

One-Shot Video-Based Person Re-Identification by Global-Local Feature and Subsampling Strategy

Shuang Huang¹, Yang Li²

¹ Shanghai University, hs99977@shu.edu.cn, Shanghai, 200444, China

² Shanghai University, liyang162@shu.edu.cn, Shanghai, 200444, China

Abstract

This paper focuses on the video-based person re-identification task under the one-shot setting, and proposes a method based on global-local feature and subsampling strategy. In order to obtain more discriminative features of people for pseudo-label estimation, global features and local features are considered together, and multiple loss functions are used to optimize the model, which improves the discriminant power of the model. In addition, a more reasonable method, the subsampling strategy, is proposed for pseudo-label estimation. By dividing the pseudo-labeled dataset into two subsets with proper proportions, the influence of false labeled samples on model training is reduced. A large number of experiments have been carried out on MARS and DukeMTMC-VideoReID, which are commonly used public video datasets. Experimental results show that our method achieves better performance than the most advanced methods, which verifies the superiority of this method.

Keywords

One-shot, video-based person re-identification, global-local feature, subsampling

1. Introduction

Person re-identification (Re-ID) is a sub-problem of image retrieval, which mainly uses computer vision technology to judge whether there is a specific person in an image or video sequence. In recent years, with the continuous development of intelligent monitoring field, person Re-ID technology has gradually attracted wide attention [1, 2]. The current person Re-ID tasks can be mainly divided into two categories, one is image-based person Re-ID, the other is video-based person Re-ID. Compared with images, video-based person Re-ID can be said to be a more natural way, because in real life, cameras usually capture video sequences, and video sequences also provide richer and more detailed features, making person modeling more accurate.

With the rise of deep neural networks [3], video-based person Re-ID has also made substantial progress. However, most of the existing methods are supervised learning methods, means they need to rely on a large number of labeled data, and datasets for labeling is a very expensive and costly undertaking, especially the correlation between multiple cameras person data, note the high cost of data is also supervised method is difficult to extend to the real application of one of the important reasons. Therefore, many scholars also begin to pay attention to semi-supervised video-based person Re-ID [4, 5].

Semi-supervised methods mainly use a small amount of labeled data to conduct network learning. This paper considers one of the semi-supervised methods based on one-shot setting. For one-shot video-based person Re-ID, only one labeled video tracklet is available for each identity, and the rest are unlabeled data. In view of the few available labeled samples and the complex information contained in video sequence, how to effectively use video sequence information for feature extraction and how to assign pseudo labels to unlabeled data have become key challenges. In recent years, there has been a

ISCIPT2022@7th International Conference on Computer and Information Processing Technology, August 5-7, 2022, Shenyang, China

EMAIL: hs99977@shu.edu.cn (Shuang Huang); liyang162@shu.edu.cn (Yang Li);

ORCID: 0000-0002-2934-5104 (Shuang Huang); 0000-0003-4915-613X (Yang Li);

© 2022 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)



lot of works based on one-shot methods [6, 7, 8, 9, 10], however most of these methods rely on global feature representation while ignoring local feature description. Only fuzzy global feature representation is likely to produce incorrect label estimation, which will have more and more influence on model training as the iteration goes on, and discriminability information is likely to be hidden in some local information in the video tracklet. In addition, for better utilization of unlabeled data, Wu et al. [8] propose a strategy of progressive sampling, in which the most reliable pseudo-labeled data is gradually selected for training. However, for one-shot tasks, there are few initial labeled data, so there are still a large number of incorrect labeled data in the pseudo-labeled data selected by this progressive sampling strategy, which affects the training of subsequent models.

In order to solve the above problems, this paper proposes a method based on global-local feature and subsampling strategy. Specifically, we adopt a progressive learning framework. When extracting features, global features and local features are taken into account, and the sample data obtained by CNN is more discriminative in global-local feature representation. In order to improve the discriminability of acquired features and enhance the discriminability of the model, we introduce the center loss function [11], minimize intra-class differences, and combine cross-entropy loss to perform identity classification tasks. In addition, exclusive loss [9] is adopted to distance unlabeled samples from each other in the feature space and reduce the influence of irrelevant features on the model. Finally, a subsampling strategy is proposed to select reliable pseudo-labeled data, which is divided into two subsets with appropriate proportions. A more reliable pseudo-labeled data subset is selected and incorporated into the training set to re-train the model, ensuring the robustness of model training.

The main contributions of this paper are summarized as follows:

- In this paper, a method based on global-local feature learning is proposed, which considers the influence of global-local feature jointly, and adopts multiple loss functions to optimize different sample data, which effectively improves the discriminant ability of the model.
- A subsampling strategy is proposed, in which the pseudo-labeled samples are divided into two subsets with proper proportion when the pseudo-labeled samples are selected dynamically, so as to reduce the influence of the false labeled samples on model training and improve the robustness of the model.
- A large number of experiments on MARS and DukeMTMC-VideoReID, two publicly available video datasets, show that our proposed method has better performance than other advanced methods.

2. Related Works

2.1. Supervised Video-Based Person Re-ID

In recent years, with the rise of deep learning [12], it has been widely used in the field of image recognition. Video-based person Re-ID has achieved great success and its performance has been significantly improved. In 2016, McLaughlin et al. [13] apply deep learning theory to video-based person Re-ID task for the first time, and a new recurrent neural network structure is proposed, which captures features from each video and then uses recurrent layer and temporal pooling layer to extract video-level features. Since then, video-based person Re-ID based on deep learning has developed rapidly and achieved remarkable results in various datasets. For the problem of occlusion in person Re-ID task, Li et al. [14] design a spatio-temporal attention model with diversity regularization. The model uses multiple spatial attention models to locate and discriminate image regions, and then collects these local features across time through temporal attention. Yang et al. [15] propose a novel spatio-temporal graph convolutional network to model the potential relationships between different parts of the human body within and across the same frame, providing more discriminative and robust information for Re-ID, and effectively overcoming the problem of occlusion and visual ambiguity.

2.2. Semi-Supervised Video-Based Person Re-ID

In order to solve the problems of high cost of dataset annotation and difficult application of actual scenes, many researchers also begin to pay attention to semi-supervised video-based person Re-ID.

Compared with supervised video-based person Re-ID, there are few studies on semi-supervised video-based person Re-ID, but in recent years, with the attention of a large number of researchers, the field of semi-supervised video-based person Re-ID has made great progress.

Ye et al. [6] propose a dynamic graph matching method in 2017, which iteratively updates image graph matching and label estimation to learn better features. Liu et al. [7] update the classifier with K-reciprocal Nearest Neighbors (KNN) in the gallery set, and refine the nearest neighbor by applying negative sample mining with KNN in the query set. It has to be mentioned that although the two methods in literature [6] and [7] claim that they are unsupervised, they require at least one labeled tracklet for each identity in the experiment, so they are regarded as a one-shot video-based person Re-ID task. Both methods adopt a static sampling strategy to determine the amount of selected pseudo-labeled data for further training, in which pseudo labels whose confidence score is higher than the predefined threshold are selected in each step, which will result in a large number of unreliable pseudo labels for training and hinder the improvement of model performance. To solve this problem, Wu et al. [8] propose a new progressive learning method. In this method, a dynamic sampling strategy is adopted to allocate pseudo labels, and the number of pseudo-labeled candidates is gradually increased, which significantly improves the performance of label estimation. On this basis, literature [9] propose a progressive method of joint learning, which divides training data into labeled data, pseudo-labeled data and index-labeled data. In the iterative process, CNN model is updated through joint training of these three parts of data, making full use of the unreliable unlabeled data of pseudo labels. Yin et al. [10] propose a method based on multi-loss learning and joint distance measurement to optimize different loss functions for different data and improve the discriminant power of the model. In label estimation, sample distance and nearest neighbor distance are considered together to further improve the accuracy of pseudo-label prediction.

