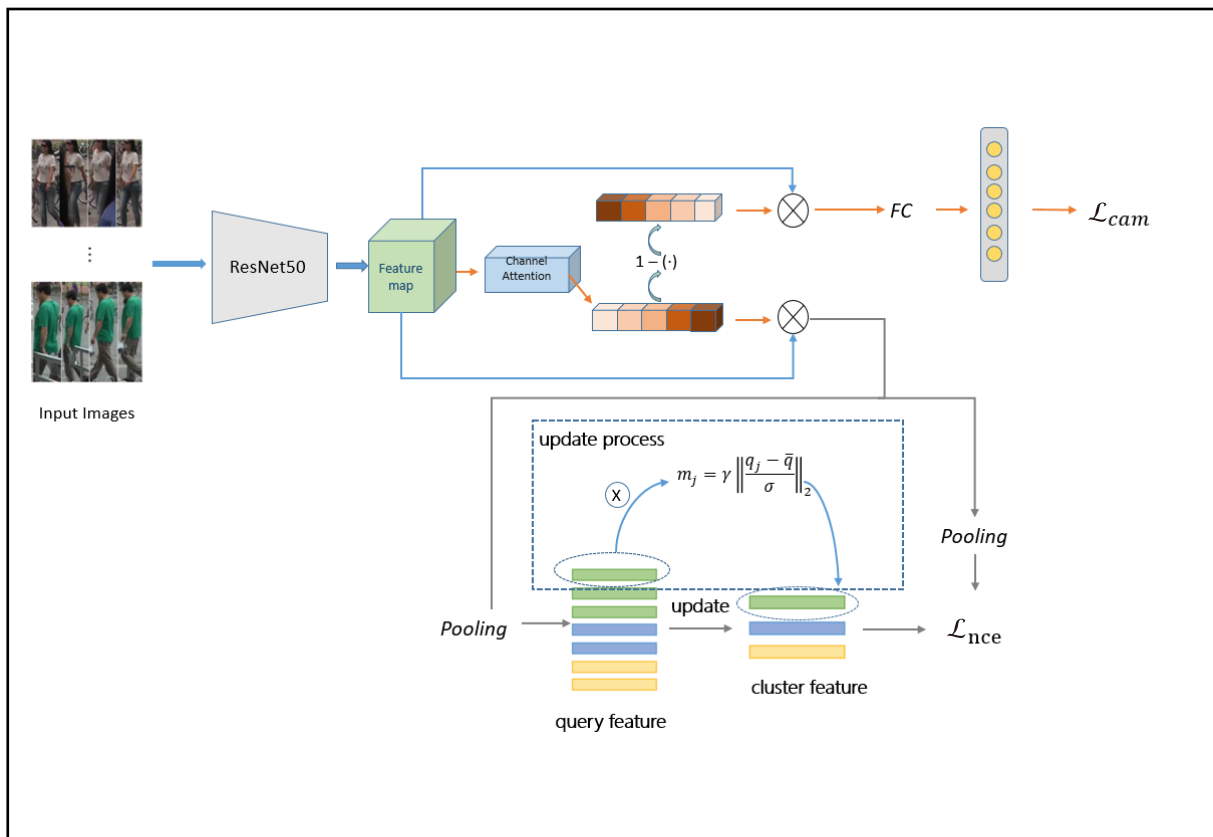
Jun Zhang, and Xinmei Tian ✉

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China

✉ Correspondence: Xinmei Tian, E-mail: xinmei@ustc.edu.cn

© 2022 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Graphical abstract



This paper improves unsupervised person reidentification by removing the camera bias and dynamically updating the memory model.

Public summary

- We design a channel attention module to separate camera-related features from global ones and obtain intra-class shared features.
- We propose a mechanism to dynamically update the memory dictionary according to the distance between the instance and cluster features.
- Extensive experiments on mainstream datasets Market1501 and DukeMTMC-Re-ID prove that our method outperforms state-of-the-art methods.

Unsupervised person re-identification based on removal of camera bias and dynamic updating of the memory bank

Jun Zhang, and Xinmei Tian 

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China

 Correspondence: Xinmei Tian, E-mail: xinmei@ustc.edu.cn

© 2022 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Cite This: *JUSTC*, 2022, 52(12): 7 (9pp)



Read Online

Abstract: In recent years, unsupervised person reidentification technology has made great strides. The technology retrieves images of interested persons under different cameras from massive repositories of unlabeled images. However, in the current research, there are some existing problems, such as the influence of pedestrians appearing across cameras and pseudo-label noise. To solve these problems, we conduct research in two ways: removing the camera bias and dynamically updating the memory model. In removing the camera bias, based on a learnable channel attention module, the features that are only related to cameras can be extracted from the feature map, thereby removing the camera bias in the global features and obtaining the features that can represent the pedestrians. In regards to dynamically updating the memory model, since the instance features do not necessarily belong to the identity represented by the pseudo-label, we adopt a method to update the memory dynamically according to the distance between the instance features and the category features so that the category features tend to be true. We combine the removal of the camera bias and the dynamic updating of the memory model to better solve problems in this field. Extensive experimentation demonstrates the superiority of our method over the state-of-the-art approaches on fully unsupervised Re-ID tasks.

Keywords: person reidentification; unsupervised learning; camera bias; dynamic update

CLC number: TP391

Document code: A

1 Introduction

Currently, with the vigorous development of the Internet of Things and advancements in the present data-saturated era, video surveillance technology has reached an intelligence-based surveillance stage. Person re-identification (Re-ID) has been a hot research topic in the field of intelligent surveillance video analysis in recent years. Re-ID identifies and locates specific pedestrians in surveillance videos, which is very useful for criminal investigations, search and rescue, etc. It is the main task in many surveillance and security applications. Therefore, person re-identification has recently attracted increased attention from academia and industry. With the development of deep learning, current supervised person re-identification methods have achieved notable performances on several benchmarks, even surpassing the discerning abilities of human eyes. However, the learning in supervised scenarios is not conducive to the generalization of the person re-identification model to other scenarios. With the increasing demands for intelligent monitoring and the huge cost of labeling, researchers have begun to focus attention on unsupervised person re-identification^[1-4].

