One-shot Video-based Person Re-identification based on SAM Attention and Reciprocal Nearest Neighbors Metric

Ji Zhang, Yuchang Yin, and Hongyuan Wang(✉)

Changzhou University, Changzhou Jiangsu 213164, China
hywang@cczu.edu.cn

Abstract. Pedestrian re-identification (Re-ID) is used to solves the recognition and retrieval of pedestrians across cameras. To solve the difficulties of label annotation in Re-ID, a one-shot video-based Re-ID method is applied. For the problem of large number of neural network parameters and weak feature extraction ability of the model, the SAM attention module is embedded. The module is plug-and-play without any additional parameters, while it can capture the hidden information in samples and improve the discriminative of model. To address the problem of low accuracy of label estimation, a reciprocal nearest neighbours metric is designed. The metric is capable of constructing closer nearest-neighbour relationships between samples, combining the Mahalanobis distance and Jaccard distance to significantly improve the accuracy of label estimation and model performance. The effectiveness of our method in this paper is extensively experimented on two video-based Re-ID datasets, MARS and DukeMTMC-VideoReID.

Keywords: Video-based Person Re-identification · One-shot Learning · Semi-supervised Learning

1 Introduction

Pedestrian re-identification (Re-ID) is a technique for detecting the presence of a specific pedestrian in a network of multiple cameras [1–5]. It plays an important role in intelligent video surveillance systems and has a wide range of applications in the field of public safety. With the developments of smart cities, a large number of cameras have been installed in cities and huge amount of video data is generated all the time. How to process video data efficiently is an important problem encountered by security personnel nowadays. Re-ID plays an important role in this regard, which reduce the consumption of manpower and economy effectively.

Most of the existing video-based pedestrian Re-ID methods use supervised based methods [6–10], which are extremely dependent on data annotation. But it is laborious and time-consuming for large-scale data annotations. Compared with supervised based methods, massive data annotations are not necessary for semi-supervised based methods. But these methods do not fully utilize the hidden information of unlabeled data, which

H. Lu et al. (Eds.): ACPR 2023, LNCS 14408, pp. 64–78, 2023.
https://doi.org/10.1007/978-3-031-47665-5_6

cause the performances of them are not as good as desired. One-shot Re-ID methods adopt a different training philosophy, which use only one labeled video clip for each identity in the training set, and the rest are unlabeled data. The current mainstream approach uses an incremental learning strategy to first assign pseudo-labels to the unlabeled data, then select some reliable pseudo-labeled data to merge with the original training set, and finally use the merged new data set to train the model again. Obviously, assignment of labels plays an important role in these methods, because it determines the assignment of pseudo-labels directly, and the correct assignment of pseudo-labels has a great impact on the next training. In other words, a correct pseudo-label sample has a positive effect on the training, while a wrong one provides wrong supervision information and prevents the model from learning the hidden information of the sample.

In this paper, a novel one-shot video-based pedestrian Re-ID method is designed based on SAM attention module and reciprocal nearest neighbor metric. The whole process can be briefly divided into three steps: (1) Initialize the model using labeled data. (2) Assign pseudo-labels for unlabeled data and select some reliable pseudo-labels to be combined with labeled data as a new training set. (3) Train the model again using the new training set. Due to the few number of initial labeled samples, the trained model is not discriminative enough. To address the problem of few number of labeled samples and low discriminative ability at the begining of training, we embed SAM attention module [11] in our network. In label estimation, in order to improve the accuracy of pseudo-label assignment, we design a reciprocal nearest neighbor metric, with k-reciprocal encoding [12], Mahalanobis metric and Jaccard metric for nearest neighbor samples, which is more robust and tighter and improves the accuracy of pseudo-label assignment. The main contributions of this paper are as follow:

(1) The SAM attention module is embedded in our network, which is flexible and effective to enhance the model discrimination without increasing the training burden.
(2) A reciprocal nearest neighbor metric is designed for pseudo-label estimation, which can effectively improve the accuracy of pseudo-label prediction.
(3) Our method in this paper focuses on the video-based pedestrian Re-ID problem under one-shot, and achieves competitive performance on two large-scale video pedestrian datasets, MARS and DukeMTMC-VideoReID.

2 Related Works

In recent years, video-based pedestrian Re-ID with deep learning has developed rapidly and achieved impressive results on major datasets [13–17]. Compared with image-based pedestrian Re-ID, video-based pedestrian Re-ID contains more pedestrian identity information and is accompanied by more noise and challenges. How to obtain sequence-level discriminant features is the core of supervised video-based pedestrian Re-ID. To address this problem, based on the spatio-temporal information of videos, many researchers extract more effective pedestrian features by integrating attention mechanisms. Li et al. propose a spatio-temporal attention module that discovers discriminative parts from pedestrian images and extract valid information without being affected by problems such as occlusion [13]; Hou et al. combine the attention mechanism with adversarial generative networks to design a spatio-temporal completion network, which recover the

occluded parts of pedestrian through the network, instead of discarding them, to effectively deal with the occlusion problem [14]; Li et al. draw on the work in image pedestrian Re-ID to explore multi-scale temporal cues in video sequences [15]; Liu et al. introduce a non-local attention module and reduce the computational complexity of extracting spatio-temporal features [16]; Eom et al. propose a spatio-temporal memory network, where spatial memory stores the features of the extracted spatial interference terms and temporal memory stores the attention used to optimize temporal patterns, this network improves the performance of attention mechanisms in the field of video-based pedestrian Re-ID [17].

Semi-supervised video-based Re-ID has been less studied, and most of the work focuses on semi-supervised video-based Re-ID under one-shot [18–24]. Zhu et al. use a semi-supervised dictionary learning approach to study the cross-view problem in video-based Re-ID by converting videos under different cameras into coding coefficients in the feature space to reduce the variation and differences between different cameras [18]; Liu et al. propose a metric boosting algorithm that iterates between model upgrading and label estimation, mainly for the problems of difficult sample mining and label propagation, and achieves good results on three datasets [19]. Wu et al. improve the framework of iterative training by drawing on the experience in [19], they also optimize the sampling strategy and the algorithm for pseudo-label assignment [20]. Similar to this iterative training approach, Ye et al. design a framework for dynamic graph matching to improve the label estimation process by continuously changing the graph structure, and also design a joint matching strategy to fully extract video information and reduce false matches [21]. Although all of the above methods are semi-supervised methods, the training and data selection strategies are different from each other. In particular, the traditional principle of nearest neighbor assignment is still used in label estimation, and pseudo-label estimation errors often occur. The method in this paper focuses on the problem of one-shot video-based Re-ID, using SAM attention as well as reciprocal nearest neighbor metric, which can effectively utilize unlabeled samples and improve the accuracy of pseudo-label prediction.

3 Method

3.1 Main Framework

Figure 1 shows the overall framework of our method proposed in this paper. The whole process consists of model initialization, label assignment, dynamic selection and model update. Firstly, the model is initialized using labeled samples. In previous works, ResNet-50 is the common choice. For the dual consideration of feature extraction and computation consumption of our model, the SAM attention module is embedded in ResNet-50. In label assignment, the reciprocal nearest neighbor metric is used, which is better in reflecting the relationships of samples and improving the accuracy of label estimation effectively, compared with the Euclidean metric. After each selection of pseudo-labeled data, the labeled data are merged with the selected pseudo-labeled data as the training set for the next training. The whole process is continuously iterated, and eventually all unlabeled data are selected.

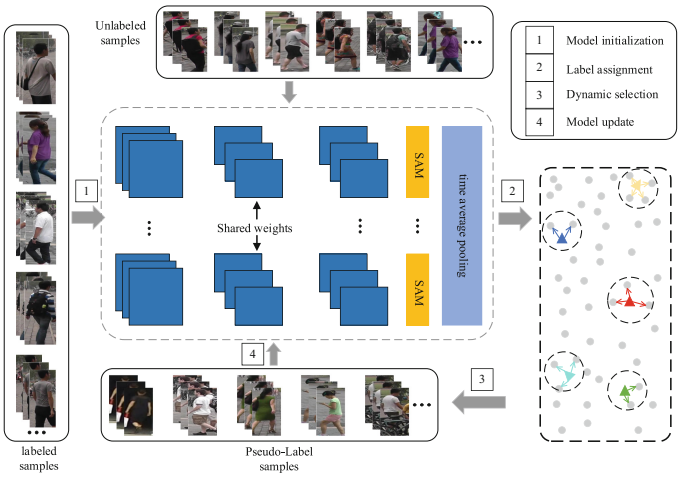


Fig. 1. The main framework

3.2 SAM Attention Module

In one-shot Re-ID, the labeled samples are unique for each identity. These samples are used to train the model firstly, then estimate labels for unlabeled ones with the model. However, due to the small amount of these labeled samples, researchers usually do not use more complex network structures. On the one hand, a complex network structure will trigger a huge amount of computation; on the other hand, a complex network structure does not necessarily enhance the feature extraction ability and discriminative ability of the model, but it may even fall into overfitting. To address the above problems, the SAM attention module is added into the ResNet-50 network, which is a plug-and-play module that does not require much adjustment of the network structure.

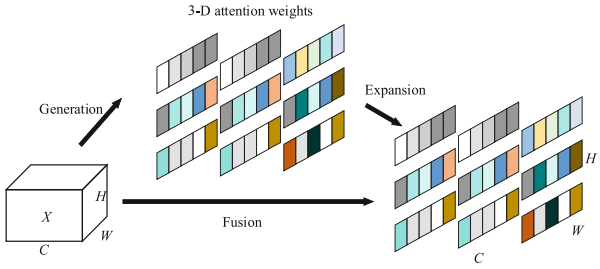


Fig. 2. 3-D attention weights

Existing attention models usually extract features from channels or image spaces only, which limit the flexibility of attention weights for network learning. In addition, the structures of these attention models are usually very complex, which bring enormous amount of computation. In contrast, the SAM module is simple and effective in that it

considers both spatial dimensions and channel dimensions and it is able to obtain 3-dimensional attention weights directly from the current neurons, as shown in Fig. 2, where $X \in \mathbf{R}^{C\times H\times W}$ is the input feature map, and C is the channel number of $\boldsymbol{X}$. The SAM energy function is shown in Eq. (1), which is derived in [11] detailly:

$$e_t^* = \frac{4\left(\widehat{\sigma}^2 + \lambda\right)}{\left(t - \widehat{\mu}\right)^2 + 2\widehat{\sigma} + 2\lambda} \tag{1}$$

$$\widehat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{2}$$

$$\widehat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}\left(x_i - \widehat{\mu}\right)^2 \tag{3}$$

In Eq. (1), t is an objective neuron in one of the channels of input feature $\boldsymbol{X}$, and x_i denotes other neurons in this channel, and i is the index of spatial dimensions. $M = H \times W$ denotes the number of neurons in the spatial dimension, and λ is a hyper-parameters. $\widehat{\mu}$ and $\widehat{\sigma}^2$ denote the mean and variance of all neurons in the channel, as shown in Eq. (2) and Eq. (3), respectively. The above formula indicates that, the smaller e_t^*, the larger t. That is, t is more pronounced than other neurons, and the network can thus learn more features.

Integrate the attention weights of each dimension with input features, and the final feature representation is shown in Eq. (4):

$$\tilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X \tag{4}$$

where, $\tilde{X}$ denotes the output feature, and E groups all e_t^* across channel and spatial dimensions. Sigmoid is added to restrict too large value in E.

SAM attention module is very flexible and modular, which is combined with the ResNet-50 network in this paper, and in order to adapt to the classification task of pedestrian Re-ID, a temporal average pooling layer is added at the end of the network to extract sequence-level features.

From Fig. 3, it can be seen that SAM attention module can be inserted into each layer, from Layer1 to Layer4. In order to extract sequence-level features, a time average pooling layer is added after Layer4 and SAM modules. With these SAM modules, many valuable clues can be captured, and the discriminative ability of the model can be enhanced, but the computational cost has not increased significantly, due to its lightweight structure.

3.3 Reciprocal Nearest Neighbor Metric

Label estimation is currently the main challenge in semi-supervised Re-ID. How to assign correct pseudo-labels for unlabeled sample plays a crucial role in model training. The nearest neighbor method with Euclidean distance metric is commonly used to assign

pseudo-labels. Equation (5) represents the Euclidean distance $d(v_i, v_j)$ between labeled sample v_i and unlabeled sample v_j:

$$d(v_i, v_j) = \|v_i - v_j\| \tag{5}$$

According to Formula (5), unlabeled data is assigned the label of the labeled data which is closest to it. However, in the process of label allocation, it is inevitable that unlabeled data is mislabeled (assigned error pseudo-label), especially when the discriminant power of the model is not strong enough. With the nearest neighbor method, when the number of mislabeled samples is large, it can greatly weaken the performance of the model. To address this problem, *k*-reciprocal encoding is used in our model for label estimation. In the field of pedestrian Re-ID, *k*-reciprocal encoding is initially used for re-ranking.

In this paper, reciprocal encoding is applied as the distance metric for one-shot label estimation of unlabeled samples. Assuming u_i is an unlabeled data, $G = \{g_i | i = 1,2,\ldots, N\}$ denotes N labeled data, and the *k*-nearest neighbor of u_i is defined as $N(u_i,k)$ as shown in Eq. (6):

$$N(u_i, k) = \{g_1^0, g_2^0, \ldots, g_k^0\}, |N(u_i, k)| = k \tag{6}$$

where, |.| denotes the number of candidates in the set. Based on the *k*-nearest neighbor of u_i, the *k*-reciprocal nearest neighbor of u_i is denoted as $R(u_i,k)$:

$$R(u_i, k) = \{g_i | (g_i \in N(u_i, k)) \bigwedge (u_i \in N(g_i, k))\} \tag{7}$$

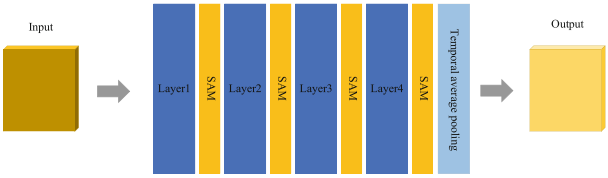


Fig. 3. Network structure diagram with SAM module

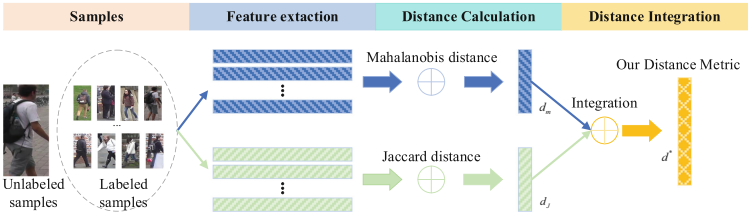


Fig. 4. Reciprocal nearest neighbour metric

As shown in Eq. (7), an unlabeled sample and a labeled sample are called *k*-reciprocal nearest neighbors, when they are *k*-nearest neighbors of each other. Compared with *k*-nearest neighbors, *k*-reciprocal nearest neighbors reflect the relationships of samples

better. At the procedure of label assignment, when an unlabeled sample and a labeled sample are k-reciprocal nearest neighbors, it can be considered that the likelihood of them belonging to the same category is very high. Based on this assumption, a reciprocal nearest neighbor metric is proposed for label estimation in this paper, as shown in Fig. 4.

It can be seen from Fig. 4 that the reciprocal nearest neighbor metric consists of three components:

(1) extract the features x_{u_i} and x_{g_i} of unlabeled data u_i and labeled data g_i, respectively,
(2) calculate the Mahalanobis distance $d_M(u_i, g_i)$ and Jaccard distance $d_J(u_i, g_i)$, as shown in Eq. (8) and Eq. (9),

$$d_M(u_i, g_i) = \sqrt{(u_i - g_i)^T \sum^{-1} (u_i - g_i)} \tag{8}$$

$$d_J(u_i, g_i) = 1 - \frac{|R(u_i, k) \cap R(g_i, k)|}{|R(u_i, k) \cup R(g_i, k)|} \tag{9}$$

In Eq. (8), $\sum$ is covariance matrix of u_i and g_i, the superscripts T and -1 denote matrix transpose and inverse. In Eq. (9), $R(\cdot, k)$ denotes the k-nearest neighbor of $\cdot$, and $|$ and $\|$ and $|\cdot|$ denotes the number of samples in $\cdot$. If sample u_i and g_i belong to the same category, there are many same samples in $R(u_i, k)$ and $R(g_i, k)$. This means that, the more the number of same samples in $R(u_i, k)$ and $R(g_i, k)$, the more similar of u_i and g_i, and the closer $d_J(u_i, g_i)$ tends to zero.

(1) Our reciprocal nearest neighbor metric $d^*(u_i, g_i)$ is defined with Mahalanobis distance $d_M(u_i, g_i)$ and Jaccard distance $d_J(u_i, g_i)$ as follow,

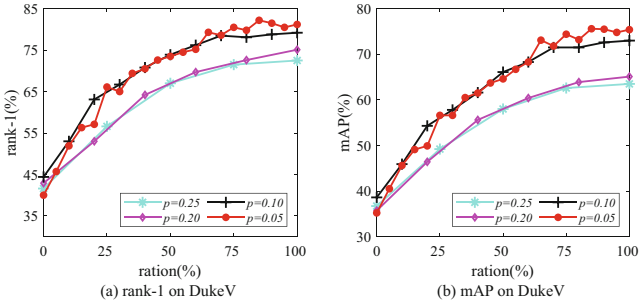
$$d^*(u_i, g_i) = (1 - \lambda) d_M(u_i, g_i) + \lambda d_J(u_i, g_i) \tag{10}$$

where, $\lambda (0 < \lambda < 1)$ is balance factor. With this metric, we can mine the information of reciprocal nearest neighbors between unlabeled and labeled data, and the accuracy of pseudo-label estimation and the robustness of the model can be improved.

4 Experiments and Results

4.1 Datasets and Evaluation Indicators

In our experiments, two mainstream large-scale datasets of video pedestrian Re-ID, MARS [25] and DukeMTMC-VideoReID [20], are used. In MARS, there are 20,478 video clips captured by 6 cameras, in which 17,503 are valid clips and the rest 3,248 are interference clips. The total number of pedestrians present in MARS is 1,261, and 625 in the training set and 636 in the test set. In DukeMTMC-VideoReID, there are 1,812 pedestrians and 4,832 video clips, wherein 702 pedestrians and 2196 clips in the training set, and 702 pedestrians and 2636 clips in the test set, in additional 408 interference pedestrians, which is the subset of DukeMTMC [26]. To demonstrate the performance of our model, two indicators, Cumulative Matching Characteristic (CMC) and Mean Average Precision (mAP), are used as evaluation indicators.



(a) rank-1 on DukeV (b) mAP on DukeV

Fig. 5. Comparison of results for different p of the DukeMTMC-VideoReID

4.2 Experimental settings

In this paper, GPU 2080ti is used and the experimental settings are the same as [20]. The stochastic gradient descent (SGD) optimization method with momentum of 0.5 and weight attenuation of 0.0005 is adopted. The overall learning rate is initialized to 0.1 and decays to 0.01 in the 15 epochs. The loss function is Cross Entropy Loss.

4.3 Experimental Results and Comparison

Parameter Selection Experiments

In this section, we analyze the parameter $p(0 < p < 1)$, which control the number of pseudo-label samples selected each time. Figure 5 shows the results with $p = 0.05$, $p = 0.10$, $p = 0.20$, $p = 0.25$ on the DukeMTMC-VideoReID dataset, respectively.

As shown in Fig. 5, it can be seen that, the smaller p selected, the better performance achieved. This is because fewer reliable pseudo-label samples are selected when p is small. Although there may be mislabeled samples contained in those selected pseudo-label samples, the number is too small to impact the performance of our model. But we also find that, p is not necessarily as small as possible. Because the fewer samples selected each time, the higher consumption in model training. In order to achieve a balance between speed and accuracy, we set $p = 0.05$ and $p = 0.10$ in the following experiments.

In addition to parameter p, k and λ can also affect the result of reciprocal nearest neighbor metric (RNM). Table 1 and Table 2 show the RNM results for different k values and different λ values on the DukeMTMC-VideoReID dataset, respectively.

As shown in Table 1, we set $p = 0.10$, $\lambda = 0.3$ and select $k = 5$, 10, 15, 20 for comparison In our experiments. Our RNM model achieves the best score in rank-1, rank-5, rank-20 and mAP, when $k = 10$. In the following experiments, we set $k = 10$.

As shown in Table 2, we set $p = 0.10$, $k = 10$ and select λ=0.3, 0.5 for comparison In our experiments. Our RNM model achieves the best score in rank-1, rank-5, rank-20 and mAP, when $\lambda = 0.3$. In the following experiments, we set $\lambda = 0.3$.

Effectiveness Experiments of SAM Attention Module

In this section, effectiveness of SAM attention module embedded in our model is compared with EGU [20] and PL [27] on DukeMTMC-VideoReID and MARS datasets.

Table 1. Comparison of RNM (%) for different k on DukeMTMC-VideoReID

k	rank-1	rank-5	rank-20	mAP
20	71.70	85.30	92.30	63.60
15	78.50	91.60	95.70	72.30
10	**79.20**	**92.20**	**95.40**	**73.00**
5	77.10	89.00	94.60	69.40

Table 2. Comparison of RNM (%) for different λ on DukeMTMC-VideoReID

λ	rank-1	rank-5	rank-20	mAP
0.30	**79.20**	**92.20**	**95.40**	**73.00**
0.50	75.40	88.00	93.90	67.50

As shown in Table 3, with SAM($p = 0.10$), rank-1 and mAP of our model achieves 75.6% and 67.6%, respectively, on DukeMTMC-VideoReID dataset, which exceeds the performance of EUG ($p = 0.05$) and PL ($p = 0.05$). Compared to EUG ($p = 0.10$), rank-1 and mAP of our model exceed 4.81% and 5.84%, respectively. Compared to PL ($p = 0.10$), rank-1 and mAP of our model exceed 4.6% and 5.7%, respectively. Similar results can also be obtained, when compared SAM ($p = 0.10$) with EUG ($p = 0.05$) and PL ($p = 0.05$).

As shown in Table 4, with SAM ($p = 0.10$), rank-1 and mAP of our model achieves 60.7% and 41.5%, respectively, on MARS dataset, which exceeds the performance of EUG ($p = 0.10$) and PL ($p = 0.10$), lower than the performance of EUG ($p = 0.05$) and PL ($p = 0.05$). Similarly,with SAM ($p = 0.05$), rank-1 and mAP of our model achieves 63.7% and 43.9%, respectively, which exceed the performance of EUG ($p = 0.05$) and PL ($p = 0.05$).

Table 3. Comparison(%) of SAM with EUG and PL on DukeMTMC-VideoReID

Methods	rank-1	rank-5	rank-20	mAP
EUG [20] ($p = 0.10$)	70.79	83.61	89.60	61.76
EUG [20] ($p = 0.05$)	72.79	84.18	91.45	63.23
PL [27] ($p = 0.10$)	71.00	83.80	90.30	61.90
PL [27] ($p = 0.05$)	72.90	84.30	91.40	63.30
SAM ($p = 0.10$)	**75.60**	**87.60**	**92.00**	**67.60**
SAM ($p = 0.05$)	**77.60**	**89.60**	**94.40**	**69.20**

Table 4. Comparison(%) of SAM with other methods on MARS

Methods	rank-1	rank-5	rank-20	mAP
EUG [20] ($p = 0.10$)	57.62	69.64	78.08	34.68
EUG [20] ($p = 0.05$)	62.67	74.94	82.57	42.45
PL [27] ($p = 0.10$)	57.90	70.30	79.30	34.90
PL [27] ($p = 0.05$)	62.80	75.20	83.80	42.60
SAM ($p = 0.10$)	**60.70**	**74.00**	**81.90**	**41.50**
SAM ($p = 0.05$)	**63.70**	**76.00**	**82.30**	**43.90**

With above experiments, it can be seen that, because of the flexible and lightweight characteristics of the SAM attention module, our model can effectively captures hidden information in video data and extracts more discriminative features, without increasing network parameters and without modifying the overall structure significantly, because of the flexible and lightweight characteristics of the SAM attention module.