Although the above methods achieve good performance, they all adopt temporal average pooling to obtain global feature representation of people during feature extraction, which may lead to incorrect label estimation and errors will accumulate with iteration. In addition, the existing methods usually use the cross-entropy loss function to classify identities, but the cross-entropy loss function only ensures that features of different categories can be separated. Therefore, in order to improve the distinguishing ability of acquired features, we consider the importance of local features, propose a learning method based on global-local features, introduce the center loss function into the identity classification task, and combine multiple loss functions to optimize the model. At the initial stage of model training, due to the small number of initial labeled samples, the model has over-fitting phenomenon. Even if the dynamic sampling strategy is adopted, there will still be a large number of mislabeled samples. Therefore, in order to reduce the influence of incorrectly labeled samples on subsequent training, we propose a subsampling strategy, which further divides the dynamically selected pseudo-labeled sample data into two subsets with appropriate proportions, and selects the more reliable pseudo-labeled sample data to be added to the training set for the next training.

3. Method

The method framework proposed by us is shown in Figure 1. In this paper, an iterative method is adopted to train the model. The specific process is as follows: First, the model is trained by using the method of global-local feature learning for labeled data, pseudo-labeled data I, pseudo-labeled data II and index-labeled data. It is worth noting that in the initial stage of model training, no label estimation has been carried out on the unlabeled data, so there are only labeled data and unlabeled data. We initialize the video-based Re-ID model according to the sample data with labels initially. Then, the method based on distance measurement is adopted in the feature space, that is, pseudo labels are assigned to each unlabeled data based on the distance between labeled data and unlabeled data, several reliable pseudo-labeled data are selected, and index labels are assigned to the unselected data as index-labeled data. Then, the subsampling strategy is adopted to divide the selected reliable pseudo-labeled data into pseudo-labeled data I and pseudo-labeled data II, and the pseudo-labeled data I is merged into the training set for the next training. In the iterative process, the size of the selected pseudo-labeled data is constantly expanded until all the unlabeled data is used for model training.

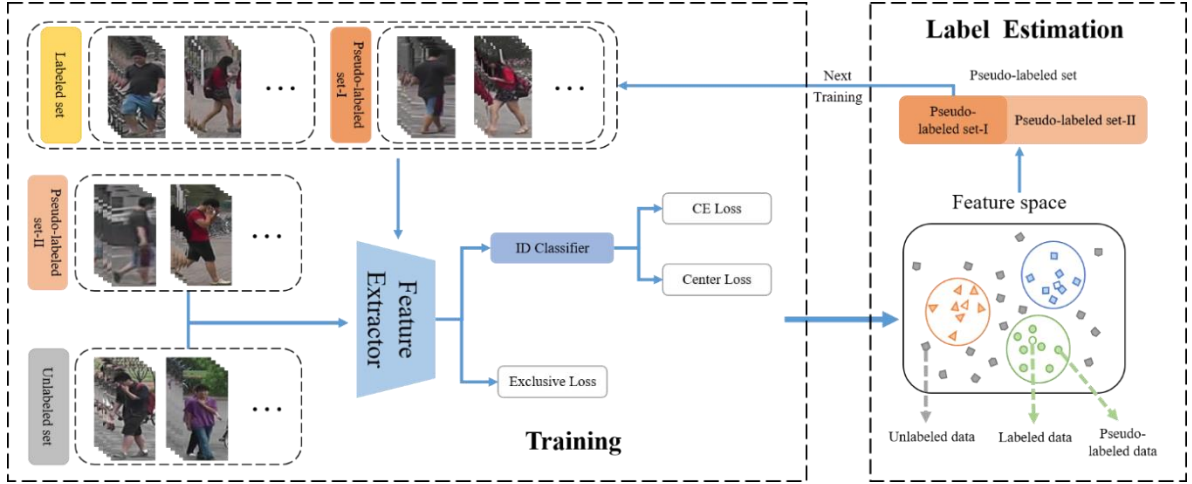


Figure 1: Frame diagram of our model

3.1. Problem Description

Under the one-shot setting, for n identities, the labeled dataset is defined as $L = \{(x_{l1}, y_{l1}), (x_{l2}, y_{l2}), \dots, (x_{ln}, y_{ln})\}$, and the unlabeled dataset is defined as $U = \{x_{u1}, x_{u2}, \dots, x_{un}\}$, where x represents the video tracklet, and y represents the label corresponding to the video tracklet. Our goal is to train feature extractor $\phi(\theta, \cdot)$ parameterized by θ using these sample data. In the t iteration, the training set contains four parts of data: labeled data L , pseudo-labeled data I P_t^I , pseudo-labeled data II P_t^{II} and index-labeled data I^t .

3.2. Global-Local Feature Learning

It is very important to obtain distinguishing feature representation for label estimation. In the previous work, the global features at the video level are usually obtained through temporal average pooling, without any fine grained local information. Therefore, this paper combines global features with local features and proposes a method based on global-local feature learning. The feature extraction module is shown in Figure 2.

For each video tracklet there are two branches: the global branch and the local branch. For global feature learning, frame-level feature embedding is first obtained through a CNN model, and then the frame-level feature is represented as the global feature of the video tracklet through global average pooling. Global features include the general features of people, but some fine grained local features are ignored. In the person Re-ID task, local information is also very effective, with strong discrimination. Therefore, we introduce local features to learn more discriminative information, and extract frame-level features from CNN model as local feature representation of video tracklets. Finally, the global-local feature representation is obtained by superposing the two level features.

We use labeled data, pseudo-labeled data I and pseudo-labeled data II for identity classification learning, and the cross-entropy loss function is as follows:

$$L_{CE} = -\frac{1}{n} \sum_{i=1}^n \log(p(y_i|x_i)), \quad (1)$$

where x_i and y_i are the input samples and their corresponding labels respectively, $p(\cdot)$ is the probability of correctly predicting sample labels, and n is the total number of sample data. It is worth mentioning that in identity classification, cross-entropy loss only ensures that features of different categories can be separated. Therefore, in order to improve the discernibility of acquired features, we introduce the center loss function [11] for identity classification, the center loss formula is as follows:

$$L_C = \frac{1}{2} \sum_{i=1}^m \|\phi(\theta; x_i) - c_{y_i}\|_2^2 \quad (2)$$

where c_{y_i} represents the grade y_i center of deep features. CNN model is trained jointly with cross-entropy loss and center loss, so the loss function of identity classification task is expressed as follows:

$$L_{ID} = L_{CE} + \alpha L_C \quad (3)$$

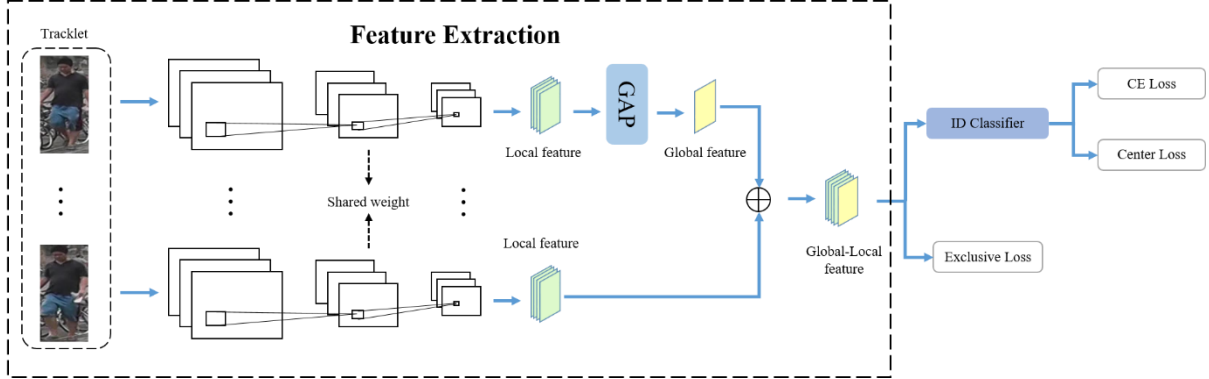


Figure 2: Feature extraction module

where α is the hyperparameter of balancing cross-entropy loss and center loss. For index-labeled data I^t , due to its unreliable pseudo-label, we hope to reduce its impact on model performance. Therefore, exclusive loss [9] is used to keep index-labeled data in feature space away from other data, so as to avoid irrelevant features learned by the model and further optimize CNN model. Exclusive loss is expressed as follows:

$$L_{Ex} = -\log \frac{\exp(v_i^T \phi(\theta; x_i)/\tau)}{\sum_{j=1}^{|I^t|} \exp(v_j^T \phi(\theta; x_i)/\tau)} \quad (4)$$

where $v_i = \phi(\theta; x_i)x_i$ is the L2 normalized feature embedding of samples, and the hyperparameter τ is mainly used to control distribution. We combine the two loss functions to train the model, and the final objective function is as follows:

$$L_{total} = \min_{\theta} \beta L_{ID} + (1 - \beta) L_{Ex} \quad (5)$$

The hyperparameter β is used to adjust the contribution of identity loss and exclusive loss.

3.3. Subsampling Strategy

Label estimation is undoubtedly a big challenge for one-shot video-based person Re-ID task, that is, how to assign correct pseudo labels to the rich unlabeled data, and the accuracy of pseudo labels will greatly affect the performance of the model. The early works [16] are to assign pseudo labels to unlabeled data through classification prediction model. However, for one-shot tasks, the classifier is easy to over-fit and cannot accurately predict the categories of unlabeled data. To this end, we adopt the distance-based nearest neighbor method in literature [9] for label estimation, as shown below:

$$d(\theta; x_u) = \min \|\phi(\theta; x_u) - \phi(\theta; x_l)\| \quad (6)$$

where x_u and x_l represent labeled data and unlabeled data respectively, and $\|\cdot\|$ represents Euclidean Distance. Specifically, the confidence degree of label estimation is defined as the Euclidean Distance between unlabeled data and labeled data in the feature space, and then pseudo labels are assigned to unlabeled data according to the labeled data nearest to it in the feature space. The enlarging factor $p \in (0,1)$ is set to dynamically select pseudo-labeled data with high confidence for the next training. In t -step iteration, the selected pseudo-labeled data set size is expressed as:

$$p^t = p^{t-1} + p \cdot U \quad (7)$$

Although the dynamic sampling strategy selecting pseudo-labeled data used to train the next step by step, to avoid the model training involves a large number of unreliable in the early years of the pseudo-labeled data, but in view of the one-shot task, each identity only one labeled video tracklet, the original labeled data is less, so the model is still easy to over-fit. Specifically, during label estimation, there are still a large number of false labeled samples in the pseudo-labeled dataset selected through the dynamic sampling strategy, which will inevitably affect the subsequent model training, and this influence will gradually increase with the iteration. Therefore, we propose a subsampling strategy based on the

dynamic sampling strategy, and divide the pseudo-labeled dataset selected by the dynamic sampling strategy into pseudo-labeled set I and pseudo-labeled set II, as shown below

$$P_I^t = \lambda P^t \quad (8)$$

$$P_{II}^t = (1 - \lambda)P^t \quad (9)$$

The hyperparameter λ is used to control the ratio of two sets. In order to select a suitable λ , we study the influence of the size of the hyperparameter λ on the model performance in Section 4.5.

4. Experiments

4.1. Datasets and Evaluation Metrics

MARS dataset [17] is the largest video dataset in the video-based person Re-ID task so far. It is mainly the person video tracklets in the university campus shot by 6 cameras, including 17,503 video tracklets, corresponding to 1263 identities, and 3248 interference clips. 625 identities are used for training and 636 for testing.

DukeMTMC-VideoReID dataset [8] is a subset of DukeMTMC, consisting of 4832 video tracklets of 1812 identities, including 2,196 video tracklets of 702 identities in the training set, 2363 video tracklets of 702 identities in the test set, and 408 interference items.

In this paper, Cumulative Matching Characteristic (CMC) and Mean Average Precision (mAP) are used to evaluate the effectiveness of the proposed method.

4.2. Implementation Details

All experiments in this paper are based on PyTorch framework, and we adopt basically the same experimental setup as literature [9]. We adopt ResNet-50 [18] with the last classification layer removed as our feature embedding model ϕ to conduct all the experiments and initialize it by the ImageNet [19] pre-trained model. α is set to 0.1 during the training phase when the model is initialized, and then to 0.001. The value of the hyperparameter τ in formula (4) is set to 0.1, β is set to 0.8 initially in formula (5) and 1 in the last 15 epochs. For label estimation, the value of λ is set to 0.4.

4.3. Comparison with the State-of-the-Art Methods

To verify the effectiveness of our method, we conduct experiments on MARS and DukeMTMC-VideoReID, two commonly used video datasets, and the experimental results are shown in Table 1. Among them, the baseline method does not use additional unlabeled data, but only uses one-shot labeled data for training. The proposed method achieved 69.8% rank-1 and 49.5% mAP (when $p = 0.05$) on MARS dataset, 1.3% and 1.7% higher than the most advanced method. In addition, the proposed method performs better on the DukeMTMC-VideoReID dataset, with rank-1 and mAP (when $p = 0.05$) reaching 87.2% and 84.5%, respectively, an improvement of 10.7% and 15.8% over the most advanced method, respectively.

Table 1

Comparison with the state-of-the-art methods on MARS and DukeMTMC-VideoReID datasets. Bold numbers are the best.

Method	MARS				DukeMTMC-VideoReID			
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
Baseline (one-shot) [11]	36.2	50.2	61.9	15.5	39.6	56.8	67.0	33.3
DGM+IDE [8]	36.8	54.0	68.5	16.9	42.4	57.9	69.3	33.6
Stepwise [9]	41.2	55.6	66.8	19.7	56.3	70.4	79.2	46.8
EUG [10] ($p = 0.10$)	57.62	69.64	78.08	34.68	70.79	83.61	89.60	61.76
EUG [10] ($p = 0.05$)	62.67	74.94	82.57	42.45	72.79	84.18	91.45	63.23
PL [11] ($p = 0.10$)	57.9	70.3	79.3	34.9	71.0	83.8	90.3	61.9
PL [11] ($p = 0.05$)	62.8	75.2	83.8	42.6	72.9	84.3	91.4	63.3

MLL+JDM [12] ($p = 0.10$)	65.5	78.5	86.6	44.2	76.2	87.2	93.3	67.5
MLL+JDM [12] ($p = 0.05$)	68.5	80.8	88.6	47.8	76.5	88.7	93.2	68.7
Our ($p = 0.10$)	66.4	79.1	86.2	45.6	86.5	96.2	97.9	83.3
Our ($p = 0.05$)	69.8	83.2	88.4	49.5	87.2	96.7	99.3	84.5

The experimental results show that the performance of the proposed method is better than the most advanced one-shot method, especially in DukeMTMC-VideoReID dataset. This may be because previous methods only focus on the global features of samples and ignore the implied local information, which often contains some distinguishing features. Secondly, MARS dataset is richer and more complex than DukeMTMC-VideoReID dataset, and there are more label noises in it, which affects the discriminant performance of the model.

4.4. Ablation Studies

For the two key parts of global-local feature learning and subsampling strategy proposed in this paper, we conduct ablation studies on MARS and DukeMTMC-VideoReID datasets, as shown in Table 2. Where "Ours w/o L" means that only global features of the sample are considered while local features are ignored during feature extraction. "Ours w/o S" means that the subsampling module is removed during label estimation, that is, only dynamic sampling is adopted. "Ours" represents the complete model presented in this paper.

Table 2

Ablation studies on MARS and DukeMTMC-VideoReID datasets.

Factor	Method	MARS		DukeMTMC-VideoReID	
		rank-1	mAP	rank-1	mAP
$p = 0.10$	Ours w/o L	65.3	44.0	79.1	73.2
	Ours w/o S	62.9	41.7	81.6	77.4
	Ours	66.4	45.6	86.5	83.3
$p = 0.05$	Ours w/o L	68.2	47.5	80.1	75.6
	Ours w/o S	66.1	45.3	82.4	78.9
	Ours	69.8	49.5	87.2	84.5

The effectiveness of global-local feature learning. Comparing our method with the global feature only method, it is found that when $p = 0.05$, our method improves rank-1 and mAP by 1.6% and 2.0%, respectively, on MARS and 7.1% and 8.9%, respectively, on DukeMTMC-VideoReID. It can be seen that the effect of global-local feature learning module on DukeMTMC-VideoReID dataset is better, mainly because there are more label noises in MARS dataset, which affects the effect of local feature extraction. In general, the comparison results on two datasets prove the effectiveness of the global-local feature learning method.

The effectiveness of the subsampling strategy. Comparing the experimental results of our method and the method of removing the subsampling module, no matter $p = 0.10$ or $p = 0.05$, the performance of our method is improved in both datasets, indicating that the subsampling strategy in label estimation can effectively reduce the impact of wrong labels on model performance compared with the dynamic sampling strategy.

In conclusion, the model performance is best when the two methods are used simultaneously on two datasets, and the global-local feature learning module plays a better role than the subsampling module in improving the performance of the whole model.

4.5. Algorithm Analysis

In this paper, we propose a subsampling strategy in the stage of label estimation, in which the hyperparameter λ determines the size of two subsets, how to select the appropriate partition ratio

becomes the key. Therefore, this paper compares the effects of different λ values on the model performance on the DukeMTMC-VideoReID dataset (where the enlarging factor $p = 0.05$), and the experimental results are shown in Table 3. It can be seen that with the increase of λ , rank-1 and mAP increase gradually. When λ is 0.4, rank-1 and mAP reach the maximum and the overall performance is the best. When λ continues to increase, rank-1 and mAP begin to decrease. Therefore, to get the best performance of the model, we set λ to 0.4.

Table 3

Comparison of different λ values on DukeMTMC-VideoReID dataset. Bold numbers are the best. (where the enlarging factor $p = 0.05$)

λ	rank-1	rank-5	rank-20	mAP
0.1	83.3	92.5	95.9	80.5
0.2	84.5	93.2	97.2	81.3
0.3	86.3	95.4	99.4	83.1
0.4	87.2	96.7	99.3	84.5
0.5	86.6	95.7	98.5	84.1

5. Conclusion

In this paper, a method based on global-local feature and subsampling strategy is proposed for one-shot video-based person Re-ID. In feature extraction, this method not only focuses on the coarse-grained global features of the video tracklet, but also considers the fine-grained local features, and combines the features of the two dimensions to obtain a more discriminative feature representation. On the basis of cross-entropy loss, joint center loss and exclusive loss further improve discriminative feature learning. In addition, in order to reduce the influence of false label samples, a subsampling strategy is proposed during label estimation, which divides the selected pseudo-labeled samples into two appropriate subsets for training. Good performance on MARS and DukeMTMC-VideoReID datasets proves the effectiveness of the proposed method.

6. References

- [1] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. ICCV, 2017.
- [2] Y. Yan et al., "Learning Multi-Granular Hypergraphs for Video-Based Person Re-Identification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2896-2905, doi: 10.1109/CVPR42600.2020.00297.
- [3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [4] X. Zhu et al., "Semi-Supervised Cross-View Projection-Based Dictionary Learning for Video-Based Person Re-Identification," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2599-2611, Oct. 2018, doi: 10.1109/TCSVT.2017.2718036.
- [5] Ma, D., Zhou, Y., Zhao, J. et al. Video-based person re-identification by semi-supervised adaptive stepwise learning. Pattern Anal Applic 24, 1769–1776 (2021).
- [6] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. ICCV, 2017.
- [7] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In ICCV, 2017.
- [8] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5177 – 5186.
- [9] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," IEEE Trans. Image Process., vol. 28, no. 6, pp. 2872–2881, Jun. 2019.

- [10] Yin YC, et al. " One-shot video-based person re-identification with multi-loss learning and joint metric." *Journal of Computer Applications* 42 (2022):764-769.
- [11] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: *European Conference on Computer Vision*, Springer, 2016, pp. 499–515.
- [12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.
- [13] N. McLaughlin, J. Martinez del Rincon and P. Miller, "Recurrent Convolutional Network for Video-Based Person Re-identification," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1325-1334, doi: 10.1109/CVPR.2016.148.
- [14] S. Li, S. Bak, P. Carr and X. Wang, "Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-identification," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 369-378, doi: 10.1109/CVPR.2018.00046.
- [15] J. Yang, W. -S. Zheng, Q. Yang, Y. -C. Chen and Q. Tian, "Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3286-3296, doi: 10.1109/CVPR42600.2020.00335.
- [16] X. Dong, D. Meng, F. Ma, and Y. Yang, "A dual-network progressive approach to weakly supervised object detection," in *ACM on Multimedia Conference*, 2017, pp. 279–287.
- [17] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*, 2016, pp. 868–884.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.