Currently, there are two experimental settings for unsupervised person Re-ID, which differ in whether or not an extra labeled dataset is used. It is generally believed that it is difficult to learn robust pedestrian features that are invariant to cameras without the corresponding identity labels. Therefore,

most works pretrain a model on a labeled source dataset and then adapt it to an unlabeled target dataset, which is called unsupervised domain adaptation (UDA) person re-identification. Another unsupervised person Re-ID method directly addresses unlabeled datasets where no additional labels are used. This subject is called fully unsupervised person re-identification (USL), which is also the main research focus of this paper.

In the current work, unsupervised domain adaptation person re-identification mainly utilizes image datasets. These studies^[1, 2, 5, 6] can be roughly divided into three methods: feature alignment methods, domain transformation methods and pseudo-label-based methods. However, the performance of UDA method is affected by the source dataset, and it is not practical to pull the features of the source dataset and the target domain dataset closer.

In contrast, fully unsupervised methods^[3, 4] do not use any labeled data. For these methods, the first step is to generate pseudo-labels for the unlabeled images, which is also the current mainstream technical approach. This kind of method is simple and clear and has good performance. In particular, some pseudo-label generation methods based on clustering can achieve performances that are close to those of supervised learning methods.

However, there are two main drawbacks in such USL methods. First, in the person re-identification dataset, a pedestrian is captured by different cameras. Due to the differences

in camera equipment, the shooting environment and lighting conditions will be different. These factors are called the camera bias. The images of the same person can differ greatly under different cameras, resulting in model recognition errors. As a remedy to the appearance difference technical problems across cameras in unsupervised person re-identifications, the previous methods^[7,8] either roughly corrected the distance between the cross-camera images or adopted an indirect method of training on the datasets divided by the cameras^[3,4,9]. This was done without bothering with the matching of the person features across the cameras after removing the camera bias. The positive effect of utilizing camera labels to generate more accurate image pseudo-labels is underestimated. Second, although these memory-based contrastive learning methods achieve good performance, there is still a large performance gap between the USL and the supervised methods. To bridge the gap, we reinvestigate the USL pipeline and argue that the memory dictionary storage mechanism, including the initialization, update and loss computation, is crucial for model optimization. According to previous methods^[2-4], the cluster feature stored in the memory dictionary is updated by the instance feature with a constant momentum. However, the instance feature may not actually belong to the identity represented by the pseudo-label since these pseudo-labels are obtained by clustering. Due to the insufficient expressive abilities of feature representations and the imperfect clustering results, dealing with an unknown proportion of false pseudo-labels is an unavoidable problem. They will adversely affect the feature learning when the cluster center is used as the positive anchor of the query image. If the instance is a false-positive instance for its pseudo label, it may update the cluster feature in the wrong way, which could degrade the model performance.

To alleviate these problems, we investigate the issues from two perspectives to facilitate the learning of the discriminative person feature representations. (i) This paper proposes a method to directly remove the camera bias from images. Based on a learnable channel attention module, the camera-related features can be extracted from the feature map to remove the camera bias in the global features. After removing the camera bias, the person features are transported to the next training process. The model can better match those cross-camera person images. Finally, we obtain the robust features that can represent the pedestrians. (ii) We propose a new mechanism to update the cluster feature in the memory. We assign different weights to the instances based on their distances from the cluster centroids. Then, we utilize the instance features to update the corresponding cluster features with their unique weights, which makes the cluster feature move toward the real feature space. In summary, the experimental results show that the two methods proposed in this paper obtain a better re-identification performance when compared to the unsupervised Re-ID state-of-the-art methods^[10,11].

2 Related work

In this section, recent works related to ours are briefly reviewed from the following perspectives: (i) deep unsupervised person Re-ID, which includes both fully unsupervised

person Re-ID and unsupervised domain adaptation person Re-ID; (ii) research on camera bias in USL; and (iii) memory dictionary-based deep learning.

Deep unsupervised person Re-ID methods can be divided into fully USL person Re-ID and UDA Re-ID. Generally, effective fully USL person Re-ID methods are clustering-based methods. These methods generate hard/soft pseudo-labels and then train the deep models based on the pseudo-labels. The pseudo-labels can be obtained by clustering the sample features or measuring the similarities among the instance features. The representatives of such pseudo-label-based methods include Refs. [2, 12, 13], which achieve good performances. In BUC^[14], a bottom-up clustering method aggregates similar samples and finds relationships between the identities or within an identity. HCT^[15] generates pseudo-labels by hierarchical clustering and selects samples for triplet loss computation. Based on contrastive learning, SPCL^[2] builds a memory to compare samples with other samples in the dataset, shortening the distance between images of the same pseudo-label and pushing away features of different labels. Because the proposed unified contrast loss function will make the samples tend to the cluster center and stay away from other cluster centers, it can achieve a good intraclass aggregation effect and a uniform distribution between the classes. However, such pseudo-label-based methods heavily depend on the quality of the estimated pseudo-labels, since noisy labels could degrade the model performance. To improve the quality of the estimated pseudo-labels and mitigate the negative effects of incorrect labels, many improved pseudo-label estimation methods have been proposed. SSG^[13] adopts human local features to assign multiscale pseudo-labels. PAST^[16] utilizes multiple regularizations to overcome this problem, and Ref. [5] proposes some reliability evaluation criteria to further modify the generated pseudo-labels. In addition to estimating the hard pseudo-labels, some researchers propose soft-label and multilabel-based USL methods for person Re-ID. Soft pseudo-labels are assigned to samples for training by mining the K-nearest neighbors of each training sample in SSL^[7]. Yu et al.^[17] proposed a soft-label-based method by measuring the similarities between person images and reference images. Wang et al.^[18] treated the USL person Re-ID as a multilabel classification problem. They gradually found the true labels with the help of some label estimation strategies and consistency regularization, which can improve the precision of the estimated pseudo-labels. Recently, some methods have introduced mutual learning among two/three collaborative networks to mutually exploit refined soft pseudo-labels with peer networks as supervision^[5]. The abovementioned two types of pseudo-label-based methods are widely used in both fully USL and UDA person Re-ID.

For the UDA person Re-ID, some domain translation methods are first proposed because the source domain labeled data can be used in this task. The representative works include the style-GAN-based^[19] translation methods, which transform the images from the source domain to match the image styles in the target domain while keeping the person identities unchanged. These methods transfer the purely USL task to the semi-supervised task, and then the previously mentioned purely USL methods can be fully used. In addition, some oth-

er methods^[20, 21] exploit the valuable information across the source and target domains, then explore some of the underlying relationships among them and finally construct the joint learning framework by using the training data in both the source and target domains.

In person re-identification dataset, a pedestrian is captured by different cameras. Due to the difference in camera equipment, the shooting environment and lighting conditions will be different. Therefore, the images of the same identity will have differences under different cameras. The appearance of different pedestrians may be more similar than the appearance of the same person. In the field of person re-identification, some researchers have explored removing camera bias. SSL^[7] introduced a regular term when calculating the distance between two features. If the cameras of the two images are the same, this method increases the dissimilarity between the two images. The CIDC^[8] method converts images from different cameras to images of the same camera style via StarGAN^[22]. Other methods divide the dataset into different data subsets according to the camera ID. They trained the model on each data subset separately and on the entire dataset. This kind of method avoids the bias caused by the camera through training on a subset of data. In this way, it is difficult to misjudge the person images of different cameras to be matched.

IICS^[9] uses the images in each camera to train the corresponding classifier and requires that the predicted probability of the same image obtained on different classifiers be consistent. CAP^[4] learns from images under different cameras separately to ensure that the model is not easily affected by hard negative samples. MetaCam^[23] introduces meta learning to enhance the model's ability to adapt to camera bias. However, they did not remove the camera bias from the image itself, and the effect of using camera labels to generate more accurate pseudo-labels is underestimated. Different from the afore-

mentioned methods, our method separates the camera-related features and camera-unrelated features in an explicit way.

The memory dictionary-based deep learning methods present promising results on unsupervised visual representation tasks. In addition, state-of-the-art unsupervised person Re-ID methods also build memory dictionaries for contrastive learning. Many strategies have been proposed to consistently update the memory dictionary, which makes deep metric learning very effective, especially for USL/UDA person Re-ID. Ge et al.^[2] proposed self-paced contrastive learning for UDA person Re-ID tasks. Zheng et al.^[6] also proposed exploiting the sample uncertainty for UDA person Re-ID task while affiliating with memory-based contrastive learning. These methods can leverage the memory bank to measure the similarity between a sample and the instances stored in the memory, which helps to mine the hard negative examples across batches, increases the contrastive power with more negatives, and finally, better trains the model. However, they neglect the unfavorable effect the false-positive instance brings to its corresponding pseudo-label class. We propose a new mechanism to update the cluster feature. Different weights are assigned to the instances on the basis of their distances from the cluster centers. In this way, the cluster feature is likely to be updated in a correct way and is less affected by the false-positive instances. Finally, we can better train the model and achieve superior performance on the unsupervised person Re-ID task.

3 Method

3.1 Clusterwise contrastive learning

As shown in Fig. 1, we propose a different framework for unsupervised person Re-ID. The overall process of our method mainly includes three stages: feature extraction, two-branch

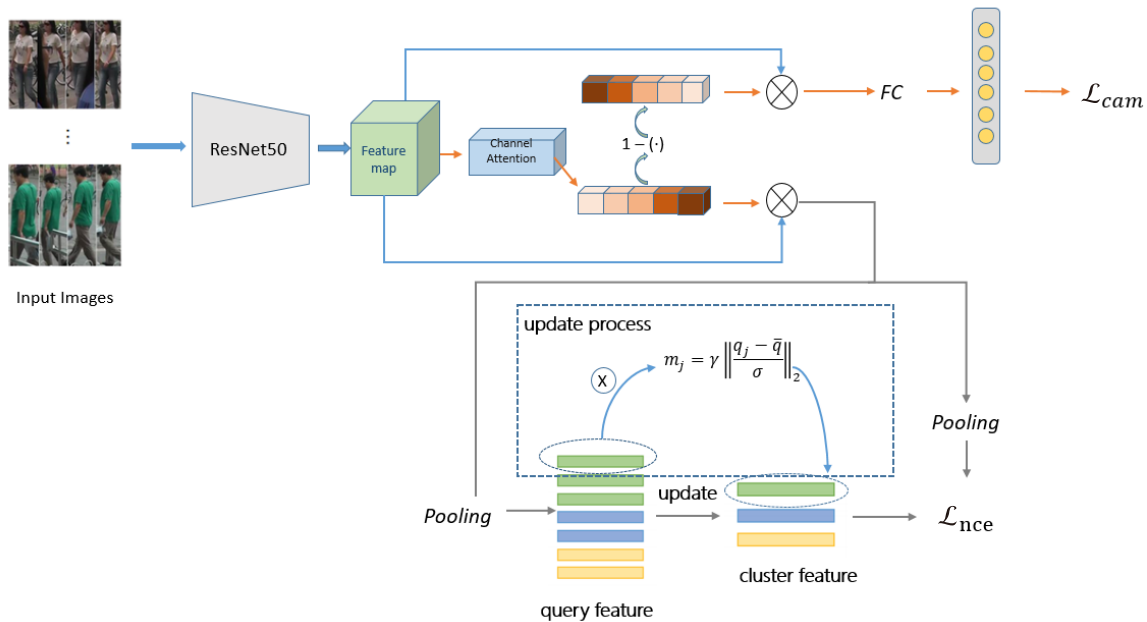


Fig. 1. Illustration of our pipeline. We first apply the channel attention module after the backbone network to explicitly separate the camera-related information from the feature maps. Then, we utilize a new mechanism of dynamically updating the memory dictionary according to the distance between the instance feature and the cluster feature.

camera bias removal, and contrastive learning.

In the feature extraction stage, we use ResNet50 to extract the feature vectors of all the images in the dataset $X = \{x_1, x_2, \dots, x_N\}$, where N is the total number of images. Specifically, for each image x_i , the corresponding feature map $F_{x_i} \in \mathbb{R}^{H \times W \times C}$ is extracted through the backbone network, where H , W and C represent the height, width and channel number of the feature map, respectively. The feature is global. In the stage of camera bias removal, we employ the feature to calculate a learnable channel attention vector. This attention module is also pretrained on ImageNet and has the ability to initially strengthen the feature channel related to the object categories and use this module to extract the features only related to persons. Then, it is multiplied by the feature vector F in the corresponding channel dimension to obtain the feature part only related to pedestrians. The complementary part of the attention module is multiplied by the feature vector F in the corresponding channel dimension to obtain the feature part only related to cameras. The camera-related feature is constrained by the proposed camera prediction loss $\mathcal{L}_{cam}$ that will drive the camera-related attention module to pay more attention to the camera features of the image so that the camera-unrelated channel attention vector can better extract the person's features and remove the camera bias. Person features can be assigned pseudo-labels by clustering and entering the next training step.

At the beginning of each epoch, the pseudo-labels $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$ of the images are obtained by DBSCAN clustering of all instance features in the dataset. Furthermore, we select reliable clusters by leaving out isolated instances. In this way, we obtain a labeled dataset. Generally, the pseudo-label is obtained from an initial pretrained model. Since the pretrained model has relatively good discerning abilities and the images of the same identity have similarities in appearance, the initial pseudo-label has a certain value. Some images are actually correctly labeled, and inevitably, some are incorrectly labeled. On the basis of the pseudo-labels, our method can obtain the common InfoNCE loss^[24] $\mathcal{L}_{nce}$ using the following equation:

$$\mathcal{L}_{nce} = -\log \frac{\exp(f_i \cdot \phi_k / \epsilon)}{\sum_{k=0}^K \exp(f_i \cdot \phi_k / \epsilon)}, \quad (1)$$

where ϕ_k is the unique representation vector of the k -th cluster and ϵ is a temperature factor. This loss brings the features closer to their corresponding cluster features and keeps them away from the other cluster features. By mining the relation-

ships between classes and within a class, the features that can represent the identity are gradually learned. The extracted instance features are generally used to update the corresponding cluster features in the memory so that the class features are consistent with the model's capabilities. Let $\mathcal{M}$ ($\mathcal{M} \in \mathbb{R}^{K \times d}$) represent the memory bank, where d is the feature dimension. After each iteration, extracted instance features f_i will update the cluster feature in the memory bank $\mathcal{M}$ with the momentum mechanism.

During the training stage, the model loss consists of two parts:

$$\mathcal{L}_{total} = \mathcal{L}_{nce} + \lambda \cdot \mathcal{L}_{cam}, \quad (2)$$

where λ is a coefficient to balance two components. In the test stage, the extracted embedding f_i is used for final matching.

3.2 Camera bias removal based on channel attention module

Due to differences in the camera equipment, the camera bias will lead to differences in shooting environments, lighting conditions, etc. Consequently, images of the same identity under different cameras have a large dissimilarity. To solve this problem, we utilize the channel attention (CA) module to remove the camera bias directly from the extracted feature maps. Enlightened by how DAAM^[1] detaches the domain-related information in UDA, we propose a new way to separate the camera-related information in our USL pipeline. There is no person identity information under the USL settings, but the camera ID information can be exploited. As shown in Fig. 1, the attention module has two branches, one path to extract the camera-unrelated features, and the complementary module to extract the camera-related features. Specifically, for each image x_i , the corresponding feature map $F_{x_i} \in \mathbb{R}^{H \times W \times C}$ is extracted through the backbone network. This feature is global. In the camera bias removal stage, this feature is used to calculate a learnable channel attention vector $a = [a_1, a_2, \dots, a_c] \in \mathbb{R}^c$, which can be obtained by the formula

$$a = g(F) = \sigma(W_2 \delta(W_1 \text{pool}(F))). \quad (3)$$

As shown in Fig. 2, in the channel attention module, the feature F goes through the global average pooling layer, the fully connected layer, the ReLU activation layer, the fully connected layer and the sigmoid function, after which we can obtain the channel attention vector. This attention module is also pretrained on ImageNet and has the ability to initially

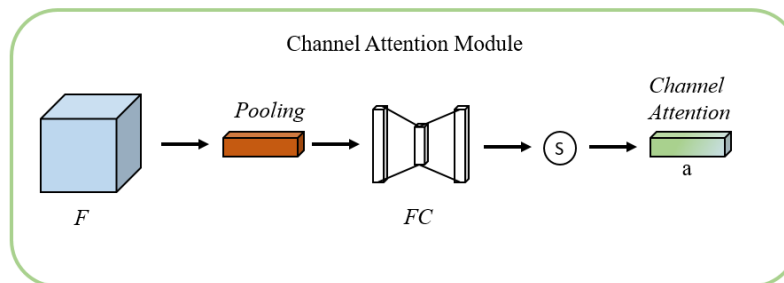


Fig. 2. The CA module.

make the channel features related to the object categories. We use this module to extract the camera-unrelated features. Then, it is multiplied by the feature map F in the corresponding channel dimension to obtain the camera-unrelated feature $F_u \in \mathbb{R}^{H \times W \times C}$. At the same time, $1 - a$ indicates an attention vector that is only concerned about the camera bias. F_r represents the part related to the camera style after removing the person feature.

$$\begin{aligned} F_u(:, :, k) &= a_k \cdot F(:, :, k), \\ F_r(:, :, k) &= (1 - a_k) \cdot F(:, :, k), \end{aligned} \quad (4)$$

where $F(:, :, k)$ is the k -th channel of the feature map $F \in \mathbb{R}^{H \times W \times C}$ and a_k is the k -th element of $\mathbf{a}$, $k = 1, 2, \dots, C$. Then, the probability of predicting the camera ID is obtained through the pooling, fully connected layer and the softmax layer.

$$P = \text{Softmax}(W \cdot (\text{GAP}(F_r)) + b). \quad (5)$$

Then, we obtain the camera prediction loss $\mathcal{L}_{\text{cam}}$:

$$\mathcal{L}_{\text{cam}} = \mathbb{E}[-y^{\text{cam}} \log P], \quad (6)$$

where y^{cam} is the camera label. The camera prediction loss will drive $1 - \mathbf{a}$ to pay more attention to the camera features of the image so that the channel attention vector $\mathbf{a}$ can better extract the person features and remove the camera bias. F_u represents the person feature to obtain the pseudo-labels by clustering, thereby entering the next training step.

In this way, we separate the camera-unrelated feature F_u and camera-related feature F_r directly from the global feature map. By optimizing $\mathcal{L}_{\text{cam}}$, the negative effect of the camera bias is reduced, and the model can extract more robust person features.

3.3 Update the memory bank with dynamic momentum

At the beginning of each epoch, the pseudo-labels are obtained by DBSCAN clustering all of the instance features in the dataset. Furthermore, we select reliable clusters by leaving out isolated instances. We suppose $\{C_1, C_2, \dots, C_K\}$ is the cluster feature corresponding to each cluster, where K is the total number of clusters. These cluster features are stored in the memory and updated by the instance features.

Due to the imperfect results of the clustering algorithm, the cluster representation inevitably incorporates information from different identities. These mislabeled instances will adversely affect the feature learning when the cluster center is used as the positive anchor of these instances. If the instance is a false-positive instance for its pseudo-label, it may update the cluster feature in the wrong way, which could degrade the model's performance.

To alleviate the negative effect of false-positive instances, we propose a new mechanism to update the cluster feature. We adopt a method to dynamically update the memory bank according to the distance between the instance and cluster features. First, we calculate the mean and variance of the instance features in a batch that belongs to the same pseudo-label. The mean represents the cluster feature after updating; then, we obtain the distance between the instance and the corresponding cluster feature. The distances are partly due to intraclass variances. However, some instance features do not

belong to this identity, so these samples are much farther from their corresponding cluster features than the other instance features. Accordingly, we decrease the importance of such instances so that it has a lower impact on the update of the cluster features, while the influence of others is increased, which will prevent the cluster features from being misled by noisy examples and make the cluster features tend toward the real feature space. The method analyzes a series of features $\{q_1, q_2, \dots, q_B\}$ belonging to a cluster. We assume the corresponding cluster feature is C_i . B is the total number of samples in the batch under this pseudo-label. After finding the mean $\bar{q}$ and variance σ of the series, the momentum of feature q_j can be updated by:

$$m_j = \gamma \left\| \frac{q_j - \bar{q}}{\sigma} \right\|_2. \quad (7)$$

Each instance is assigned its own update momentum. The update factor is variable. γ is the scaling factor which is empirically set to 0.3. Therefore, the update factor is within a reasonable range. The update momentum is obtained and then the cluster features are updated:

$$C_i = m_j \cdot C_i + (1 - m_j) \cdot q_j, \quad (8)$$

where C_i is the cluster feature in the memory. Based on the distance between each instance feature and the cluster feature, each instance is assigned a different importance weight to update the cluster feature in the memory dictionary. This method allows us to utilize different instances discriminately and makes the cluster features more inclined toward the real feature space. Consequently, the optimization of the model is ensured in the right direction.

4 Experiments

4.1 Datasets and implementation details

4.1.1 Datasets

Datasets. We evaluate our methods on three commonly used person Re-ID datasets. Market1501^[10] has a total of 1501 person identities and 32668 images. There are 12936 images in the training set, 751 pedestrians and 19732 images in the test set. The images are taken with six cameras. DukeMTMC-ReID^[11] is a large-scale dataset captured from 8 cameras, containing 36411 images with 1404 identities. It is divided into 16522 images of 702 identities for training, and 19889 images of 702 identities for testing. MSMT17^[19] has 4101 identities and 126411 images in total. There are 30248 images of 1041 identities in the training set, while the remaining 3060 identities are used for testing. The images are taken by 15 cameras. By convention, the cumulative matching characteristic (CMC) and mean average precision (mAP) are used to indicate the performance.

4.1.2 Implementation details

Implementation details. ResNet-50, pretrained on ImageNet, is used as the backbone for our encoder. During training, the input image is resized to 256×128 , and various data augmentations are applied, including random flipping, random

cropping, and random erasing. At the start of each epoch, DBSCAN clustering is used to generate pseudo-labels. For Market1501/DukeMTMC-Re-ID/MSMT17, the ϵ in DBSCAN is set to 0.45/0.5/0.4. The scale factor γ is 0.3/0.2/0.3 on Market1501/DukeMTMC-Re-ID/MSMT17. In $\mathcal{L}_{ncc}$, the temperature coefficient τ_{ncc} is fixed to 0.07, and epochs is set to 80. The λ in $\mathcal{L}_{total}$ is empirically set to 0.02. The batch size is set to 256 with the Adam optimizer. The initial learning rate is set to 0.00035 and is decayed by 1/10 every 20 epochs.

4.2 Comparison with the state-of-the-art

In this section, we compare our method with state-of-the-art unsupervised Re-ID methods. The experimental results are shown in Table 1.

As shown in Table 1, our method achieves better performance than the previous methods under the fully unsupervised setting. The introduction of the channel attention module and the camera loss function proposed in this paper can more directly remove the camera bias from the image, and the extracted pedestrian's features are more robust. In this paper, after removing the camera bias, the extracted person's features can better explore the intraclass and interclass relations without the influence of the camera. Methods, such as CAP^[4] and ICE^[3] are directly trained on the datasets and divided by the cameras. This method avoids camera bias by training on datasets divided by the cameras. Consequently, it is difficult to judge the person images of the different cameras that are to be matched. However, they are not direct and do not bother with the matching of the person's features across cameras after re-

moving the camera bias. The method in this paper directly separates the camera-related features and camera-unrelated features from the feature map and then trains the person's features after removing the camera bias, which can better match the cross-camera pedestrian images. This is also the reason the mAP of the method in this paper is higher than other methods on Market1501. At this time, it better matches those cross-camera images with a lower retrieval rank. The reason rank1 is lower than ICE is that ICE is trained on a single camera dataset, which reduces the adverse effects of those pedestrian images judged to match due to the camera bias, and the rank-1 results are more reliable.

4.3 Ablation study

In this section, we specifically analyze and discuss the effectiveness of key components, including the channel attention module and the dynamic updating of cluster features.

The baseline (Index-1) is the pipeline without our proposed CA module and dynamic update (d-up) method. As shown in Table 2, compared with the baseline, the CA module yields a general improvement on the datasets. mAP/Rank1 is increased by 2.4%, 1.1% on Market1501, and 0.4%, 0.8% on DukeMTMC-Re-ID. These results demonstrate the effectiveness of the CA module in removing the camera bias. The improvements of d-up (Index-2) over baseline are 1.1%, 0.2% on Market and 0.8%, 0.3% on Duke. In addition, d-up is complementary to CA. The best performance can be obtained when combining both CA and d-up (Index-4), with improvements of 2.7%, 1.1% on Market and 0.9%, 0.8% on Duke, re-

Table 1. Performance comparison with recent methods.

Method	Market1501		DukeMTMC-Re-ID		MSMT17	
	mAP (%)	Rank1 (%)	mAP (%)	Rank1 (%)	mAP (%)	Rank1 (%)
BUC ^[14]	38.3	66.2	27.5	47.4	—	—
SSL ^[7]	37.8	71.7	28.6	52.5	—	—
MMCL ^[18]	45.5	80.3	40.2	65.2	11.2	35.4
HCT ^[15]	56.4	80.0	50.7	69.6	—	—
SpCL ^[2]	73.1	88.1	65.3	81.2	19.1	42.3
IICS ^[9]	72.1	88.8	59.1	76.9	18.6	45.7
OPLG ^[25]	78.1	91.1	65.6	79.8	28.4	54.9
CAP ^[4]	79.2	91.4	67.3	81.1	36.9	67.4
MCRN ^[26]	80.8	92.5	69.9	83.5	31.2	63.6
MGH ^[27]	81.7	93.2	70.2	83.7	—	—
ICE ^[3]	82.3	93.8	69.9	83.3	38.9	70.2
Ours	84.3	93.3	73.1	85.2	40.1	72.2

Table 2. Ablation studies on individual components.

Index	Method	Market1501		DukeMTMC-Re-ID	
		mAP (%)	Rank1 (%)	mAP (%)	Rank1 (%)
1	Ours-d-up- $\mathcal{L}_{cam}$	82.1	92.3	72.6	84.9
2	Ours- $\mathcal{L}_{cam}$	83.0	92.5	72.8	85.2
3	Ours-d-up	84.1	93.2	72.9	85.4
4	Ours	84.3	93.3	73.2	85.6

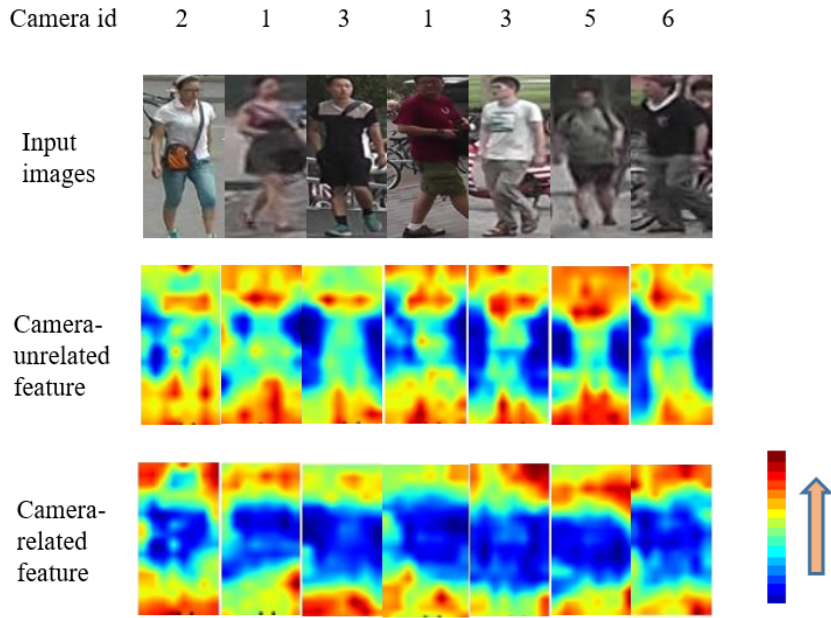


Fig. 3. Visualization of the attention map in the CA module.

spectively. This shows that CA and d-up can improve the model together.

To show the focus of CA module and its complementary part, we compare the attention maps of camera-related and camera-unrelated features. Since the location and angle of the camera are fixed, each camera has its own background. The background of the image is an important aspect of identifying the camera ID. As shown in Fig. 3, the first line is the ID of the camera that takes the image. The third row is a visualization of the camera-unrelated features. Red indicates the part with strong attention, and blue indicates the part with weak attention. The camera-unrelated feature map attaches more importance to the person and pays more attention to the middle part of the image, ignoring the influence of the camera bias. The fourth row is a visualization of camera-related features. We can see that the camera-related feature map emphasizes environmental factors, which are mainly distributed on both sides of the pedestrians and the upper and lower ends of the images.