Effectiveness Experiments of Reciprocal Nearest Neighbor Metric
In this section, effectiveness of RNM (reciprocal nearest neighbor metric) used in our model is compared with EGU [20] and PL [27] on DukeMTMC-VideoReID and MARS datasets.

As shown in Table 5, with RNM ($p = 0.10$), rank-1 and mAP of our model achieves 79.2% and 73.0%, respectively, on DukeMTMC-VideoReID dataset, which exceeds the performance of EUG ($p = 0.05$) and PL ($p = 0.05$). Compared to EUG ($p = 0.10$), rank-1 and mAP of our model exceed 8.41% and 11.24%, respectively. Compared to PL ($p = 0.10$), rank-1 and mAP of our model exceed 8.2% and 11.1%, respectively. Similar results can also be obtained, when compared RNM ($p = 0.10$) with EUG ($p = 0.05$) and PL ($p = 0.05$).

As shown in Table 6, with RNM ($p = 0.10$), rank-1 and mAP of our model exceed the performance of EUG ($p = 0.05$) and PL ($p = 0.05$) on MARS dataset. Compared to EUG ($p = 0.10$), rank-1 and mAP of our model exceed 7.18% and 8.52%, respectively. Compared to PL ($p = 0.10$), rank-1 and mAP of our model exceed 6.9% and 8.3%, respectively. Similar results can also be obtained, when compared RNM ($p = 0.05$) with EUG ($p = 0.05$) and PL ($p = 0.05$).

In RNM, it is assumed that, two samples that are k-nearest to each other are highly likely to belong to the same category. Experiments have shown that, compared to the commonly used Euclidean distance, RNM makes labeled data and pseudo-labeled data to connect much more closer to each other, which can improve the accuracy of pseudo-label estimation.

Ablation Experiments
In ablation experiments, in order to demonstrate the effectiveness of SAM and RNM, we use EUG as baseline and the framework used in our model is the same as that in EUG.

Table 5. Comparison(%) of RNM with other methods on DukeMTMC-VideoReID

Methods	rank-1	rank-5	rank-20	mAP
EUG [20] ($p = 0.10$)	70.79	83.61	89.60	61.76
EUG [20] ($p = 0.05$)	72.79	84.18	91.45	63.23
PL [27] ($p = 0.10$)	71.00	83.80	90.30	61.90
PL [27] ($p = 0.05$)	72.90	84.30	91.40	63.30
RNM ($p = 0.10$)	**79.20**	**92.20**	**95.40**	**73.00**
RNM($p = 0.05$)	**81.20**	**92.70**	**95.60**	**75.40**

Table 6. Comparison (%) of RNM with other methods on MARS

Methods	rank-1	rank-5	rank-20	mAP
EUG [20] ($p = 0.10$)	57.62	69.64	78.08	34.68
EUG [20] ($p = 0.05$)	62.67	74.94	82.57	42.45
PL [27] ($p = 0.10$)	57.90	70.30	79.30	4.90
PL [27] ($p = 0.05$)	62.80	75.20	83.80	42.60
RNM ($p = 0.10$)	**64.80**	**79.50**	**86.90**	**43.20**
RNM ($p = 0.05$)	**66.30**	**80.40**	**87.50**	**44.80**

As shown in Table 7, when $p = 0.10$, rank-1 of SAM, RNM and SAM + RNM are 75.6%, 79.2% and 80.5%, and mAP of them are 67.6%, 73% and 74.4%, respectively. All of the scores exceed the scores with EUG. Similar results can be obtained, when compared SAM, RNM and SAM + RNM with EUG. It can be seen that, performance improvements are achieved on DukeMTMC-VideoReID dataset, whether using SAM, RNM or SAM + RNM, which demonstrates the effectiveness of our method.

As shown in Table 8, when $p = 0.10$ and 0.05, SAM, RNM and SAM + RNM are better than EUG, which demonstrates the effectiveness of our method.

Generally speaking, both SAM and RNM outperform EUG on both datasets, which demonstrates the effectiveness of SAM and RNM. Specifically, when using both SAM and RNM simultaneously, the performance of our model achieves an better performance than using SAM or RNM alone. With above experiments, we can see that, compared SAM with RNM, the latter contributes more to the network. This is because RNM acts on the process of label estimation directly, and significantly improves the accuracy of pseudo-label assignment, which reduces the proportion of erroneous samples in the model and helps to generate a robust model. SAM, on the other hand, focuses on extracting hidden features from video data and relies on models with strong discriminative power to play its role.

Table 7. Ablation Experiments on DukeMTMC-VideoReID

Methods	rank-1	rank-5	rank-20	mAP
EUG [20] ($p = 0.10$)	70.79	83.61	89.60	61.76
SAM ($p = 0.10$)	75.60	87.60	92.00	67.60
RNM ($p = 0.10$)	79.20	92.20	95.40	73.00
SAM + RNM ($p = 0.10$)	**80.50**	**92.70**	**96.20**	**74.40**
EUG [20] ($p = 0.05$)	72.79	84.18	91.45	63.23
SAM ($p = 0.05$)	77.60	89.60	94.40	69.20
RNM ($p = 0.05$)	81.20	92.70	95.60	75.40
SAM + RNM ($p = 0.05$)	**82.20**	**92.50**	**96.40**	**75.60**

Table 8. Ablation Experiments on MARS

Methods	rank-1	rank-5	rank-20	mAP
EUG [20] ($p = 0.10$)	57.62	69.64	78.08	34.68
SAM ($p = 0.10$)	60.70	74.00	81.90	41.50
RNM ($p = 0.10$)	64.80	79.50	86.90	43.20
SAM + RNM ($p = 0.10$)	**66.80**	**80.90**	**87.70**	**45.80**
EUG [20] ($p = 0.05$)	62.67	74.94	82.57	42.45
SAM ($p = 0.05$)	63.70	76.00	82.30	43.90
RNM ($p = 0.05$)	66.30	80.40	87.50	44.80
SAM + RNM ($p = 0.05$)	**67.60**	**81.20**	**88.70**	**47.90**

Comparison with Other Advanced Methods

In order to demonstrate the superiority of our method, a large number of experiments are conducted on two commonly used large datasets in this section to compare it with the current advanced one-shot video-based pedestrian Re-ID methods, including Stepwise [19], EUG [20], DGM + IDE [21], SCLU [22], LGF [23], PL [27], and BUC [28]. Among them, Baseline (one shot) [20] indicates that only the initial labeled samples are used for training, without any progressive learning methods, and Supervised [20] indicates that all labeled samples are used for training the baseline. From Table 9, it can be seen that the method SAM + RNM proposed in this paper achieves significant performance improvement compared to the Baseline (one shot) and unsupervised method BUC (only use single labeled samples for training), and also surpasses traditional semi-supervised methods such as DGM + IDE and Stepwise. Compared with some one-shot methods, including EUG, SCLU, and PL, the performance of our SAM + RNM method exceeds all above methods, which demonstrate the effectiveness of the proposed method.

In summary, the method proposed in this paper has achieved excellent performance on DukeMTMC-VideoReID and MARS datasets, which indicates that our model, with

Table 9. Comparison with Some Advance Methods on Large-Scale Datasets

Methods	MARS				DukeMTMC-VideoReID			
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
Baseline(one-shot) [20]	36.20	50.20	61.90	15.50	39.60	56.80	67.00	33.30
Supervised [20]	80.8	92.1	96.1	63.7	83.6	94.6	97.6	78.3
Stepwise [19]	41.20	55.60	66.80	19.70	56.30	70.40	79.20	46.80
EUG($p = 0.10$) [20]	57.62	69.64	78.08	34.68	70.79	83.61	89.60	61.76
EUG($p = 0.05$) [20]	62.67	74.94	82.57	42.45	72.79	84.18	91.45	63.23
DGM + IDE [21]	36.80	54.00	68.50	16.90	42.40	57.90	69.30	33.60
SCLU($p = 0.10$) [22]	61.97	76.52	84.34	41.47	72.79	84.19	91.03	62.99
SCLU($p = 0.05$) [22]	63.74	78.44	85.51	42.74	72.79	85.04	90.31	63.15
LGF [23]	58.80	69.00	78.50	36.20	86.30	96.00	98.60	82.70
PL($p = 0.10$) [27]	57.90	70.30	79.30	34.90	71.00	83.80	90.30	61.90
PL($p = 0.05$) [27]	62.80	75.20	83.80	42.60	72.90	84.30	91.40	63.30
BUC [28]	55.10	68.30	-	29.40	74.80	86.80	-	66.70
SAM + RNM($p = 0.10$)	**66.80**	**80.90**	**87.70**	**45.80**	**80.50**	**92.70**	**96.20**	**74.40**
SAM + RNM($p = 0.05$)	**67.60**	**81.20**	**88.70**	**47.90**	**82.20**	**92.50**	**96.40**	**75.60**

SAM and RNM, can effectively improve the performance of feature extraction, reduce the mislabel in pseudo-label assignment, and enhance the robustness and discrimination.

5 Conclusion

In this paper, we focus on the problem of one-shot video-based pedestrian Re-ID. To address the issue of insufficient labeled data under this one-shot settings, SAM attention module is embedded in the network to improve the ability of feature extraction and the discriminative power. In order to further improve the accuracy of label prediction, we design RNM, which can assign pseudo-labels for unlabeled samples more accurately. The excellent performance on two large-scale datasets proves that, the effectiveness of our method. But it should be noted that, although our method proposed in this paper achieves competitive performance, there are still issues with incorrect label assignment and long training time in the later stage, which require further exploration.

References

1. Liu, J., Zha, Z.J., Wu, W., et al.: Spatial-temporal correlation and topology learning for person re-identification in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, pp. 4370–4379 (2021)

2. Zhang, Y.P., Wang, H.Y., Zhang, J., et al.: One-shot Video-based Person Re-identification Based on Neighborhood Center Iteration Strategy. J. Softw. **32**(12), 4025–4035 (2021)
3. Dai, C.C., Wang, H.Y., Ni, T.G., et al.: Person Re-Identification Based on Deep Convolutional Generative Adversarial Network and Expanded Neighbor Reranking. Comput. Res. Dev. **56**(8), 1632–1641 (2019)
4. Ni, T., Gu, X., Wang, H., et al.: Discriminative deep transfer metric learning for cross-scenario person re-identification. J. Electron. Imaging **27**(4), 043026 (2018)
5. Wang, H., Ding, Z., Zhang, J., et al.: Person reidentification by semisupervised dictionary rectification learning with retraining module. J. Electron. Imaging **27**(4), 043043 (2018)
6. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, pp. 608–617 (2019)
7. Zhang, Z., Lan, C., Zeng, W., et al.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, pp. 667–676 (2019)
8. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future[EB/OL]. (2016–10–10)[2022–06–11]. https://arxiv.org/pdf/1610.02984.pdf
9. Ye, M., Shen, J., Lin, G., et al.: Deep learning for person re-identification: a survey and outlook. In: IEEE Transactions on Pattern Analysis and Machine Intelligence: 1–1 (2021)
10. Navaneet, K.L., Todi, V., Babu, R.V., et al.: All for one: Frame-wise rank loss for improving video-based person re-identification. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, pp. 2472–2476 (2019)
11. Yang, L., Zhang, R.Y., Li, L., et al.: Simam: A simple, parameter-free attention module for convolutional neural networks. In:International Conference on Machine Learning. New York: ICML, pp. 11863–11874 (2021)
12. Zhong, Z., Zheng, L., Cao, D., et al.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, pp. 1318–1327 (2017)
13. Li, S., Bak, S., Carr, P., et al.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, pp. 369–378 (2018)
14. Hou, R., Ma, B., Chang, H., et al.: Vrstc: Occlusion-free video person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, pp. 7183–7192 (2019)
15. Li, J., Wang, J., Tian, Q., et al.: Global-local temporal representations for video person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, pp. 3958–3967 (2019)
16. Liu, C.T., Wu, C.W., Wang, Y.C.F., et al.: Spatially and temporally efficient non-local attention network for video-based person re-identification[EB/OL]. (2019–08–05)[2022–06–11]. https:// arxiv.org/ pdf/1908.01683. pdf
17. Eom, C., Lee, G., Lee, J., et al.: Video-based Person Re-identification with Spatial and Temporal Memory Networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, pp. 12036–12045 (2021)
18. Zhu, X., Jing, X.Y., Yang, L., et al.: Semi-supervised cross-view projection-based dictionary learning for video-based person re-identification[J]. IEEE Trans. Circuits Syst. Video Technol. **28**(10), 2599–2611 (2017)
19. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, pp. 2429–2438 (2017)

20. Wu, Y., Lin, Y., Dong, X., et al.: Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, pp. 5177–5186 (2018)
21. Ye, M., Li, J., Ma, A.J., et al.: Dynamic graph co-matching for unsupervised video-based person re-identification. IEEE Trans. Image Process. **28**(6), 2976–2990 (2019)
22. Yin, J., Li, B., Wan, F., et al.: A new data selection strategy for one-shot video-based person re-identification. In: 2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, pp. 1227–1231 (2019)
23. Zhao, C., Zhang, Z., Yan, J., et al.: Local-global feature for video-based one-shot person re-identification. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, pp. 3662–3666 (2020)
24. Liu, M., Qu, L., Nie, L., et al.: Iterative local-global collaboration learning towards one-shot video person re-identification. IEEE Trans. Image Process. **29**, 9360–9372 (2020)
25. Zheng, L., Shen, L., Tian L, et al. Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision. Piscataway: IEEE, pp. 1116–1124 (2015)
26. Ristani, E., Solera, F., Zou, R., et al.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision, pp. 17–35. Springer, Berlin (2016)
27. Wu, Y., Lin, Y., Dong, X., et al.: Progressive learning for person re-identification with one example. IEEE Trans. Image Process. **28**(6), 2872–2881 (2019)
28. Lin, Y., Dong, X., Zheng, L., et al.: A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 33(01), pp. 8738–8745 (2019)