Self-Label Refining for Unsupervised Person Re-Identification

Xiaoting Yu, Lijun Guo, *Member, IEEE*, and Rong Zhang, *Member, IEEE*

***Abstract*—Fully unsupervised person Re-ID is a challenging task. State-of-the-art methods perform model training with the pseudo labels generated by clustering algorithms on the unlabeled dataset. However, the label noise caused by clustering limits the performance of person Re-ID tasks. To alleviate the problem, this paper proposes a Self-Label Refining Network (SLRNet). It is considered that the local parts naturally mitigate the variation of intra-identity samples caused by cross-view. Thus, the Self-Label Refining (SLR) module estimates the similarities between global and local pseudo labels with clustering consensus, and then it refines the global pseudo labels by integrating propagated local pseudo labels into global pseudo labels. Meanwhile, a symmetric InfoNCE loss is further proposed to enhance the robustness of the network to noisy labels. Extensive experiments show that our method achieves state-of-the-art performance on three widely used person Re-ID datasets.**

***Index Terms*—Person re-identification, fully unsupervised learning, label refining**

I. Introduction

As a task to identify the same person from non-overlapping cameras [9], [14], [15], [23], person re-identification (Re-ID) has been widely applied to surveillance and public security. Due to the high cost of annotating data, a plethora of studies have been conducted on the unsupervised person Re-ID problem. According to the use of external labeled datasets, unsupervised person Re-ID can be divided into unsupervised domain adaptation (UDA) person Re-ID and fully unsupervised (FU) person Re-ID.

This paper focuses on FU person Re-ID, which does not leverage any labeled data and therefore is more challenging. Most of the recent popular FU person Re-ID methods adopt clustering algorithms to generate pseudo labels for unlabeled samples and thus train the model in a supervised manner [21], [22], [5]. However, clustering algorithms cannot guarantee that intra-identity samples are assigned with the same pseudo label, which will inevitably introduce noisy labels. Efforts have been made to improve the quality of pseudo labels. BUC [11] proposed a bottom-up clustering scheme to optimize the relationship among samples. SpCL [7] introduced a self-paced method to create reliable clusters progressively. More recently, IICS [20] decomposed the sample similarity computation into intra-camera and inter-camera computation to generate more reliable pseudo-labels.

This work was supported in part by the Zhejiang Provincial Public Welfare Technology Research Project under Grant LGF21F020008.*(Corresponding author: Lijun Guo.)*

The authors are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315000, China (e-mail: yxt_always@163.com; guolijun@nbu.edu.cn; zhangrong@nbu.edu.cn).

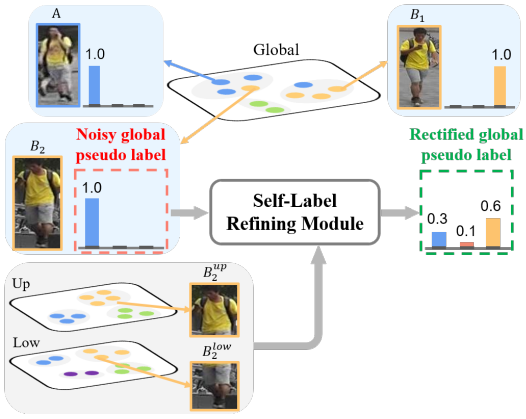


Fig. 1. The function of the SLR scheme. Person image B_1 and B_2 belong to the same identity while A with a similar appearance is from another person. B_2 is misclassified, and its noisy global pseudo-label can be corrected through SLR module to some extent.

Despite the success of these FU person Re-ID approaches, they ignore part-based features containing finer information of a person. Based on this observation, this paper proposes a novel framework called Self-Label Refining Network (SLRNet). SLRNet applies a Self-Label Refining (SLR) scheme to rectify the hard global pseudo labels via part-based features. It should be indicated that part-based features naturally alleviate the intensive variation of intra-identity samples caused by drastic cross-view because local parts usually occupy smaller areas in images [21]. Then, by clustering the part-based features, those intra-identity samples from different camera views can be clustered into the same cluster. The SLR scheme aims to better correct the noisy global pseudo labels, as shown in Fig. 1. Usually, a contrastive loss (e.g., InfoNCE loss) is adopted in unsupervised visual representation learning. It is similar to cross-entropy loss, which uses hard labels and may lead the model overfitting to noisy labels [18]. Although the noisy labels can be corrected through SLR to some extent, it cannot be ensured that corrected labels are absolutely right. Inspired by the symmetric idea from KL-divergence, a symmetric InfoNCE loss (SNCE) is proposed in this paper to act as a noise-tolerant counterpart of InfoNCE loss (NCE), which further enhances the robustness of the network to noisy labels.