To observe the dynamic update function on the memory,

we visualize the feature representations of several classes. As shown in Fig. 4, there are some false-positive examples after the clustering process. If a previous update strategy on the memory is used, the cluster feature is inevitably affected by noisy examples. Our method can gradually remove the negative effect of noisy examples, making the cluster feature more accurate. The instances inside a cluster are also more compact.

Here, we further explore the effect of different batch sizes on network performance. The model is trained with different batch sizes from 32 to 256. From the experimental results shown in Fig. 5, it can be seen that larger batches will achieve better results. When a larger batch is used, each identity contains more instance samples, and updating the cluster center in the memory with these samples will make the cluster center more balanced and robust. Moreover, the instance features in the same batch will be regularized. The more samples there are, the better it is to remove the task-independent biases and focus on the task-related features.

In general, by using two components together, our method can extract more robust person features under the USL settings.

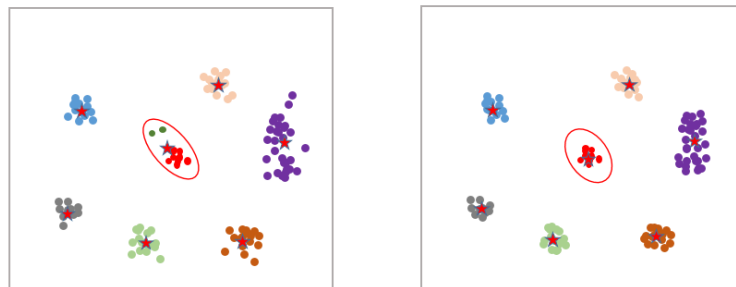


Fig. 4. T-SNE visualization of 7 classes in the Market1501 test set between our method with d-up removed (left) and our method (right). We utilize a new mechanism of dynamically updating the memory dictionary according to the distance between the instance and cluster feature.

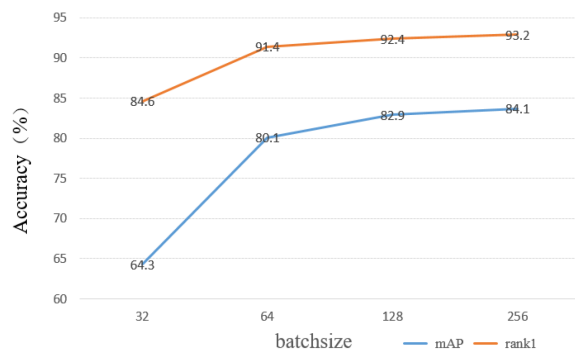


Fig. 5. The effect of batch size.

5 Conclusions

In this paper, we propose a channel attention module to distill the camera-unrelated features and a dynamic momentum update mechanism for the fully unsupervised person Re-ID tasks. Our mechanisms effectively eliminate the camera bias from the feature map through our proposed channel attention module and camera prediction loss. Moreover, we assign different weights to samples based on their distances from the class centers to obtain the cluster features that are closer to the real feature space. The experimental results demonstrate that our method shows superior performance against the existing state-of-the-art methods.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (WK3490000005).

Conflict of interest

The authors declare that they have no conflicts of interest.

Biographies

Jun Zhang received the B.E. degree from the University of Science and Technology of China (USTC) in 2019. He is currently pursuing a master's degree at the College of Information Engineering in USTC. His main research interests include deep learning and computer vision.

Xinmei Tian is an Associate Professor in the CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application Systems, University of Science and Technology of China (USTC). She received her B.E. degree and Ph.D. degree from USTC in 2005 and 2010, respectively. She received the Excellent Doctoral Dissertation of Chinese Academy of Sciences award in 2012 and the Nomination of National Excellent Doctoral Dissertation award in 2013. Her current research interests include multimedia information retrieval and machine learning.

References

- [1] Huang Y, Peng P, Jin Y, et al. Domain adaptive attention learning for unsupervised person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34: 11069–11076.
- [2] Ge Y, Zhu F, Chen D, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID. In: NIPS'20: 34th International Conference on Neural Information Processing Systems. BC, Canada: Curran Associates Inc, 2020: 11309–11321.

- [3] Chen H, Lagadec B, Brémond F. ICE: Inter-instance contrastive encoding for unsupervised person re-identification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021: 14940–14949.
- [4] Wang M, Lai B, Huang J, et al. Camera-aware proxies for unsupervised person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35 (4): 2764–2772.
- [5] Ge Y, Chen D, Li H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. 2020. <https://doi.org/10.48550/arXiv.2001.01526>. Accessed December 3, 2021.
- [6] Zheng K, Lan C, Zeng W, et al. Exploiting sample uncertainty for domain adaptive person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35 (4): 3538–3546.
- [7] Lin Y, Xie L, Wu Y, et al. Unsupervised person re-identification via softened similarity learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 3387–3396.
- [8] Tian J, Tang Q, Li R, et al. A camera identity-guided distribution consistency method for unsupervised multi-target domain person re-identification. *ACM Transactions on Intelligent Systems and Technology*, 2021, 12 (4): 1–18.
- [9] Xuan S, Zhang S. Intra-inter camera similarity for unsupervised person re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021: 11921–11930.
- [10] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 1116–1124.
- [11] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking. In: Hua G, Jégou H, editors. Computer Vision—ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 17–35.
- [12] Song L, Wang C, Zhang L, et al. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 2020, 102: 107173.
- [13] Fu Y, Wei Y, Wang G, et al. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 6111–6120.
- [14] Lin Y, Dong X, Zheng L, et al. A bottom-up clustering approach to unsupervised person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33 (1): 8738–8745.
- [15] Zeng K, Ning M, Wang Y, et al. Hierarchical clustering with hard-batch triplet loss for person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 13654–13662.
- [16] Zhang X, Cao J, Shen C, et al. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 8221–8230.
- [17] Yu H X, Zheng W S, Wu A, et al. Unsupervised person re-identification by soft multilabel learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 2143–2152.
- [18] Wang D, Zhang S. Unsupervised person re-identification via multi-label classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 10978–10987.
- [19] Wei L, Zhang S, Gao W, et al. Person transfer GAN to bridge domain gap for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 79–88.
- [20] Zhao L, Li X, Zhuang Y, et al. Deeply-learned part-aligned

- representations for person re-identification. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, **2017**: 3239–3248.
- [21] Zou Y, Yang X, Yu Z, et al. Joint disentangling and adaptation for cross-domain person re-identification. In: Vedaldi A, Bischof H, Brox T, editors. *Computer Vision—ECCV 2020*. Cham: Springer International Publishing, **2020**: 87–104.
- [22] Choi Y, Choi M, Kim M, et al. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 8789–8797.
- [23] Yang F, Zhong Z, Luo Z, et al. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, **2021**: 4853–4862.
- [24] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. **2018**. <https://doi.org/10.48550/arXiv.1807.03748>. Accessed December 12, 2021.
- [25] Zheng Y, Tang S, Teng G, et al. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, **2021**: 8351–8361.
- [26] Wu Y, Huang T, Yao H, et al. Multi-centroid representation network for domain adaptive person re-ID. *Proceedings of the AAAI Conference on Artificial Intelligence*, **2022**, 36 (5): 2750–2758.
- [27] Wu Y, Wu X, Li X, et al. MGH: metadata guided hypergraph modeling for unsupervised person re-identification. In: MM '21: Proceedings of the 29th ACM International Conference on Multimedia. New York: Association for Computing Machinery, **2021**: 1571–1580.

(Continued from page 5-8)

References

- [1] An Q, Chen H, Wu J, et al. Measuring slacks-based efficiency for commercial banks in China by using a two-stage DEA model with undesirable output. *Annals of Operations Research*, **2015**, 235 (1): 13–35.
- [2] Cook W D, Liang L, Zha Y, et al. A modified super-efficiency DEA model for infeasibility. *Journal of the Operational Research Society*, **2009**, 60 (2): 276–81.
- [3] Liang X, Zhou Z. Cooperation and competition among urban agglomerations in environmental efficiency measurement: A cross-efficiency approach. *JUSTC*, **2022**, 52 (4): 3.
- [4] Chen Y, Tsionas M G, Zelenyuk V. LASSO+DEA for small and big wide data. *Omega*, **2021**, 102: 102419.
- [5] Lee C Y, Cai J Y. LASSO variable selection in data envelopment analysis with small datasets. *Omega*, **2020**, 91: 102019.
- [6] Golany B, Roll Y. An application procedure for DEA. *Omega*, **1989**, 17 (3): 237–250.
- [7] Boussofiene A, Dyson R G, Thanassoulis E. Applied data envelopment analysis. *European Journal of Operational Research*, **1991**, 52 (1): 1–15.
- [8] Bowlin W F. Measuring performance: An introduction to data envelopment analysis (DEA). *The Journal of Cost Analysis*, **1998**, 15 (2): 3–27.
- [9] Cooper W W, Seiford L M, Tone K. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. New York: Springer, **2007**.
- [10] Sehra S, Flores D, Montañez G D. Undecidability of underfitting in learning algorithms. In: 2021 2nd International Conference on Computing and Data Science (CDS). Stanford, CA: IEEE, **2021**: 28–29.
- [11] Ueda T, Hoshiai Y. Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. *Journal of the Operations Research Society of Japan*, **1997**, 40 (4): 466–478.
- [12] Adler N, Golany B. Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operational Research Society*, **2002**, 53 (9): 985–991.
- [13] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **1996**, 58 (1): 267–288.
- [14] Rosa G J M. The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Hastie T, Tibshirani R, and Friedman J. *Biometrics*, **2010**, 66 (4): 1315–1315.
- [15] Li S, Fang H, Liu X. Parameter optimization of support vector regression based on sine cosine algorithm. *Expert Systems with Applications*, **2018**, 91: 63–77.
- [16] Breiman L. Random forests. *Machine Learning*, **2001**, 45 (1): 5–32.
- [17] Friedman J H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **2001**, 29 (5): 1189–1232.
- [18] Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, **2003**, 3: 1157–1182.
- [19] Mézard M, Montanari A. *Information, Physics, and Computation*. Oxford: Oxford University Press, **2009**: 584.
- [20] Proflidis V A, Botzoris G N. Chapter 5: Statistical methods for transport demand modeling. In: *Modeling of Transport Demand*. Amsterdam: Elsevier, **2019**: 163–224.
- [21] Biswas S, Bordoloi M, Purkayastha B. Review on feature selection and classification using neuro-fuzzy approaches. *International Journal of Applied Evolutionary Computation*, **2016**, 7: 28–44.
- [22] Fraser A M, Swinney H L. Independent coordinates for strange attractors from mutual information. *Physical Review A*, **1986**, 33 (2): 1134–1140.
- [23] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. *Science*, **2011**, 334 (6062): 1518–1524.
- [24] Zhang Z, Dong J, Luo X, et al. Heartbeat classification using disease-specific feature selection. *Computers in Biology and Medicine*, **2014**, 46: 79–89.
- [25] Soares F, Anzanello M J. Support vector regression coupled with wavelength selection as a robust analytical method. *Chemometrics and Intelligent Laboratory Systems*, **2018**, 172: 167–173.
- [26] Friedman J H. Multivariate adaptive regression splines. *The Annals of Statistics*, **1991**, 19 (1): 1–67.
- [27] Breiman L. Bagging predictors. *Machine Learning*, **1996**, 24 (2): 123–140.