International Journal of Machine Learning and Cybernetics (2024) 15:1779–1793
https://doi.org/10.1007/s13042-023-01997-1

ORIGINAL ARTICLE

A new robust contrastive learning for unsupervised person re-identification

Huibin Lin[1] · Hai-Tao Fu[1] · Chun-Yang Zhang[1] · C. L. Philip Chen[2]

Received: 7 May 2023 / Accepted: 30 September 2023 / Published online: 16 November 2023

Abstract
Unsupervised person re-identification (Re-ID) is more substantial than the supervised one because it does not require any labeled samples. Currently, the most advanced unsupervised Re-ID models generate pseudo-labels to group images into different clusters and then establish a memory bank to calculate contrastive loss between instances and clusters. This framework has been proven to be remarkably efficient for unsupervised person Re-ID tasks. However, clustering operation inevitably produces misclassification, which brings noises and difficulties to contrastive learning and affects the initialization and updating of the prototype features stored in the memory bank. To solve this problem, we propose a new robust unsupervised person Re-ID model with two developed modules: Cluster Sample Aggregation module (CSA) and Hard Positive Sampling strategy (HPS). The CSA module aggregates each sample in the same cluster through the multi-head self-attention mechanism. This process enables the initialization of prototypes based on the similarities observed within clusters. Additionally, the HPS strategy extracts the dispersion degree of each sample by means of a self-attention aggregation module (SAA) that has been trained by CSA module. According to the obtained indicators, the hardest positive sample is sampled to update the prototype feature stored in the memory bank. With the self-attention mechanism fusing the information among instances in each cluster, the implicit relationships between samples can be better explored in a more refined way. Experiments show that our method achieves state-of-the-art results against existing unsupervised baselines on Market-1501, PersonX, and MSMT17 datasets.

Keywords Person Re-ID · Contrastive learning · Unsupervised learning · Self-attention

1 Introduction

Given a query image of one person, the Re-ID task aims to identify this person from a large number of non-overlapping images recorded by different cameras, contributing to the person tracking and retrieval. Supervised person Re-ID has achieved significant performance improvement on several real-world datasets over many years of development, but it is still far from being as effective in applications. Because it takes a lot of human efforts to annotate the person bounding box and identity for each image and has poor generalization and adaptation across various environments, unsupervised person Re-ID has received more attention in recent years. At present, there are two categories for unsupervised person Re-ID. One is Unsupervised Domain Adaptation (UDA), which is given with a set of labeled samples in the source domain and a collection of unlabeled samples in the target domain. The goal of UDA is to train a model with good cross-domain performance [1, 2]. Due to the significant variations between domains,

Hai-Tao Fu, Chun-Yang Zhang have contributed equally to this work.

✉ Hai-Tao Fu
13067413633@163.com

✉ Chun-Yang Zhang
zhangcy@fzu.edu.cn

Huibin Lin
huibinlin@outlook.com

C. L. Philip Chen
Philip.Chen@ieee.org

[1] College of Computer and Data Science, Fuzhou University, Fuzhou 350116, Fujian, China

[2] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, Guangdong, China

its performance is somehow unstable. The other is pure Unsupervised learning (USL), similar to other unsupervised tasks, only unlabeled samples are used to extract generalization and discriminative features to obtain a model [3–5].

The images of different person Re-ID datasets are collected under different camera angles, illumination, and distances. All of them have their unique styles, which brings challenges for recognition accuracy. The UDA focuses on invariant features of the two datasets from the source domain and the target domain. Therefore, if the data distribution of the source domain is entirely different from the target domain, there will be a remarkable decrease in performance. In contrast, USL is more flexible and challenging than UDA, because it does not need to import additional source domain datasets and model can be learned without sample annotation. The training and test sets for unsupervised person Re-ID are consistent in style, that is, training and test sets are independent and identically distributed (iid). To this end, USL mainly extracts the generalized features of all images. The performance of the unsupervised person Re-ID model depends on the strength of image representation learning. Different images belonging to the same ID should be close in the feature space, while those with different IDs should be as far as possible. The difficulty of unsupervised person Re-ID is how to effectively learn image representations without the guidance of labels.

The state-of-the-art USL for person Re-ID [5, 6] mainly obtains pseudo-labels of instances by clustering algorithms [7, 8] and further determines positive and negative samples for contrastive learning. At the same time, a dynamic dictionary based on memory bank is built to store the features of samples or prototypes (cluster centers) [9, 10]. This framework has achieved excellent performance in unsupervised person Re-ID, even better than supervised ones. First, the methodology entails the selection of a pre-trained image feature extractor, typically a deep convolution network such as ResNet-50 [11], where the feature extractor is fixed to produce a feature vector for each image from the training dataset. Second, a density-based clustering algorithm is employed to obtain pseudo-labels for all images. Meanwhile, a small number of outliers will be discarded to reduce the risk of misclassifications. Third, the features of each cluster center are calculated, and then all the features of images and cluster centers are stored in a memory bank, which can be continuously updated to improve the diversity of samples. Many previous works [12, 13] take each image as a separate class in the memory bank, yet some others [6, 14] only record cluster prototype features. The former focuses on discrimination between different samples, whereas the latter significantly reduces training complexity and provides a broader range of perspectives for cluster centers. For each instance, this framework updates its features in the memory bank by contrasting it with corresponding prototypes or samples from different cluster centers.

By incorporating the memory bank mechanism, the existing contrastive models undergo training using multiple augmented views generated from each sample compared to prototype features. These augmented views are subsequently compared to prototype features, yielding expected losses, such as InfoNCE loss [15], Triplet loss [16], or other multi-categorization losses [17]. Augmented views contain rich semantic information, allowing the model to mine potential relationships among samples and be robust during multiple iterations of learning [18]. To facilitate this, in each gradient backpropagation of loss, the encoder is used to extract augmented features of images to update representations stored in the memory bank, which not only keeps the speed of updating the memory bank consistent with that of the model [14], but also compares images from different perspectives of the same type in contrastive learning. Although there are several ways to update a memory bank, such as using all samples in each batch to update the corresponding features in the memory bank [19] or selecting the hardest positive sample to update [14]. During training at each batch iteration, the common way is to use all of its shuffled samples to update prototype features with momentum. However, the order of images is randomly sampled during each training iteration, causing the feature information of the last image to occupy most of the weight in the memory bank when updating with momentum. Therefore, the update of memory bank features can be further optimized by considering the choice or order of sampling.

The initialization and updating of the memory bank rely on pseudo-labels. However, the acquisition of pseudo-labels through clustering algorithms without the guidance of labels will be inevitably prone to misclassification. This is due to the fact that images captured by the same person under different camera conditions can also be indistinguishable. Besides, some different persons have high similarities, such as similar clothing, resulting in extracted features that are too close to each other and thus misclassified as the same category. After that, if the samples in the same cluster are saved in the memory bank and fused with the misclassified features together, it will affect the representation ability of prototypes of clusters. In order to solve this problem, we propose a new Cluster Sample Aggregation module (CSA), which allocates different attention to the original features to reduce the weight of misclassified and increase the importance of hard positive samples containing rich semantics. In this way, we can obtain more excellent prototype features in the initialization stage. Secondly, when unsupervised person Re-ID tasks update the prototype features stored in the memory bank during training, most of them use all features in the batch to update prototypes with momentum. Similar to the

above problem, if the misclassified samples are substituted into the update process, the model will suffer from confirmation bias. In addition, it is also crucial to select some excellent augmented samples with rich semantics to update the memory bank for improving model's performance. To optimize the problem of iterative update of cluster prototypes, we propose an attention-based Hard Positive Sampling strategy (HPS) module to learn more representative cluster prototypes. As shown in Fig. 1, when CSA and HPS are not introduced, misclassified samples in the cluster easily become noise during model learning. On the other hand, CSA and HPS suppress the noise effects on the initialization and update of the memory bank, respectively.

In summary, the contributions of this paper are as follows.

- We propose a self-attention module that aggregates all the samples in a cluster based on the importance between them to generate the corresponding cluster center, which avoids the impact of error pseudo-labels generated by clustering algorithms.
- The proposed multi-head Self-Attention Aggregation module (SAA) is used to extract the degree of dispersion between samples in the same cluster, and the hardest positive sample is selected as the cluster center to update prototype feature with momentum. The hardest samples can learn more meaningful semantics and reduce the impact of misclassification.
- Extensive experiments on the current famous large-scale person Re-ID dataset using identical test criteria demonstrate that our model outperforms the most advanced methods and improves the performance of unsupervised person Re-ID.

2 Related works

Person Re-ID is a mature and extensively studied field accompanied by a large amount of works. To lay the stage for the proposed method SACL, in this section we provide a brief introduction for researches that are closely related to SACL. In particular, the most advanced unsupervised person Re-ID methods usually adopt techniques including memory bank and self-attention. We describe related works from the following three perspectives.

2.1 Unsupervised person Re-ID

Currently, the most recent unsupervised person Re-ID tasks are based on pseudo-labels and a contrastive learning framework, contrasting with features stored in the memory bank. Typical approaches include MoCo [20] and SimCLR [21]. The application in an unsupervised person Re-ID task can usually be divided into four stages: clustering to generate pseudo-labels, memory dictionary initialization, memory dictionary updating, and model training. The model's performance can be improved to a certain extent by refining the different stages.

SpCL [3] proposes a new self-paced contrastive learning framework using hybrid memory to leverage multi-domain information. This method stores different domain features in the hybrid memory dictionary, where the cluster-level contrastive loss is utilized to update the instance-level memory dictionary. However, when computing the cluster center of the same group instances to acquire cluster prototype, it is inevitable to introduce noise generated by misclassified samples. Furthermore, the inconsistent speed in updating

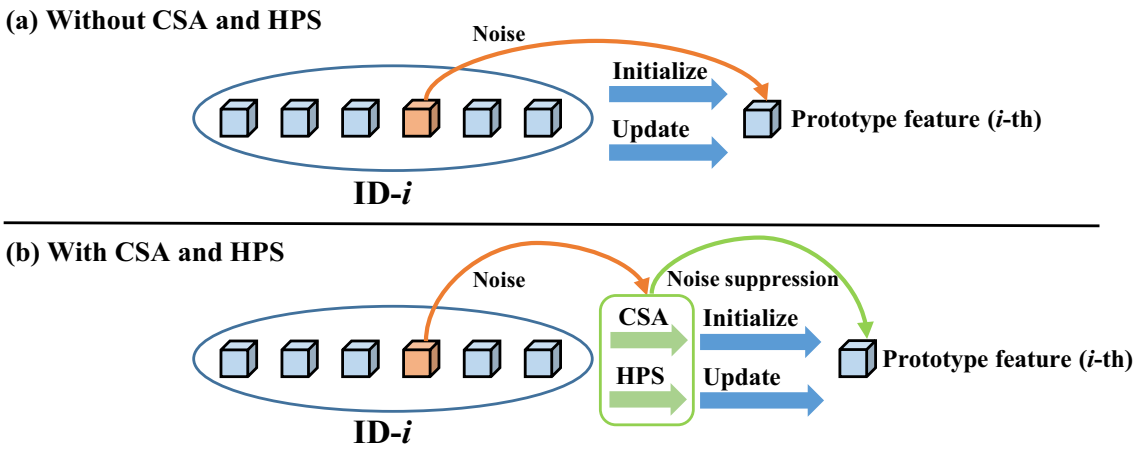

Fig. 1 The impact of error clustering on the model under noise conditions. **a** Without CSA and HPS. **b** With CSA and HPS. CSA is adopted to learn the relationship between samples to initialize prototype features. Besides, HPS selects the hardest positive sample to update the prototype features in the memory bank, allowing the cluster prototypes more refinement. Both two modules play a key role in suppressing the influence of noise on the model

the instances stored in the memory can lead to poor performance. In order to alleviate this situation, CCL [14] proposes to store the cluster-level dynamic memory and the contrastive loss between instances and clusters. CCL has shown that the inconsistency of prototype representation can be effectively resolved through the cluster-level memory dictionary. However, the problem of dynamically updating memory bank still persists when using an individual feature with noise. To mitigate this issue, DCC [4] establishes both instance-based and centroid-based update mechanisms in a unified cluster contrastive framework, which maintains both the instance memory and cluster centroid memory.

Besides, RLCC [22] proposes to refine pseudo-labels with aggregated soft labels to improve accuracy. Zhou et al. [23] propose a heterogeneous dual model framework with two asymmetric networks to mitigate the pseudo-label noise and thus maintain the consistency of the feature space. Recently, Liu et al. [24] propose CACHE algorithm to refine clustering results, which includes complementary attention-driven contrastive learning and hard-sample exploration modules.

It can be seen that the experimental performance is heavily influenced by the results of clustering. In this paper, we focus on reducing the impact of low-quality pseudo-labels generated by clustering in USL. Furthermore, the subsequent sections of this study explore optimizing the prototype features in the memory from different perspectives.

2.2 Memory dictionary

Contrastive learning combined with memory banks [9] can augment sample diversity. Recent unsupervised and weakly supervised vision tasks [10, 25] have demonstrated that considerable results can be achieved by constructing dynamic dictionaries. Initialization of memory bank features before model training and continuous update with data-augmented samples can enhance the diversity of learning samples and facilitate the training of unsupervised models.

In recent years, the memory bank mechanism has been widely used in unsupervised person Re-ID tasks, and various different methods being employed to construct memory dictionaries have a large performance gap. Some models [3, 12] achieve this goal by storing each instance sample in a memory bank, which requires a large amount of memory space to save these features and leads to the problem of inconsistent feature updates. In contrast, some other models [4, 5, 14] save cluster centers into a memory bank, which reduces memory space requirements but entails selecting an appropriate prototype generation method to avoid the noise produced by misclassified samples.

During the model training process, the memory dictionary undergoes constant updates, leading to the change of positive and negative sample pairs in each iteration. This enables the model to learn the features of the same person from different people. Since the prototype features in the memory bank represent each cluster, the dictionary is updated by leveraging feature vectors from the same group. The sample features used for updating are extracted from the model after real-time updates, thus ensuring the real-time nature of the updated samples. The latest experiments indicate that using samples to update prototype features in the memory bank directly is significantly better than storing all instances in each iteration [14, 19]. This approach avoids the problem of updating some samples in the same cluster while leaving others unchanged during the training process. However, how to generate a good prototype feature that can represent all samples in a cluster at the time of storage is a crucial consideration.

2.3 Self-attention

Self-attention is a technique for the calculation of dependencies between different units of the input sequence, using the relationships between the sequences to generate focal regions like human attention. The most classic self-attention model is the Transformer model [26], which has been widely used for feature representation computation in sequential tasks. It is not only excellent for traditional NLP tasks but also suitable for migrating to computer vision [27, 28] and graph networks [29, 30]. Transformer calculates the relationship between different units in the sequence as the attention weight and iteratively aggregates the information from similar units. Self-attention in Transformer is usually combined with multi-head attention to divide the sequence into several parts, each of which maps to a different feature space. Multi-head attention mechanism enables the model to focus on different aspects of information in different feature subspaces, allowing for more comprehensive and effective results obtained from multiple perspectives.

This paper mainly applies the multi-head self-attention mechanism to mine the implied relationships between samples within a cluster beyond the original features and view the connections between samples from multiple perspectives. In person Re-ID task, the photos of the same person taken by different cameras can vary considerably. In contrast, different persons captured by cameras at a certain angle can reach a level of recognition that even surpasses the human visual system. In such cases, it is difficult to mine hard positive and negative samples solely relying on the original spatial similarity without the right metrics to guide the model. Therefore, it is crucial to learn implicit information among samples through a multi-head self-attention mechanism.

3 Proposed method

3.1 Overview

Training an unsupervised person Re-ID model needs to give an unlabeled dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, where N represents the number of images in the dataset. The goal is to train a model that can generate powerful discriminative features, allowing the images of the same person under different shooting conditions to be as close as possible.

This paper proposes an unsupervised person Re-ID model, which is illustrated in Fig. 2. As can be seen from the figure, the learning of the model is mainly divided into two parts: initialization and training. It is worth mentioning that the gray SAA and encoder modules are shared with the green ones that are updated during the initialization and training phases, respectively. In other words, gray modules do not participate in backpropagation for parameter updates.

During initialization, given training set $\mathcal{X}$ is passed through the feature extractor ResNet-50 [11], which has been pre-trained on the ImageNet [31] dataset, to obtain the feature vector $F = \{f_1, f_2, \ldots, f_N\}, f_i \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width of the corresponding an image feature, respectively. The image features F are subjected to Global Average Pooling (GAP) or Generalized Mean pooling (GeM) to obtain the higher-level representations of the image features $V = \{v_1, v_2, \ldots, v_N\}, v_i \in \mathbb{R}^D$. The Jaccard distance [32] between all images is then computed to produce different clusters using the clustering method. Most samples are assigned pseudo-labels, yet a small number of outliers that cannot be clustered together are ignored. The method of density-based clustering is adopted to group images that are very close in the feature space into the same cluster. To a certain extent, the supervised learning method can be considered for unlabeled datasets with pseudo-labels.

As for the training of each epoch progress, it is necessary to split each batch of the dataset into $P \times K$ samples, where P is the number of sample clusters defined before each epoch training phase, K is the number of samples per cluster, and $P \times K$ is equivalent to the batch size. This division allows the positive and negative sample pairs to maintain a constant proportion. Consequently, if each sample in the batch has a fixed number of positive and negative sample pairs, we can calculate the contrastive loss among samples. Specifically, the prototype corresponding to the samples within the same group is taken as positive sample pair, and other cluster centers of the same batch are regarded as negative sample pairs. The contrastive loss between samples and clusters can be expressed as

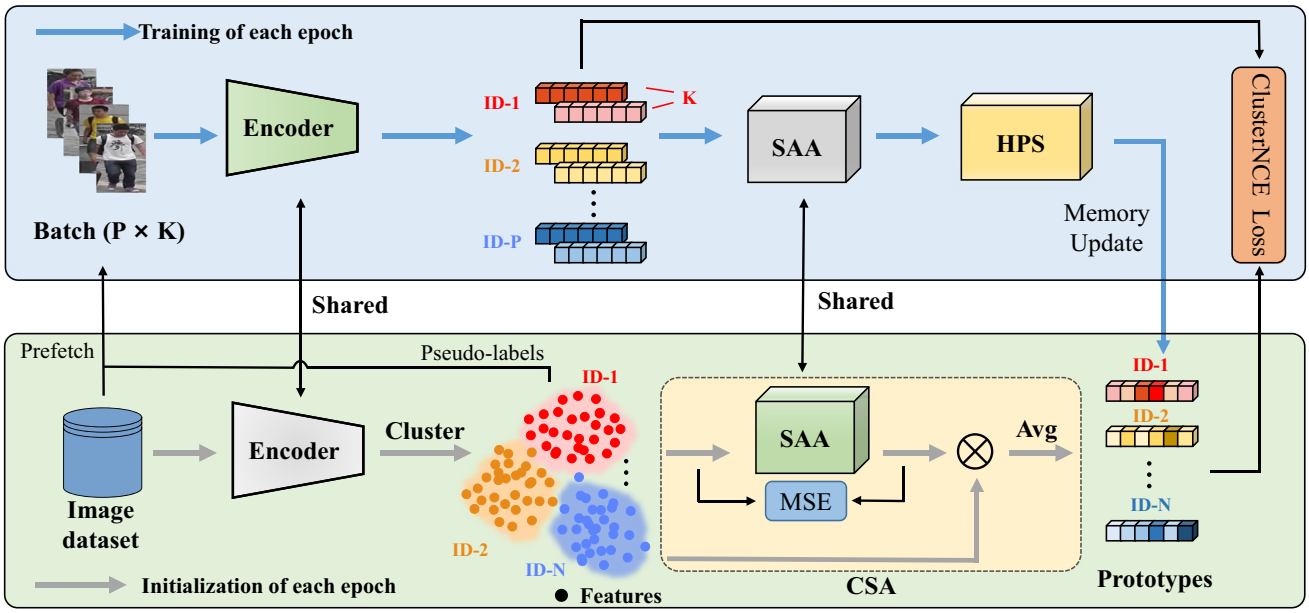

Fig. 2 The overall diagram of our unsupervised person Re-ID method. During each training epoch, a batch of P × K samples are randomly prefetched according to the pseudo-labels obtained at each initialization period, where P and K represent P clusters and the number of samples in each cluster, respectively. There are mainly two modules, which are represented by Cluster Sample Aggregation module (CSA), and the Hard Positive Sampling strategy (HPS), respectively. The Self-Attention Aggregation (SAA) module is updated during the initialization phase by the Mean Square Error (MSE) loss between the input feature and output global cluster feature (bottom), while the encoder and HPS module (top) are updated during training loop by the defined clusterNCE loss between different IDs and the prototypes saved in the memory bank

$$\mathcal{L}_{\mathcal{IC}} = \mathbb{E}\left[-log(\frac{exp(< v_a \cdot c_a > /\tau_c))}{\left(\sum_{i=1}^{|P|} exp(< v_a \cdot c_i > /\tau_c) \right.}\right], \quad (1)$$

where the symbol $\cdot$ represents an element-wise product, v_a denotes the sample feature belonging to the a-th cluster, c_i indicates the i-th prototype feature stored in the memory bank, and τ_c is a temperature hyperparameter.

In addition to the cluster contrastive loss, the overall loss also includes $\mathcal{L}_{\mathcal{HPS}}$ loss for suppressing the influence of noise using hard positive sampling strategy (introduced in the subsequent Sect. 3.3). The overall loss of our method is defined as

$$\mathcal{L}_{\mathcal{A}\updownarrow} = \alpha \mathcal{L}_{\mathcal{IC}} + \beta \mathcal{L}_{\mathcal{HPS}}. \quad (2)$$

The existing benchmark models employ contrastive learning by using different augmented views of various class samples, which significantly develops the representation ability of the feature extraction model. Despite the advancements achieved by current models, there is still great potential for improvement in these models. First, these models heavily relied on the assumption that pseudo-labels are equal to ground-truth labels. There are some challenging samples that are very difficult to identify and may be easily classified into other classes, which can affect the learning and training of the model. Therefore, it is crucial to find hard positive and negative sample pairs to improve the representation ability of model. The second one is to select an augmented view with an intense expression ability while updating the feature vectors stored in the memory bank to further optimize the contrastive loss.

As shown in Fig. 2, our self-attention contrastive learning model consists of two modules: Cluster Sample Aggregation module (CSA) and Hard Positive Sampling strategy (HPS). First, CSA is designed to optimize and initialize prototype features stored in the memory bank, which are acquired by aggregating all samples of the same cluster. Second, HPS strategy is adopted to update the prototype features by computing the hard positive sample of each cluster at each batch iteration. The hard positive samples contain more substantial information that can further improve the performance of the model.

3.2 Cluster sample aggregation module (CSA)

The initialization of prototype features plays a significant role in the early iterations of contrastive learning. Suppose a good prototype representation can be generated according to pseudo-label categories. In this context, it can guide the model to learn and cluster more high-quality pseudo-labels, with continuous interaction playing a key role in model learning. Therefore, initializing a prototype feature with refined cluster representation in the memory bank can effectively improve the performance of the model. Additionally, mislabeling in clusters is one of the critical factors affecting the discriminative of the model. To this end, it is necessary to assign different levels of attention to samples clustered as the same class, which avoids the influence of mislabeling and reduces the proportion of misclassified samples in the clusters to a certain extent.

The proposed CSA module is used to solve the above problem. As shown in Fig. 3, we concatenate all samples in the same cluster into a group, and insert a blank token with the same dimension as sample features in the head to form an input sequence $Z_i = \{z_i^0, z_i^1, z_i^2, \ldots, z_i^{t_i}\} \in \mathbb{R}^{(t_i+1)\times D}$, where Z_i denotes the sequence of tokens generated by the i-th cluster, z_i^0 represents the blank token at the first position of the sequence, and z_i^j indicates the j-th sample in the i-th cluster, containing a total of t_i samples. Then, the sequence Z_i is fed into the Self-Attention

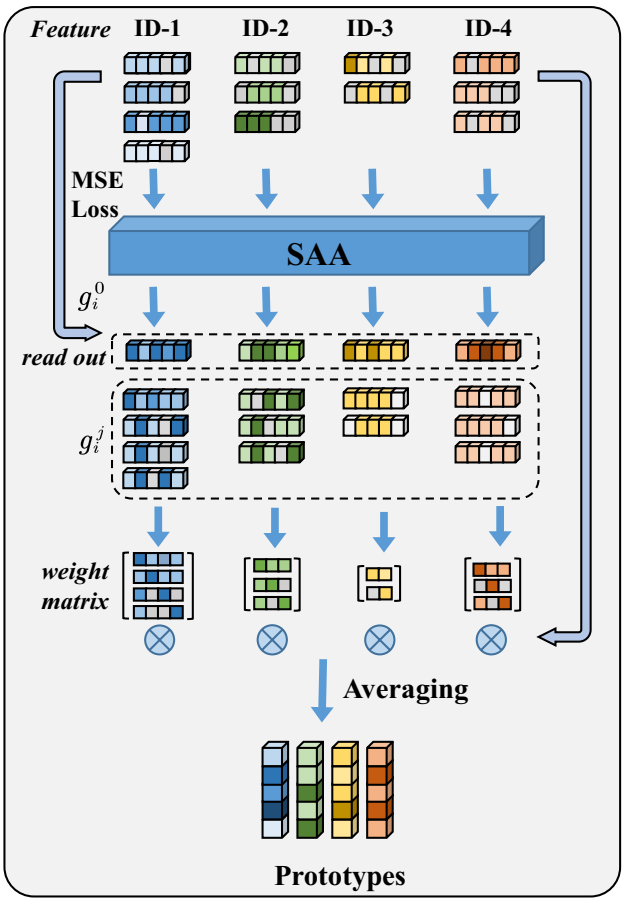

Fig. 3 Structure diagram of Cluster Sample Aggregation module (CSA), where SAA refers to the Self-Attention Aggregation model (SAA). Each sample with the same cluster ID is fed into the SAA module in parallel to obtain the corresponding global feature g^0. The prototype feature of each cluster is achieved through a process of multiplying the relationship weight matrix with the features of all samples within the cluster, followed by an average pooling process

Aggregation model (SAA) to obtain the aggregation feature $G_i = \{g_i^0, g_i^1, g_i^2, \ldots, g_i^{t_i}\} \in \mathbb{R}^{(t_i+1)\times D}$, which is a variant of self-attention mechanism [26]. G_i is composed of the aggregated global feature g_i^0 and the sample aggregation feature g_i^j, where $j \in [1, t_i]$. g_i^0 is the global cluster feature generated by the i-th cluster, which ideally should be as close as possible to all samples in the same cluster in the original feature space.

Specifically, as shown in Fig. 4, SAA first maps the sequence data to a new feature space, and the mapping feature $\widetilde{Z}_i$ is formulated as

$$\widetilde{Z}_i = Norm\left(Z_i W_{in}\right), \tag{3}$$

where $\widetilde{Z}_i \in \mathbb{R}^{(t_i+1)\times D_v}$, D_v is the dimension size of the mapping space, and $W_{in} \in \mathbb{R}^{D\times D_v}$ denotes the learnable fully connected parameters.

Then, a LayerNorm (LN) operation and multi-head self-attention mechanism (MHA) are performed on the sequence $\widetilde{Z}_i$. Then, the middle feature is obtained via residual connection as shown:

$$\overline{Z}_i = MHA\left(LN\left(\widetilde{Z}_i\right)\right) + \widetilde{Z}_i. \tag{4}$$

Finally, G_i is generated by mapping $\overline{Z}_i$ to a new feature space after the residual connection and multi-layer perceptron (MLP) operation, with the same dimensionality as the initial feature.

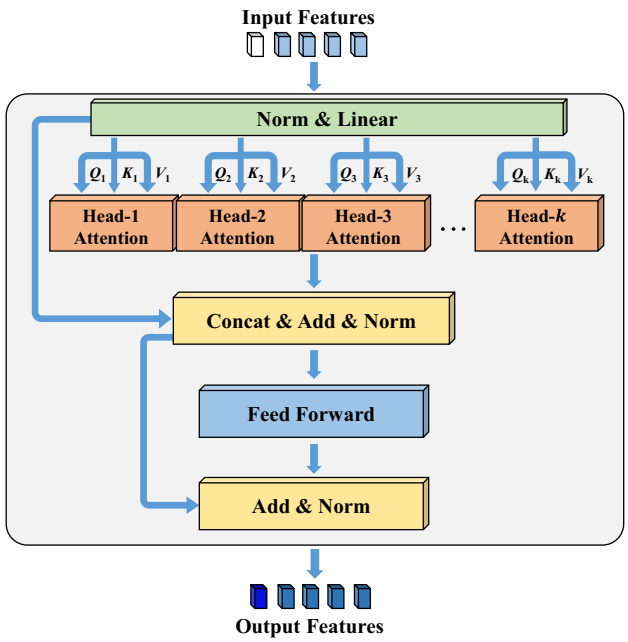

Fig. 4 Structure diagram of Self-Attention Aggregation model (SAA). An empty token is incorporated into the input clustering ID feature to gain the global cluster feature g^0

$$G_i = \left(MLP\left(LN\left(\overline{Z}_i\right)\right) + \overline{Z}_i\right) W_{out}, \tag{5}$$

where $W_{out} \in \mathbb{R}^{D_v\times D}$ also represents the learnable fully connected parameters, and D_v is the dimension of the sequence unit of the SAA module.

Denote $HA_h(X)$ as the self-attention function of the h-th head. It is designed by using the parameters W_{Q_h}, W_{K_h}, and W_{V_h} on the input mapping to obtain Q, K, and V, then the interrelationships is calculated on the mapping space of Q and K as weights assigned to V as

$$HA_h(X) = \text{Softmax}\left(\frac{\left((XW_{Q_h})(XW_{K_h})^T\right)}{\sqrt{(D_v)}}\right)\left(XW_{V_h}\right). \tag{6}$$

The $MHA(X)$ consists of multiple self-attention heads, called multi-head self-attention modules, where k self-attention heads are concatenated together and a learnable weight parameter W_{con} is adopted to measure the degree of influence of each self-attention head. Multi-head attention allows the model to focus on information from different representation subspaces. Formally, $MHA(X)$ is defined as

$$MHA(X) = \left[HA_1(X); HA_2(X); \ldots; HA_k(X)\right] W_{con}. \tag{7}$$

For a given cluster, the g_i^0 in the sequence G_i can be calculated as the global cluster feature, making it as close as possible to all samples in the same cluster. Hence, this part of learning can be regarded as a regression problem rather than a classification problem. In particular, we employ Mean Squared Error (MSE) as the measure of loss,

$$\mathcal{L}_{CSA} = \sum_{j=1}^{t_i} \frac{\left(g_i^0 - v_j\right)^2}{t_i}. \tag{8}$$

The mean square error between the global cluster feature g_i^0 and the original feature v_j of each sample within the cluster is calculated to promote g_i^0 as a suitable cluster center. The g_i^0 acquired from z_i^0 with an empty initial vector after executing SAA can fairly learn the self-attention among samples. This loss can guide g_i^0 closer to the expected cluster center, and the model will also induce other g_i^h $(h = 1, 2, \ldots, t_i)$ to fusion neighborhood samples from different perspectives. In practical, a prototype feature can be achieved by calculating the relationship between g_i^h to replace the global cluster feature g_i^0. Since the model involves multiple residual connections, g_i^h not only preserves the relationship in the original feature space but also includes the one between multi-angle domains. By calculating the relationship among the feature mapping g_i^h as the weight, we can get a refined attention relationship among cross-domain samples in the same cluster. In SAA mapping space, the relationship weight a_{ij} and the aggregated features H_i are defined as

$$H_i = A_i V_i \ \ (i \in 1, 2, \ldots, T), \tag{9}$$

$$a_{m,n}^{(i)} = \frac{exp(Norm(g^m \cdot g^n))}{\sum_{k=1}^{t_i} exp\left(Norm(g^m \cdot g^k\right)}, \tag{10}$$

where $A_i \in \mathbb{R}^{t_i \times t_i}$ is the relationship weight matrix of the i-th cluster composed of entry $a_{m,n}^{(i)}$, and T denotes the total number of clusters. V_i and $H_i \in \mathbb{R}^{t_i \times D}$ are the original features of the i-th cluster and the aggregated SAA features, respectively. Finally, the prototype feature can be acquired for i-th cluster by averaging the aggregated features H_i.

The relationship among g_i^h is used to generate the prototype feature in the initialization phase instead of the global cluster feature g_i^0 obtained by SAA. This is mainly due to the fact that large-scale models (e.g., self-attention) require plenty of data sources to achieve superior results. However, SAA only inputs the sample sequence of the same cluster in each iteration, and the number of iterations depends on the size of clusters. These data volumes are far from enough to support the training model in the early stage. If the original feature space is directly fused with the insufficient aggregation space, a significant deviation may occur, leading the model parameters to develop in an unpredictable direction. Although SAA projects each sample from the original feature space to a new feature space, the inter-sample relationships within the new feature space remain consistent with those of the original feature space. In a nutshell, the feature mapping of SAA module with few samples in the early stage can cause significant changes in the feature space, while the inter-sample dependencies in any feature space can maintain relatively invariant in terms of their relationship. As training progresses, the relationship among each feature in multi-angle will be gradually mined, and the new relationship will be assigned to the original features as weights.

3.3 Hard positive sampling strategy (HPS)

Before training, we use the CSA module to generate initialized prototype features and store them in the memory bank. During the training period, each batch includes $P \times K$ augmented images fed into the model, where the memory bank is updated with the corresponding P cluster prototypes when the loss gradient is returned. In this way, the diversity of cluster prototypes can be augmented through the features of the same person from different angles.

The method using momentum to update the corresponding cluster center is based on the assumption that the ground-truth labels of all samples in the same cluster belong to the identical person. However, this assumption may lead to some degree of error impact when updating cluster prototypes involving misclassified samples. In addition, giving equivalent rewards to all samples of the same cluster in the batch to update cluster prototypes hinders the model from focusing on the differences and regions of interest for clustering, which is detrimental to feature representation. Furthermore, a cluster with a large number of samples updates its corresponding prototype more frequently than a cluster with a small amount of samples because of the difference in cluster magnitude and the update time. Therefore, it is particularly crucial to select representative samples for a cluster with a smaller number of samples.

To address the issue, we propose a novel strategy named Hard Positive Sampling Strategy (HPS), as shown in Fig. 5, in which the most representative and informative samples in the cluster are selected to participate in the update. This allows the model to learn more discriminative and comprehensive prototype representation. Actually, the self-attention mechanism is calculated for all samples within a cluster, and then a sample with the lowest scores is selected as positive sample after sorting. In this way, we can obtain the sample that measures the affinity with other samples in the same cluster.

Firstly, the features in each batch divided into P clusters are parallel passed into the trained SAA from CSA module in the initialization stage to obtain the aggregated features of P clusters. For the convenience of description, we will refer to these clusters as "batch clusters". The internal self-attention is then calculated in terms of units of these batch clusters. The calculation method is shown in Eqs. (3)–(5) and Eq. (10) to obtain $\mathcal{A} \in \mathbb{R}^{P \times K \times K}$, where $\mathcal{A}$ and P are the self-attention matrix and the number of batch clusters in

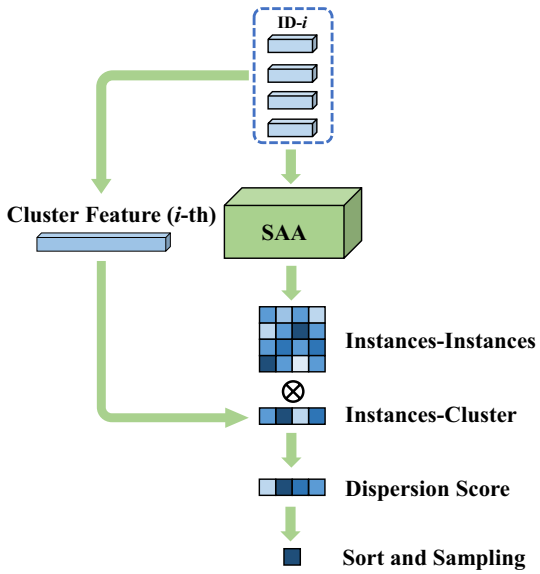

Fig. 5 Structure diagram of hard positive sampling strategy (HPS). The hardest positive sample is selected according to the index with the lowest score in dispersion score, which is used to update the corresponding prototype in the memory bank

each batch respectively, and K is the number of samples in each batch cluster. Since $\mathcal{A}$ represents the relationship between instances, the multi-angle consideration with SAA will be limited to the relationship between instances. Therefore, it is necessary to evaluate the degree of outlier for each sample in combination with the relationship between calculation instances and clusters. Sample dispersion score of the i-th batch cluster $S_i \in \mathbb{R}^K$ $(i = 1, 2, \ldots, P)$ in each batch is given by

$$S_i = \mathcal{A}_i \cdot \text{Softmax}(V_i c_i), \tag{11}$$

where $V_i \in \mathbb{R}^{K \times D}$ indicates the K sample features of the i-th batch cluster, and $c_i \in \mathbb{R}^D$ is the prototype feature corresponding to each cluster stored in the memory bank. The obtained score S_i demonstrates the relationship between instances-instances and instances-cluster into account to better measure the dispersion degree of each sample in the cluster and its structure is depicted in Fig. 6.

Finally, according to the index with the lowest score in S_i, we select the sample with this index as the feature v_{hard} of the hardest positive sample to update the corresponding prototype feature in the memory bank.

$$c_a \leftarrow \alpha c_a + (1 - \alpha) v_{hard}. \tag{12}$$

Compared with updating every sample to memory bank in the batch for $P \times K$ times, the method of updating only the hardest sample for P times is more effective. CCL [14] only uses the relationship between instances and clusters to select hard samples. However, our method also considers the self-attention relationship among instances through SAA to evaluate the degree of outliers from different angles to achieve better results.

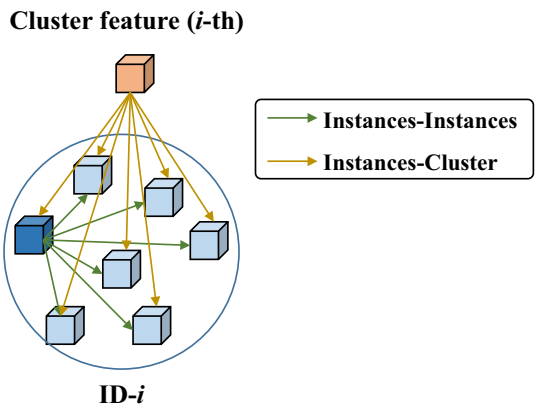

Fig. 6 Calculation of schematic diagram of dispersion score S_i. The green arrow indicates the relationship among instances, and the brown arrow indicates the relationship between instances and the cluster prototype

Although hard sample updating is better than the previous methods, it also suffers from a lack of robustness. Hard positive and negative samples are often difficult to distinguish due to the lack of label guidance. The fault-tolerant rate of hard samples that rely on outlier scores is meager. If the model selects non-positive samples in some exceptional cases, it will further affect the performance of the model. For example, the model is stuck in the situation of confirmation bias. In order to prevent the model from selecting non-positive samples as hard positives, we modify the triplet loss [16] to improve the classification ability and the validity of hard samples. The $\mathcal{L}_{\mathcal{HPS}}$ is defined as

$$\mathcal{L}_{\mathcal{HPS}} = \left[d_{hp} - d_{hn} + m \right]_+, \tag{13}$$

where m is the boundary of triplet loss, and d_{hp} and d_{hn} denote the distances between an anchor sample and its corresponding hardest positive sample, as well as the anchor sample and its corresponding hardest negative sample, respectively. It is worth mentioning that anchor samples are derived from prototypes stored in the memory bank, and the hardest positive and negative samples are selected from v_{hard}. If the cluster ID of a sample is identical to the anchor, the sample is considered a positive sample; otherwise, it is classified as a negative one. In addition, the hardest negative sample is chosen based on the criterion that the negative sample with the closest Euclidean distance to the anchor point.

On the one hand, because P hardest positive samples sampled from P batch cluster in a mini-batch can be seen as new various prototypes to some extent, the triplet loss over P hardest negative samples is calculated to facilitate the fit of the new cluster centers. On the other hand, the triplet loss further pulls the hardest positive samples and anchor points closer together, and pushes away the hardest negative samples from anchor points, thereby improving the representation ability of the model. As a result, this loss effectively mitigates the influence of non-positive samples, thereby facilitating HPS module in guiding the model to learn more robust and discriminative features for the memory update operation. As shown in Eq. (2), $\mathcal{L}_{\mathcal{IC}}$ and $\mathcal{L}_{\mathcal{HPS}}$ are taken as our final loss. The learning algorithm of the proposed model is summarized in Algorithm 1.

4 Experiments

4.1 Datasets and Evaluation Metrics

To evaluate the performance of the proposed SACL, we conduct comparative experiments on three widely-used person Re-ID datasets, including Market-1501 [33], MSMT17 [34]

Algorithm 1 Learning algorithm for SACLÍ

Input: Dataset $\mathcal{X}$, model parameter θ, hyperparameters.
Output: Feature mapping f_θ, SAA module S_θ.

for epoch in $\{1, \cdots, n\}$ **do**
  $\mathcal{X}$ obtains T pseudo-labels through DBSCAN algorithm;
  for SAA iterative training in $\{1, \cdots, T\}$ **do**
    Compute the CSA loss $\mathcal{L}_{\mathcal{CSA}}$ with Eq. (8);
    Update the parameters of S_θ by minimizing $\mathcal{L}_{\mathcal{CSA}}$;
  end for
  Calculate the weights in each cluster to initialization prototype features (CSA);
  Cut the batch with $P \times K$ size;
  for batch in $\{1, \cdots, batchnum\}$ **do**
    Compute the total loss $\mathcal{L}_{All}$ with Eq. (2);
    Update the parameters of f_θ by minimizing $\mathcal{L}_{All}$;
    Select hard samples to update prototype features (HPS);
  end for
end for

and PersonX [35]. The statistics of datasets are shown in Table 1 and the details are described as follows.

Market-1501 [33] is composed of a total of 32,668 images with 1,501 person identities, and each person is captured by up to 6 cameras, where 750 identities are used for training and 751 identities are assigned for testing. The training set comprises 12,936 images, the query set has 3,368 images, and the gallery set consists of 19,732 images. All images are cropped with a pedestrian detector. There are also some poorly detected samples as distractors in this dataset.

MSMT17 [34] is the largest available dataset and more challenging for research in person Re-ID task, which contains 126,441 images of 4101 identities taken from 15 camera views over a 4-day period. Specifically, 32,621 images with 1041 identities are used for training, and 11,659 and 82,621 images consist of 3,060 identities for validation and testing, respectively. Extreme lighting variations can be observed in different camera views.

PersonX [35] is a dataset synthesized using Unity [36] technology, which contains 1266 manually modeled pedestrians, including 547 women and 719 men, covering various types of clothing, walking posture, age, body shape and skin tones. There are 9840 images of 410 identities in the training set, and 5136 and 30,816 images of 856 identities in the query and test sets, respectively.

Table 1 Statistics of datasets

DataSet	Market-1501	MSMT17	PersonX
Train images	12,936	32,621	9840
Query images	3368	11,659	5136
Gallery	19,732	82,161	30,816
Train IDs	751	1041	410
Query IDs	750	3060	856
Cameras	6	15	6

To measure the performance of model, the Cumulative Matching Characteristic (CMC) [37] measurement including Rank-1, Rank-5 and Rank-10, and mean Average Accuracy (mAP) are employed as common criteria. The matching accuracy Rank-k ($k = 1, 5, 10$) indicates the top k similar features with matching values for a given query image feature. Moreover, person Re-ID is a multi-category retrieval problem, thus it is suitable to use mAP to evaluate the effect of the model.

4.2 Baselines

In this subsection, we provide a brief overview of eleven baseline models for comparative analysis, including BUC [13], SSL [38], JVCT [39], MMCL [17], GCL [40], SpCL [3], CAP [41], CACL [42], CCL [14], DCC [4], and HHCL [5]. The descriptions of each baseline model are as follows:

BUC [13] utilizes a bottom-up clustering approach to concurrently optimize the relationship between a convolutional neural network and individual samples. A softened similarity learning method is proposed by SSL [38] that aims to mitigate the effects of hard quantization losses incurred by clustering. JVTC [39] integrates local one-hot classification and global multi-class classification to jointly enforce visual and temporal consistency to ensure the quality of label prediction, thereby enhancing the quality of label prediction. An innovative loss function termed memory-based multi-label classification loss (MMCL) [17] is introduced with the purpose of generating discriminative features for person Re-ID tasks. GCL [40] combines generative adversarial networks and contrastive learning modules within a joint training framework, aiming to acquire view-invariant representations. SpCL [3] presents a novel self-paced contrastive learning framework, which is introduced in Sect. 2.1. In each cluster, CAP [41] introduces the concept of camera-aware proxies to generate reliable pseudo-labels, which are

further used for intra- and inter-camera contrastive learning. An asymmetric contrastive learning framework is designed in CACL [42] to help the siamese network efficiently mine invariants in representation learning. A dynamic memory bank cluster contrastive loss is proposed by CCL [14] and introduced in Subsection 2.1. DCC [4] introduces a novel dual cluster contrastive framework that leverages two memory banks for exchanging cross-view information during optimization. HHCL [5] presents bootstrap cluster-level and instance-level with hard samples for loss calculation.

4.3 Implementation Details

We choose ResNet-50 [11] pre-trained on ImageNet [31] as the feature extractor, and the extracted feature maps are passed through the Global Average Pooling (GAP) and Generalized Mean Pooling (GeM) [43] to obtain 2048-dimensional representation feature for each image. Before the training at each epoch, the features of all images in the training set are extracted under the current model parameters. The Jaccard distances [32] between image features are calculated. Pseudo-labels are then generated by using clustering methods, such as DB-SCAN [7] and Infomap [44], which find the $k = 30$ nearest neighbors of each image. There are only images with pseudo-labels are regarded as input samples for training, while samples that cannot be clustered are considered outliers without any pseudo-label assigned. Further, each image with pseudo-label is resized to 256×128, and several image-augmented operations are further performed, including random flipping, pixel filling, random clipping, and random erasing.

The augmented image features are randomly sampled from knowledge of pseudo-labels with batch size 128 and passed to ResNet-50. Each batch comprises a $P \times K$ tensor, where P is the number of clusters in the batch and K denotes the number of samples in each cluster. The hyperparameters of the proposed SACL are set up the same as CCL [14], which is a Re-ID model trained with the Adam optimizer. The initial learning rate is set to $3.5e-4$ and the weights are decayed to $5e-4$. The model is trained for a total of 60 epochs with a reduction of 10 percent in the learning rate after every 20 epochs. The SAA module has an initial learning rate of $1.2e-3$, the dimension of self-attention mapping space is fixed at 768, and the head number of multi-head attention is set to 16.

4.4 Comparison results

We compare our proposed method with the state-of-the-art unsupervised Re-ID models mentioned above. The comparative results on the Market-1501 dataset are reported in Table 2.

As shown in Table 2, our unsupervised person Re-ID model SACL has achieved SOTA performance in mAP of 84.8% and Rank-1 of 93.6%, while keeping comparable performance to other baselines on Market-1501. Although the proposed SACL model becomes a new state-of-the-art model with only a 0.6% improvement in mAP, it is still prominent for unsupervised models. The merits of our proposed method in terms of mAP metric can be visually demonstrated through Fig. 7, providing a more intuitive understanding of its effectiveness.

To investigate the generalizability of our model across diverse styles of datasets, we conducted verification on the 3D animation dataset PersonX, and the results are demonstrated in Table 3. The method with source domain is an unsupervised domain adaptation person Re-ID task, which requires a source dataset with labeled information for auxiliary training. Notably, the experimental outcomes demonstrate the robust stability of our proposed SACL method, exhibiting significantly superior performance compared to other approaches. To further validate the generalization

Table 2 Unsupervised person Re-ID performance comparison on Market-1501

Method	Full name	Market-1501			
		mAP	R1	R5	R10
BUC [13]	Bottom-Up Clustering	38.3	66.2	79.6	84.5
SSL [38]	Softened Similarity Learning	37.8	71.7	83.8	87.4
JVTC [39]	Joint Visual and Temporal Consistency	41.8	72.9	84.2	88.7
MMCL [17]	Memory-based Multi-label Classification Loss	45.5	80.3	89.4	92.3
GCL [40]	Generative and Contrastive Learning	66.8	87.3	93.5	95.5
SpCL [3]	Self-paced Contrastive Learning	73.1	88.1	95.1	97.0
CAP [41]	Camera-aware Proxy	79.2	91.4	96.3	97.7
CCL [14]	Cluster Contrast Learning	82.6	93.0	97.0	98.1
DCC [4]	Dual Cluster Contrastive	83.8	93.4	97.1	98.1
CACL [42]	Cluster-guided Asymmetric Contrastive Learning	80.9	92.7	97.4	**98.5**
HHCL [5]	Hard-sample Guided Hybrid Contrast Learning	84.2	93.4	**97.7**	**98.5**
SACL	Ours	**84.8**	**93.6**	97.5	98.2

The best performance is highlighted in bold

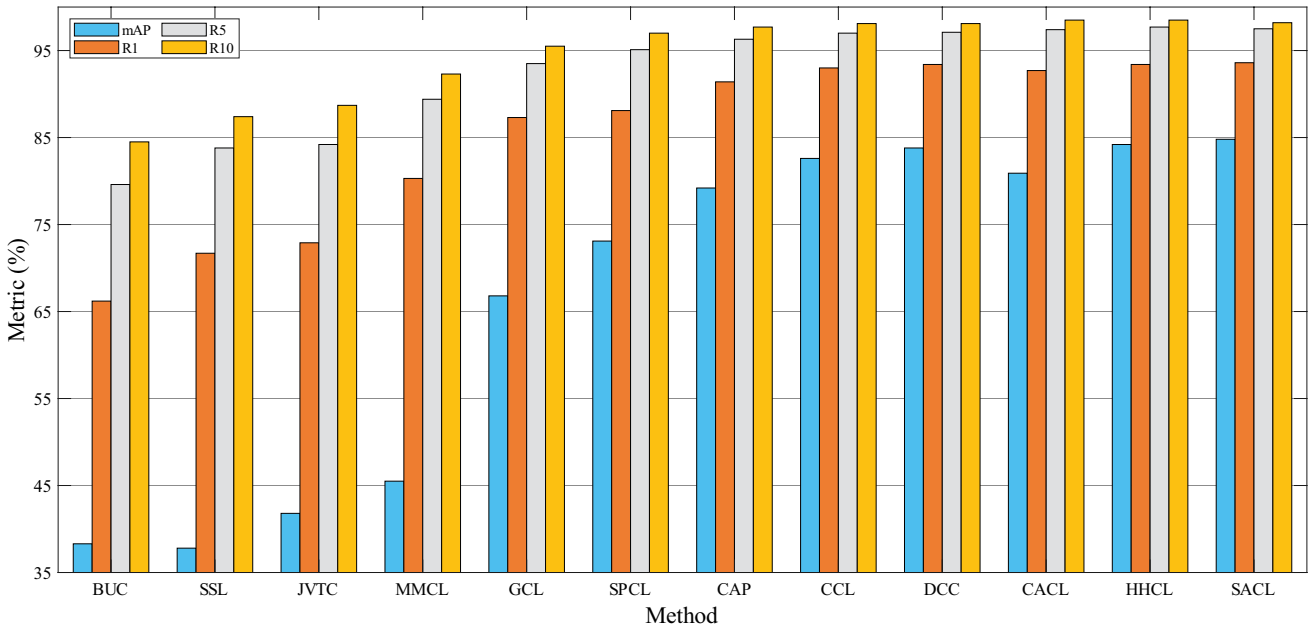

Fig. 7 Intuitive comparison of unsupervised person Re-ID on Market-1501

Table 3 Unsupervised person Re-ID performance comparison on PersonX

Method	Source Domain	PersonX			
		mAP	R1	R5	R10
SpCL [3]	None	72.3	88.1	96.6	98.3
MMT [45]	Market-1501	78.9	90.6	96.8	98.2
SpCL [3]	Market-1501	78.5	91.1	97.8	99.0
CCL [14]	None	84.8	94.5	98.4	99.2
SACL	None	**86.0**	**94.7**	**98.8**	**99.4**

The best performance is highlighted in bold

Table 4 Unsupervised person Re-ID performance comparison on MSMT17

Method	Source Domain	MSMT17			
		mAP	R1	R5	R10
ECN [12]	DukeMTMC-reID	10.2	30.2	41.5	46.8
MMCL [17]	None	11.2	35.4	44.8	49.8
TAUDL [46]	None	12.5	28.4	–	–
UGA [47]	None	21.7	49.5	–	–
MMT [45]	Market-1501	24.0	50.1	63.5	69.3
CycAs [48]	None	26.7	50.1	–	–
SpCL [3]	None	19.1	42.3	55.6	61.2
SpCL [3]	Market-1501	26.8	53.7	65.0	69.8
CCL [14]	None	27.6	56.0	66.8	71.5
SACL	None	**31.8**	**58.0**	**69.6**	**74.2**

The best performance is highlighted in bold

capability, we conducted a comparative analysis between SACL and existing USL and UDA models using the more challenging MSMT17 dataset, renowned for its large-scale and complex background environment. Based on the results presented in Table 4, it is evident that the evaluation metrics on MSMT17 dataset generally exhibit relatively low values. However, even in the face of this challenging dataset, the proposed method demonstrates a substantial improvement in performance. Therefore, Table 3 and Table 4 confirm our SACL is more effective in unsupervised person Re-ID task.

4.5 Ablation Studies

The SACL achieves state-of-the-art performance in unsupervised person Re-ID tasks mainly due to the two proposed modules: CSA and HPS. In the following, we perform ablation studies on Market-1501 to verify their validity, as shown in Table 5.

During the initialization phase, we explored three different methods for prototype generation. The first method involved averaging all samples with the same cluster ID to generate prototypes, called Mean (the baseline CCL model). This method yields 83.2% (the latest version of source code opened by the authors) in mAP on Market-1501 dataset. The second method, denoted as Self-Attention (Original features), where the self-attention mechanism is used in the original feature space, and the method has a slight improvement in mAP. Thirdly, it can be observed that the model's

Table 5 Ablation studies of different key components of SACL on Market-1501

Method	Market-1501			
	mAP	R1	R5	R10
Initialization method				
Mean (BaseLine)	83.2	92.9	97.0	97.7
Self-Attention (Original features)	83.4	92.8	96.9	97.9
Global feature g_i^0	81.1	92.0	96.7	97.6
CSA	84.0	93.5	97.1	98.0
Update method				
Random (BaseLine)	83.2	92.9	97.0	97.7
Mean	83.0	92.6	96.7	97.7
Hard Sample (Instances-Clusters)	83.7	93.0	97.0	98.1
HPS	84.3	**93.6**	97.2	98.0
Initialization and Update Method				
CSA + HPS (w/o Triplet)	84.4	93.2	97.4	**98.3**
CSA + HPS (w/ Triplet)	**84.8**	**93.6**	**97.5**	98.2

The best performance is highlighted in bold

performance degrades when using g_i^0 as the initial prototype, thereby providing empirical evidence that supports our explanation in Subsection 3.2. Finally, the model using CSA module has a mAP of 84.0% and a Rank-1 of 93.5%. Compared with the baseline models, this module is confirmed to improve mAP and Rank-1 by 0.8% and 0.4% in mAP and Rank-1 respectively.

As for the training phase, there are four ways to update the prototypes stored in the memory bank. Because the data in a mini-batch is divided into $P \times K$ samples, it is necessary to update prototype features based on the current K samples within the same cluster. The first update prototype method is the baseline model, which employs a momentum-based approach to randomly update the prototype using K samples from the same cluster. Due to the sequential update process, the last image in the sequence takes up a very large proportion. Moreover, the K samples in the batch cluster are selected by random sampling, thus the selection order and results exit random situation. It can be seen that the results of the baseline model on the Market-1501 are 83.2% mAP and 92.9% Rank-1. The second method is to calculate the mean feature of K samples within the same cluster, and use the mean feature as the basis for updating. However, this method achieves lower performance compared to the baseline model's update approach. The reason may be due to the fact that the model is not confident enough in the initial stage, and the frequent update of the mean feature in K clusters leads the model to fall into the confirmation bias. The third method adopts the HPS strategy without considering the instance-instance relationship. Notably, this method shows significant improvement in mAP, but is slightly changed in Rank-k metrics, which indicates that the strategy of selecting hard positive samples proves to be effective. Upon incorporating the two modules CSA and HPS, it is easy to observe that the mAPs of the SACL are up to 84.0% with 0.8% improvement and 84.3% with 1.1% gain respectively. In addition, it is noteworthy that the introduced triplet loss has a certain impact on HPS module. It not only optimizes HPS module to some extent, but also enhances the robustness brought by the method of updating only with the hardest sample. As a result of incorporating the triplet loss, the mAP scores on Market-1501 dataset show a significant improvement, reaching 84.4% (1.2% increase) and 84.8% (1.6% increase), respectively.

Furthermore, because there is only one prototype assigned to each cluster ID in the memory bank, we adopt HPS strategy to find the hardest positive sample within each batch cluster for prototype update. In order to investigate whether selecting the hardest negative sample is the best way, we conducted an additional evaluation by

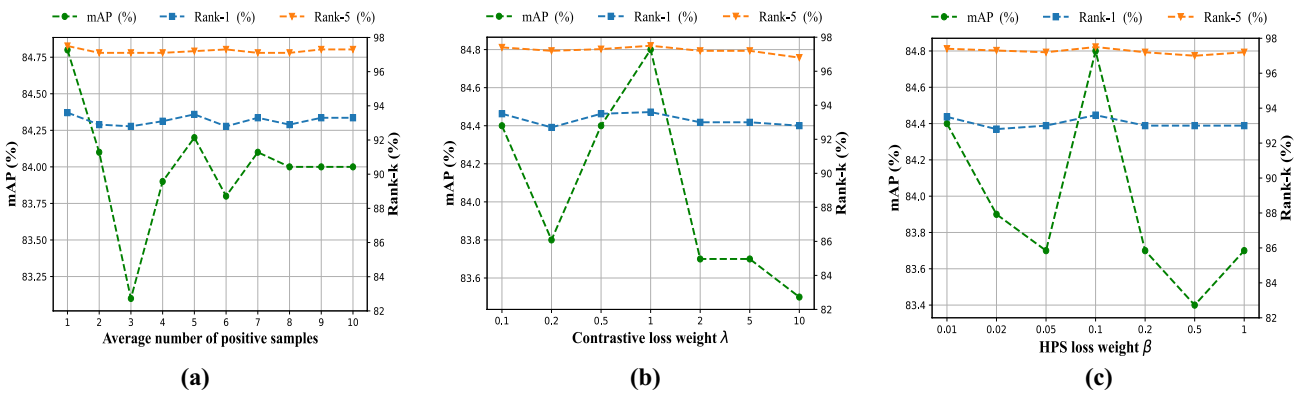

Fig. 8 Parameter sensitivity analysis. **a** Denotes the average number of different positive samples based on the top-k lowest scoring. **b**, **c** Denote the effect of different weight parameters on the model, respectively

considering an alternative approach, that is, selecting the lowest top-k positive samples based on the ranking of scores in S_i. Subsequently, a fused feature is formed by averaging these samples for comparative analysis. The results are depicted in Fig. 8a, which is easy to find that the proposed HPS strategy outperforms the fused feature significantly, especially in terms of mAP metric. We believe this is due to the inclusion of noise in the fused feature, which adversely affects their performance compared to HPS strategy. Figure 8b, c present the results of the parameter sensitivity analysis for prototype contrastive loss and HPS loss on the Market-1501 dataset, respectively. The experimental results indicate that the model exhibits remarkable stability concerning the R1 and R5 metrics. However, when it comes to the mAP metric, the optimal parameter values are approximately around 1 for contrastive loss and 0.1 for HPS loss, respectively, and any increase or decrease will lead to a decrease in model performance. The experimental results of each component mentioned above show the effectiveness of the proposed CSA and HPS modules.

5 Conclusion

In this paper, a novel unsupervised person Re-ID model, named SACL, is proposed. SACL comprises two key modules, i.e., Cluster Sample Aggregation module (CSA) and Hard Positive Sampling strategy (HPS). On the one hand, SACL reduces the risk of instance misclassification when assigning pseudo-labels for contrasting in the dominant unsupervised person Re-ID framework. On the other hand, it focuses on the hidden attentions among cluster samples at the initialization of the memory bank, and updates the initialized prototype features by considering the different importance of each sample via the multi-head self-attention mechanism. The learned CSA module is leveraged on samples in each batch to evaluate the closeness of each sample in the current fixed number of batch clusters. Besides, the prototype features in the memory bank are dynamically updated with the closest samples in the batch clusters using HPS module, so that hard samples can better facilitate contrastive learning. By introducing these two modules, SACL shows the SOTA performance, which is clearly confirmed by extensive experiments on benchmarks: Market-1501, PersonX, and MSMT17.

A potential limitation of our approach is that model clustering relies on DBSCAN algorithm to assign pseudo-labels. The choice of an effective clustering method is beneficial to prevent the model from getting stuck to confirmation bias. Therefore, in the future, we plan to explore transfer learning and combine it with current existing clustering methods to further improve the model performance. We also hope the proposed model can provide a promising direction for future research in the unsupervised person Re-ID task.

Acknowledgements This research is sponsored in part by the National Natural Science Foundation of China under Grant no. 62076065 and the Natural Science Foundation of Fujian Province under Grant no. 2020J01495.

Data availability The dataset used in this paper is publicly accessible based on cited references. In addition, the data that support the findings of this study are available on request from the first author, [Huibin Lin], upon reasonable request.

References

1. Zhai Y, Ye Q, Lu S, Jia M, Ji R, Tian Y (2020) Multiple expert brainstorming for domain adaptive person re-identification. In: Proceedings of the European Conference on computer vision, pp 594–611
2. Song X, Jin Z (2022) Domain adaptive attention-based dropout for one-shot person re-identification. Int J Mach Learn Cybern 13:255–268
3. Ge Y, Zhu F, Chen D, Zhao R, et al (2020) Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In: Advances in neural information processing systems, vol. 33, pp 11309–11321
4. Yao H, Xu C (2021) Dual cluster contrastive learning for object re-identification. arXiv preprint arXiv:2112.04662
5. Hu Z, Zhu C, He G (2021) Hard-sample guided hybrid contrast learning for unsupervised person re-identification. In: 2021 7th IEEE International Conference on network intelligence and digital content, pp 91–95
6. Chen H, Lagadec B, Bremond F (2021) Ice: inter-instance contrastive encoding for unsupervised person re-identification. In: Proceedings of the IEEE International Conference on computer vision, pp 14960–14969
7. Ester M, Kriegel H-P, Sander J, Xu X, et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol. 96, pp. 226–231
8. Xie K, Wu Y, Xiao J, Li J, Xiao G, Cao Y (2021) Unsupervised person re-identification via k-reciprocal encoding and style transfer. Int J Mach Learn Cybern 12:2899–2916
9. Bachman P, Hjelm RD, Buchwalter W (2019) Learning representations by maximizing mutual information across views. In: Advances in neural information processing systems, vol. 32
10. Tian Y, Krishnan D, Isola P (2020) Contrastive multiview coding. In: Proceedings of the European Conference on computer vision, pp 776–794
11. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 770–778
12. Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2019) Invariance matters: exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 598–607
13. Lin Y, Dong X, Zheng L, Yan Y, Yang Y (2019) A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on artificial intelligence, vol. 33, pp 8738–8745
14. Dai Z, Wang G, Zhu S, Yuan W, Tan P (2021) Cluster contrast for unsupervised person re-identification. arXiv preprint arXiv:2103.11568

15. Oord A, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv e-prints, 1807
16. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 815–823
17. Wang D, Zhang S (2020) Unsupervised person re-identification via multi-label classification. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 10981–10990
18. Ren Z, Zhang Y, Wang S (2022) Lcdae: data augmented ensemble framework for lung cancer classification. Technol Cancer Res Treat 21:1–14
19. Xiao T, Li S, Wang B, Lin L, Wang X (2017) Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 3415–3424
20. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 9729–9738
21. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International Conference on machine learning, pp 1597–1607
22. Zhang X, Ge Y, Qiao Y, Li H (2021) Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 3436–3445
23. Zhou H, Kong J, Jiang M, Liu T (2023) Heterogeneous dual network with feature consistency for domain adaptation person re-identification. Int J Mach Learn Cybern 14(5):1951–1965
24. Liu Y, Ge H, Sun L, Hou Y (2022) Complementary attention-driven contrastive learning with hard-sample exploring for unsupervised domain adaptive person re-id. IEEE Trans Circuits Syst Video Technol 33(1):326–341
25. Ren Z, Wang S, Zhang Y (2023) Weakly supervised machine learning. CAAI Trans Intell Technol 8:1–32
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems 30
27. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. In: International Conference on machine learning, pp 10347–10357
28. Zhang Y, Deng L, Zhu H, Wang W, Ren Z, Zhou Q, Lu S, Sun S, Zhu Z, Gorriz JM et al (2023) Deep learning in food category recognition. Inform Fusion, p 101859
29. Cai D, Lam W (2020) Graph transformer for graph-to-sequence learning. In: Proceedings of the AAAI Conference on artificial intelligence, vol. 34, pp 7464–7471
30. Yun S, Jeong M, Kim R, Kang J, Kim HJ (2019) Graph transformer networks. In: Advances in neural information processing systems, vol. 32
31. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 248–255. Ieee
32. Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 1318–1327
33. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on computer vision, pp 1116–1124
34. Wei L, Zhang S, Gao W, Tian Q (2018) 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
35. Sun X, Zheng L (2019) Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 608–617
36. Riccitiello J (2015) John riccitiello sets out to identify the engine of growth for unity technologies (interview). VentureBeat. Interview with Dean Takahashi. Retrieved January 18(3)
37. Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, pp. 1–7
38. Lin Y, Xie L, Wu Y, Yan C, Tian Q (2020) Unsupervised person re-identification via softened similarity learning. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 3390–3399
39. Li J, Zhang S (2020) Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In: Proceedings of the European Conference on computer vision, pp 483–499
40. Chen H, Wang Y, Lagadec B, Dantcheva A, Bremond F (2021) Joint generative and contrastive learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp 2004–2013
41. Wang M, Lai B, Huang J, Gong X, Hua X-S (2021) Camera-aware proxies for unsupervised person re-identification. In: Proceedings of the AAAI Conference on artificial intelligence, vol. 2, p 4
42. Li M, Li C-G, Guo J (2022) Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. IEEE Trans Image Process 31:3606–3617
43. Radenović F, Tolias G, Chum O (2018) Fine-tuning cnn image retrieval with no human annotation. IEEE Trans Pattern Anal Mach Intell 41(7):1655–1668
44. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123
45. Ge Y, Chen D, Li H (2020) Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526
46. Li M, Zhu X, Gong S (2018) Unsupervised person re-identification by deep learning tracklet association. In: Proceedings of the European Conference on computer vision, pp 737–753
47. Wu J, Yang Y, Liu H, Liao S, Lei Z, Li SZ (2019) Unsupervised graph association for person re-identification. In: Proceedings of the IEEE International Conference on computer vision, pp 8321–8330
48. Wang Z, Zhang J, Zheng L, Liu Y, Sun Y, Li Y, Wang S (2020) Cycas: self-supervised cycle association for learning re-identifiable descriptions. In: Proceedings of the European Conference on computer vision, pp 72–88