The contributions of this paper are summarized as follows: 1) SLRNet is proposed, which covers both global and local information and adopts the SLR scheme to rectify the noisy hard global pseudo labels via part homogeneity. This scheme effectively improves the quality of pseudo labels. 2) A symmetric InfoNCE loss (SNCE) is proposed, which further alleviates the negative effect caused by noisy labels and promotes the

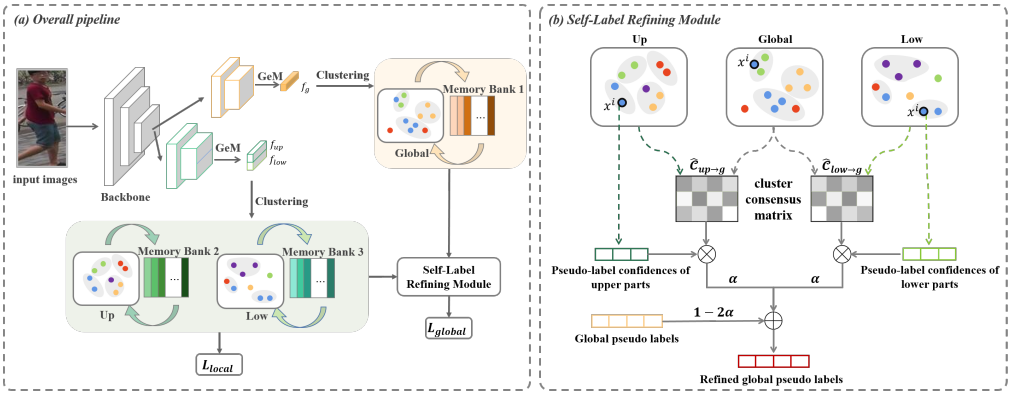


Fig. 2. Illustration of our method. (a) The architecture of the proposed SLRNet, where the SLR module and symmetric InfoNCE loss is adopted to achieve robust training. (b) The detailed structure of the SLR module, which rectifies the hard global pseudo labels via part homogeneity. $\otimes$ and $\oplus$ represent matrix multiplication and element-wise add with weights, respectively.

robustness of the model. 3) The proposed method achieves a significant performance improvement compared to the state-of-the-art methods on three benchmarks for unsupervised person Re-ID.

II. Methodology

A. The Overall Framework

In this paper, the unlabeled dataset is denoted as $\mathcal{D} = \{x_i\}_{i=1}^{N}$, where x_i is the i-th image, and N is the number of images. The structure of the SLRNet is illustrated in Fig. 2. It is a two-branch network architecture consisting of a global branch and a local branch with dedicated parameters from the 4th residual stage of the ResNet50 backbone. In the global branch, global feature vectors are obtained by applying generalized-mean pooling (GeM) to feature maps. In the local branch, referring to the method in [17], local feature vectors, including upper-part features and lower-part features, are obtained by dividing the feature maps into two horizontal strips and GeM. After feature extraction for all training samples, a multi-granularity feature set $\mathcal{F} = \{f_g^i, f_{up}^i, f_{low}^i\}_{i=1}^{N}$ is obtained. Then, a clustering algorithm (e.g., DBSCAN [4]) is used to generate multi-granularity pseudo labels for unlabeled images based on the feature set $\mathcal{F}$. Finally, a new labeled dataset $\mathcal{D}' = \left\{\left(x_i, y_g^i, y_{up}^i, y_{low}^i\right)\right\}_{i=1}^{N'}$ is formed. Differently, the local pseudo labels are generated by concatenating global and local features, which resists the bias in local clustering results caused by local occlusion.

Due to the label noise introduced by clustering, the SLR module is applied to rectify the hard global pseudo labels via part homogeneity. This process will be comprehensively described in Section II-B. In this way, the reliability of global pseudo labels can be improved by the finer information provided by the local branch. After the global noisy labels are rectified to some extent, this paper proposes SNCE to further enhance the noise-robust ability of the model, and this will be described in Section II-C. The above-proposed schemes aim to mitigate the negative effects caused by noisy samples.

During training, three memory banks $\mathcal{K}_g$, $\mathcal{K}_{up}$, $\mathcal{K}_{low}$ are adopted to store the feature centroids of the clusters so that the model can adapt to the changing pseudo-labels after each iteration. The memory banks [8] are updated consistently on the fly to facilitate unsupervised learning. Finally, the combination of SNCE with NCE and soft softmax-triplet loss is adopted to train the model until it converges. In the above process, the local branch is mainly used to provide valuable supervision information to improve the global feature representation. Meanwhile, local pseudo labels are not refined in the training process. Therefore, only global features are adopted in the inference stage. The detailed experiments are formulated in the supplemental materials.

B. Rectification with the SLR scheme

Due to the drastic cross-view, the intra-identity samples from different cameras may suffer from the changes of viewpoint and other environmental factors. As a result, these hard positive samples are possibly assigned with wrong pseudo labels, which degrades the model's ability to identify people in detail. Since local features provide finer information than the global ones, by combining with the information from global pseudo labels, the pseudo-label confidence obtained from local features can provide reliable supervision information. To this end, this paper applies the SLR scheme to rectify the noisy global pseudo labels by exploiting the pseudo-label confidence from local features.

Since the global and local pseudo label sets do not overlap, local pseudo labels cannot be propagated and aggregated to global pseudo labels through the existing label ensemble methods [10]. Therefore, through clustering consensus, this paper establishes the similarities between the pseudo labels of the global pseudo-label group Y_g and local pseudo-label groups, Y_{up} and Y_{low}. Specifically, the global sample set with a pseudo label i is denoted as $I_g[i]$, where $i \in [1, Z_g]$, and Z_g is the cluster number of the global clustering results. Similarly, the sample sets with a pseudo label j in upper-part clusters and k in lower-part clusters are respectively denoted as $I_{up}[j]$ and $I_{low}[k]$, where $j \in [1, Z_{up}]$ and $k \in [1, Z_{low}]$. The clustering consensus matrixes $\mathcal{C}_{up\to g} \in \mathbb{R}^{Z_{up}\times Z_g}$ and $\mathcal{C}_{low\to g} \in \mathbb{R}^{Z_{low}\times Z_g}$ store the Intersection over Union (IoU) criterion between global label sets and local label sets, respectively. Take $\mathcal{C}_{up\to g}$ as an example, $\mathcal{C}_{up\to g}[j,i]$ is calculate as,

$$\mathcal{C}_{up\to g}[j,i] = \frac{|I_{up}[j] \cap I_g[i]|}{|I_{up}[j] \cup I_g[i]|} \in [0,1] \tag{1}$$

where $|\cdot|$ denotes the number of samples in a set. Intuitively, $\mathcal{C}_{up\to g}[j,i]$ measures the consensus or similarity between global pseudo class i and upper-part pseudo class j. The pseudo classes that are consistent between global and local pseudo-label groups will have higher confidence. Then, the consensus matrix $\mathcal{C}_{up\to g}$ is normalized in row to obtain the normalized matrix $\hat{\mathcal{C}}_{up\to g}$. Similarly, $\hat{\mathcal{C}}_{low\to g}$ can be obtained as well. Based on the cross-granularity consensus matrixes, $\hat{\mathcal{C}}_{up\to g}$ and $\hat{\mathcal{C}}_{low\to g}$, the process of propagating local pseudo labels to global pseudo classes is determined as follows. For the soft labels that make samples more robust against label noise [1], this paper refines the global pseudo labels based on the soft local pseudo-label confidence. As shown in Fig. 2(b), the confidences are firstly calculated with the class proxies in memory as $\mathcal{K}_{up}^{\mathrm{T}} f_{up}^i$ and $\mathcal{K}_{low}^{\mathrm{T}} f_{low}^i$. Then, the propagated soft local pseudo labels $\hat{y}_{up}^i \in \mathbb{R}^{Z_g}$ and $\hat{y}_{low}^i \in \mathbb{R}^{Z_g}$ are calculated, as shown in Eq. (2).

$$\begin{aligned} \hat{y}_{up}^i &= \hat{\mathcal{C}}_{up\to g}^{\mathrm{T}} softmax\left(\tau \cdot \mathcal{K}_{up}^{\mathrm{T}} f_{up}^i\right) \\ \hat{y}_{low}^i &= \hat{\mathcal{C}}_{low\to g}^{\mathrm{T}} softmax\left(\tau \cdot \mathcal{K}_{low}^{\mathrm{T}} f_{low}^i\right) \end{aligned} \tag{2}$$

where τ is a temperature hyper-parameter for sharpening the class confidence. The propagated labels $\hat{y}_{up}^i$ and $\hat{y}_{low}^i$ will be integrated into global pseudo labels. Finally, the noisy hard global pseudo label y_g^i can be defined as $\tilde{y}_g^i \in \tilde{Y}_g$ via Eq. (3).

$$\tilde{y}_g^i = (1-2\alpha)\cdot y_g^i + \alpha \cdot \hat{y}_{up}^i + \alpha \cdot \hat{y}_{low}^i \tag{3}$$

where $\alpha \in [0,1]$ is a momentum value for ensembling. Based on the propagated soft local pseudo labels, the original noisy global pseudo labels can be rectified through the cross-granularity pseudo-label similarities calculated before. Well-rectified global pseudo labels lead to more robust training.

C. