2022 The 22st International Conference on Control, Automation and Systems (ICCAS 2022)
BEXCO, Busan, Korea, Nov. 27∼Dec. 01, 2022

Semantic Foreground Feature Extraction and Camera-aware Re-allocation Clustering for Unsupervised Person Re-identification

Ge Cao[1] and Kanghyun Jo[1*]

[1]Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, 44610, Korea (caoge@ulsan.ac.kr, acejo@ulsan.ac.kr)

Abstract: Unsupervised person re-identification has achieved great performance in recent years, especially the clustering-based method. However, there are still two challenging problems in this field. First, most methods generate pseudo labels with input samples that contain background noise. Second, the common clustering method cannot guarantee enough reliable clustering quality. To overcome these problems, we propose a Semantic Feature Extraction-Camera-aware Re-allocation (SFE-CR) framework for unsupervised person re-identification. Especially, in the semantic feature extraction module, we utilize a parsing model to extract semantic local features for training samples, so as to eliminate the background noise. In the camera-aware re-allocation module, we split the samples using their camera id and re-allocate the pseudo label generated by the common clustering method. Extensive experiments on Market-1501 and DukeMTMC-reID datasets demonstrate that our method's effection outperforms the baseline evidently.

Keywords: Unsupervised person re-identification, Semantic feature extraction, Label Re-allocation

1. INTRODUCTION

Person Re-identification (Re-ID) is the task of searching the queried pedestrian from a non-overlapping multi-camera system, which is the foundation of a wide range of applications, such as person tracking [1] and human activity analysis [2]. In order to reduce the time-consuming and cumbersome work of manual annotation, large numbers of existing works focus on unsupervised person re-ID cases. Thanks to the rapid development of deep learning, the research in the unsupervised field also performed great results.

In the unsupervised person re-ID field, one popular solution is Unsupervised Domain Adaptation (UDA) [3], which aims to train a powerful backbone model for the target domain dataset by using the annotated label in the source domain dataset. However, this kind of work also leverages the annotated label, and this paper proposed a method that only focuses on the target dataset, which is called fully unsupervised and therefore is more challenging than the UDA-based idea. Many existing methods mainly resort to generating reliable pseudo labels to train under supervised cases. For producing sufficiently reliable pseudo label, clustering-based methods [9], [4] have been widely used to divide the training instances into multiple clusters and treat each cluster as a pseudo-identity class. Thus, the clustering quality is vital for unsupervised training. As shown in Fig.1, it's obvious that the clustering quality would be affected by the varying backgrounds, illuminations, clothing styles, and occlusions across different cameras. Those negative factors cause a large domain gap which harmed the training performance. And the pipeline of most clustering-based methods will generate large quantities of wrong pseudo labels in the beginning epochs of the training. As shown in Fig. 2, there are four sets of samples that are taken from the clustering results of the first training epoch. The four samples in each blue dotted box are wrongly clustered as a pseudo identity mainly due to the similar background. And most clusters contain the samples captured from the same camera. To solute this problem, we proposed a framework named SFE-CR, which contains a Semantic Foreground feature Extraction module (SFEM) for eliminating the background noise, and a Camera-aware Re-allocation Module (CRM) to allocate a more robust and reliable pseudo label for each sample.

Fig. 1. Illustration of the domain gaps between Market-1501 [14] and DukeMTMC-reID [13]. The images of each row is captured from the same identity. It is obvious that the backgrounds, illuminations, clothing style vary greatly across different domains.

The remaining content is organized as follows. Section 2 summarized the related work. The details of SFEM

Fig. 2. Illustration of the negative affection of the similar background for clustering quality on Market-1501 [14]. The samples in each blue dotted box are wrongly clustered as a pseudo identity in the clustering process.

and CRM in the whole architecture are introduced in Section. 3. Section 4 provides the implementation details and the results of the experiments. Finally, Section. 5 concludes the paper.

2. RELATED WORKS

Traditional unsupervised person re-ID works can be classified into three categories. The first category can be regarded as a collection of traditional unsupervised person re-ID methods hand-craft features design [5], localized salience statistic exploit [20] and dictionary learning based methods [21]. However, it is difficult to design robust and discriminative features in cross-camera conditions. And different illumination and viewpoint would hinder these kinds of methods to pursue better performance. Recent unsupervised person re-ID methods with deep learning ideas achieved great performance and can be roughly categorized into unsupervised domain adaptation (UDA) and fully unsupervised methods.

A typical representative framework is proposed by Yang *et al.* [22] which mined the potential similarity of unlabeled samples and achieved great success. Apart from that, Lin *et al.* [9] innovatively proposed a bottom-up clustering framework that trains networks with pseudo labels generated by unsupervised learning iteratively. ECN [19] proposed three types of underlying invariance to reduce the feature distribution gap between the source and target domains. MAR [23] adopted a different strategy and used the source dataset as a reference to learn soft labels as ground truth to supervise the re-ID training. The MMCL [11] does not need any label dataset as an auxiliary reference and achieves better performance than most transfer learning methods.

Although the above approaches have achieved remarkable performance on unsupervised person re-identification, most of them perform feature extraction that contains both foreground information and background noise. In addition, the works which utilize the common clustering method like DBSCAN [4] and k-nearest [24], have no more innovative contribution.

3. METHODOLOGY

In this section, we introduce our proposed framework detailedly. Section 3.1 is the problem formulation and overview of this paper. The Sections 3.2 and 3.3 demonstrate the Semantic Foreground feature Extraction module (SFEM) and Camera-aware Re-allocation Module (CRM) respectively.

3.1 Problem Formulation and Overview

Given an unlabeled person re-ID dataset $X = \{x_1, x_2, ..., x_N\}$, where N means the total number of identities, our goal is to train a person re-ID model on $\mathcal{X}$. The unlabeled dataset X consists of three subsets: training set $\mathcal{T}$, query set $\mathcal{Q}$,and gallery set $\mathcal{G}$. When giving any query image q from query set $\mathcal{Q}$, we expect the re-ID model produces a feature vector f_i to retrieve image g in gallery set $\mathcal{G}$ containing the same identity. As shown in Fig. 3, when giving any input image x_i, the model extracts the feature vector as representation. For inference, the representation of the same person are supposed to be as close as possible.

At the beginning of each training epoch, we use the backbone network to extract appearance representations $M = \{m_1, m_2, ..., m_N\}$ of all the samples in the training set X . We use a clustering algorithm DBSCAN [4] on these appearance representations to generate pseudo identity labels $Y = \{y_1, y_2, ..., y_N\}$. We only consider clustered inliers for training, while un-clustered outliers are discarded.

The centroid of cluster a is defined as the averaged momentum representations of all the instances belonging to this cluster:

$$p_a = \frac{1}{N_a} \sum_{m_i \in y_a} m_i \tag{1}$$

where N_a is the number of instances belonging to the cluster a. Then for an input image in X, we can extract the representative feature f_a belonging to cluster a. The contrastive loss is a softmax log loss with one positive cluster p_a and all the negatives in the memory:

$$Ł = E\left[-log \frac{exp(f_a \cdot p_a/\tau)}{\sum_{i=1}^{|p|} exp(f_a \cdot p_i/\tau)}\right] \tag{2}$$

where $|p|$ is the number of clusters in a training epoch and τ is a temperature hyper-parameter. Different from unified contrastive loss, outliers are not considered as single instance clusters. In such a way, outliers are not pushed away from clustered instances, which allows us to mine more hard samples for our proposed hard instance contrast.

This is the baseline algorithm we applied in this paper. The performance of the baseline and more comparisons is shown in section 4.

3.2 Semantic Foreground Feature Extraction Module

The background could affect the clustering results' quality as we showed in Fig.2, we construct a semantic foreground feature extraction module to eliminate the

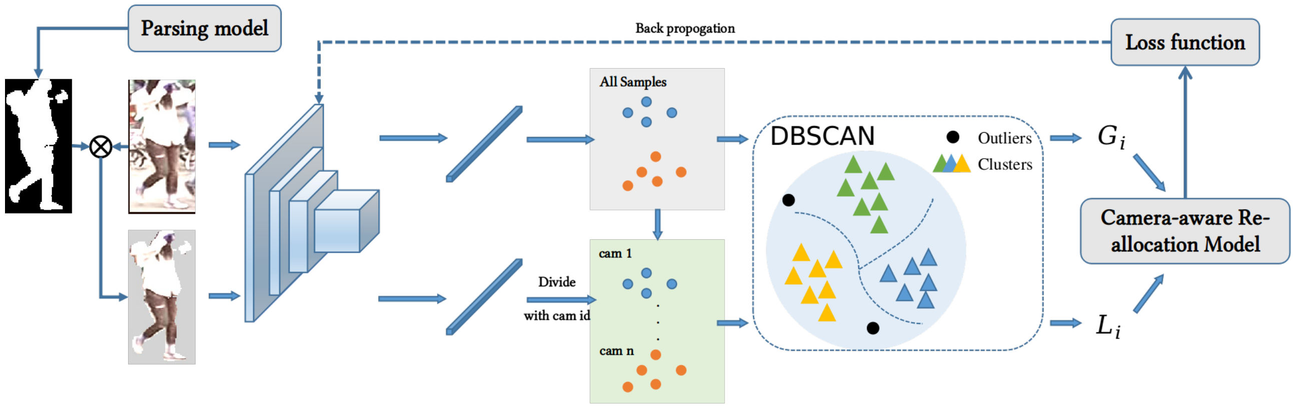

Fig. 3. The overview of the proposed SFE-CR framework.

background noise. Given a sample x_i, we utilize a parsing model to get the human body foreground mask P_i. The foreground mask is a two-dimension matrix with binary values. Then we perform element-wise multiplication on the input sample x_i and foreground mask P_i to get the input sample without background noise G_i, which keeps the same size as the input sample,

$$G_i = x_i \bigotimes F_i \tag{3}$$

where $\bigotimes$ denotes the element-wise multiplication.

Then given the input sample x_i and generated sample with no background noise G_i, the backbone network extracts the D-dim feature F_i^x and F_i^g, respectively. The F_i^x, which is extracted from the input sample, is a vital feature for providing a discriminative feature to training loss. We directly utilize the clustering-based method DBSCAN [4] to generate the global clustering pseudo label. Then subsection 3.3 will introduce the usage of F_i^g in the camera-aware re-allocation module.

3.3 Camera-aware Re-allocation Module

To generate a more robust and reliable global pseudo label, we propose the camera-aware re-allocation module to make the global clustering results more accurate. Given the feature F_i^g, we divide the samples into n groups, where n denotes the number of cameras in the training process, for example, in Market-1501 [14] dataset, the samples are captured by 6 cameras. So in this step, the samples are divided into 6 groups. And we apply DBSCAN in each group. Then every sample has a local cluster pseudo label. The process of re-allocation is shown in Fig. 4. If the majority of samples in local clusters appear in the same global cluster, we consider all samples in local clusters belonging to a global cluster. So given the local cluster L_i and global cluster G_i, we compute the overlap degree $P(L_i \to G_i) by$:

$$P(L_i \to G_i) = \frac{|L_i \cap G_i|}{|L_i|} \tag{4}$$

where $|\cdot|$ denotes the number of samples in the set. So the $P(L_i \to G_i)$ is the indicator for deciding whether L_i should be included into G_i. The process can be formu-

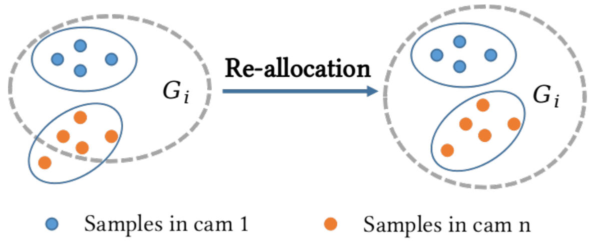

Fig. 4. Illustration of the camera-aware re-allocation module. Due to the dividing operation (with camera id), each identity is naturally split into multiple local clusters. If the majority samples in local clusters appear in the same global cluster, we consider all samples in local clusters belonging to a global cluster.

lated as:

$$\begin{aligned} L_i &\to G_i \\ s.t. P(L_i \to G_i) &\geqslant \delta \end{aligned} \tag{5}$$

where δ is the threshold.

For instance, in Fig. 4, $n = 2$, when given the local cluster label and global cluster label of the orange sample. The overlap degree $P(L_i \to G_i) = \frac{4}{5}$, so the orange sample which originally not belong to the global cluster should be re-allocated into G_i.

Discussions: About the condition when the overlap rate is too low, this paper didn't remove the samples which originally belong to the global cluster. Because of the bad performance of setting the bottom-threshold for removing the terrible clustered sample, we have reason to believe that the re-allocation algorithm is not wholly satisfactory. And in future work, we will focus on that and improve the performance.

4. EXPERIMENTS

4.1 Dataset and evaluation metrics

Market-1501 [14], *DukeMTMC-reID* [13] datasets are applied in this paper. *Market-1501* contains 32,668 labeled person images of 1,501 identities collected from 6 non-overlapping camera views. *DukeMTMC-reID* contains 36,411 annotated images of 1,404 identities with

Table 1. Unsupervised person re-ID performance comparison with other methods on Market-1501 and DukeMTMC-reID.

Method	reference	Market-1501					DukeMTMC-reID				
		Source	Rank-1	Rank-5	Rank-10	mAP	Source	Rank-1	Rank-5	Rank-10	mAP
LOMO [5]	CVPR15	None	27.2	41.6	49.1	8	None	12.3	21.3	26.6	4.8
BOW [6]	ICCV15	None	35.8	52.4	60.3	14.8	None	17.1	28.8	34.9	8.3
UDML [7]	CVPR16	None	34.5	52.6	59.6	12.4	None	18.5	31.4	37.6	7.2
DECAMEL [8]	TPAMI18	None	60.2	76	81.1	32.4	-	-	-	-	-
BUC [9]	AAAI19	None	66.2	79.6	84.5	38.3	None	47.4	62.6	68.4	27.5
DBC [10]	BMVC19	None	69.2	83	87.8	41.3	None	51.5	64.6	70.1	30
MMCL [11]	CVPR20	None	80.3	89.4	92.3	45.5	None	65.2	75.9	80	40.2
Ours(SFE-CR)	This paper	None	**83.1**	**92.5**	**95.3**	**65.4**	None	**77.7**	**87.2**	**90.7**	**61.8**

Table 2. Unsupervised person re-ID performance comparison with the baseline on Market1501, DukeMTMC-reID datasets.

Model	Market-1501				DukeMTMC-reID			
	mAP	R1	R5	R10	mAP	R1	R5	R10
baseline	62.9	79.7	88.3	91.2	57.5	74.3	82.7	86.0
SFE-CR	65.4(**+2.5**)	83.1(**+3.4**)	92.5(**+4.2**)	95.3(**+4.1**)	61.8(**+4.3**)	77.7(**+3.4**)	87.2(**+4.5**)	90.7(**+4.7**)

Table 3. The information of two person re-identification datasets (Market-1501 and DukeMTMC-reID) used in this work.

Dataset	Identities	Training	Gallery	Query	Cameras
Market [14]	1,501	12,936	19,732	3,368	6
Duke [13]	1,404	16,522	17,611	2,228	8

8 cameras. These datasets are constructed with a large amount of annotated images collected from different view cameras, illumination, indoor or outdoor scene, and other variations. More details can be found in Table. 3. In this paper, we follow the standard settings with [11], [19]. We don't use any other labeled dataset when training and testing and the performance is evaluated by Mean average precision (mAP) and the Cumulative Matching Characteristic (CMC) Rank-1/5/10 matching accuracy.

4.2 Implementation Details

We employ the ResNet-50 [12] pretrained on ImageNet [16] as the backbone network in all the experiments. Based on it, we remove the fully-connected classification layer, and add a Batch Normalization (BN) layer after the Global Average Pooling (GAP) layer. The L_2 normalized feature is used for the updating of proxies in the memory during training, and also for the distance ranking during inference. For each input image, we resize it into 256×128.

In the camera-aware re-allocation module, the threshold $\delta = 0.85$ in this paper.

The parsing model is the state-of-th-art SCHP pretrained model [17] trained on LIP dataset [18] to get the human body masks for all samples in advance. The original human mask has 6 parts for background, head, leg, arm, upper-body and the foreground. In this paper, we only utilize the foreground part.

At the beginning of each epoch, we compute Jaccard distance with k-reciprocal nearest neighbors and use DBSCAN with a threshold of 0.5 for the global clustering.

We use ADAM as the optimizer. The initial learning rate is 0.00035 with a warmup scheme in the first 10 epochs, and is divided by 10 after each 20 epochs. The total epoch number is 50. We set the batch size to 32 in training and testing and all experiments are implemented on PyTorch 1.4 with CUDA 10.1. Each training batch consists of 32 images randomly sampled from 8 proxies with 4 images per cluster. Random flipping, cropping, and erasing are applied as data augmentation.

4.3 Test results on Person Re-ID datasets

We compare the proposed method against state-of-the-art unsupervised learning works on *Market-1501* [14], *DukeMTMC-reID* [13]. The comparison results are shown in Table. III.

We only compare with the methods which don't need any other labeled dataset: LOMO [5], BOW [6], UDML [7], DECAMEL [8], BUC [9], DBC [10] and MMCL [11]

In the comparison methods, LOMO and BOW used traditional unsupervised learning methods which utilize hand-crafted features and got lower re conclusion section is not required. Compared with others. UDML proposed a multi-task dictionary learning method to learn dataset-shared but target-data-biased representation. DECAMEL, BUC and DBC use the clustering method to train their networks. MMCL proposed memory-based multi-label classification loss.

It is obvious that our proposed model outperforms other works with a large margin. For instance, Table. 1 compared the baseline with the other methods. Compared with the baseline, our approach obtains 3.4% Rank-1 and 2.5% mAP gain on Market-1501 dataset, 3.4% Rank-1 and 4.3% mAP gain on DukeMTMC-reID.

5. CONCLUSION

In this paper, we proposed a Semantic Feature Extraction-Camera-aware Re-allocation (SFE-CR) framework for unsupervised person re-identification tasks. The SFE-CR contains a semantic foreground feature extraction module and a camera-aware re-allocation module, which eliminate the background noise and improve the clustering quality, respectively. This paper demonstrates the effectiveness of the proposed method and shows the performance compared with other methods. And we still have great research potential in re-allocation algorithms in future work.

ACKNOWLEDGMENT

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

REFERENCES

[1] J. Nino, A. Frias Velazquez, N. Bo, M. Slembrouck, J. Guan, G. Debard, B. Vanrumste, T. Tuytelaars, and W. Philips, Scalable semi-automatic annotation for multi-camera person tracking, IEEE Transactions on Image Processing, vol. 25, pp. 11, 05 2016.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, Person Re-identification: Past, Present and Future, arXiv e-prints, p. arXiv:1610.02984, Oct. 2016.

[3] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. PR, 2020. 1, 2

[4] 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.

[5] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, Person Re-identification by Local Maximal Occurrence Representation and Metric Learning, arXiv e-prints, p. arXiv:1406.4216, Jun. 2014.

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, Scalable person re-identification: A benchmark, 12 2015, pp. 11161124

[7] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, Unsupervised cross-dataset transfer learning for person reidentification, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 13061315.

[8] H. Yu, A. Wu, and W. Zheng, Unsupervised person re-identification by deep asymmetric metric embedding, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 956973, 2020.

[9] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 87388745, 07 2019.

[10] G. Ding, S. H. Khan, and Z. Tang, Dispersion based clustering for unsupervised person re-identification, in BMVC, 2019.

[11] D. Wang and S. Zhang, Unsupervised Person Re-identification via Multi-label Classification, arXiv e-prints, p. arXiv:2004.09228, Apr. 2020.

[12] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In CVPR.

[13] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, 2016. 5, 7

[14] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In ICCV, 2015. 2, 5, 6, 7, 8.

[15] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In AAAI, 2021. 2, 3, 4, 7, 8, 11.

[16] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248255.

[17] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. arXiv preprint arXiv:1910.09777, 2019. 3

[18] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. TPAMI, 41(4):871885, 2018. 3

[19] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification, arXiv e-prints, p. arXiv:1904.01990, Apr. 2019.

[20] H. Wang, S. Gong, and T. Xiang, Unsupervised learning of generative topic saliency for person re-identification, BMVC 2014 - Proceedings of the British Machine Vision Conference 2014, 01 2014.

[21] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, Person re-identification by unsupervised 11 graph learning, vol. 9905, 10 2016, pp. 178195.

[22] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. Huang, Self-similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-identification, arXiv e-prints, p. arXiv:1811.10144, Nov. 2018.

[23] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, Unsupervised Person Re-identification by Soft Multilabel Learning, arXiv e-prints, p. arXiv:1903.06325, Mar. 2019.

[24] Zhong, Z.; Zheng, L.; and Li, S. 2017. Re-ranking Person Re-identification with k-Reciprocal Encoding. In CVPR.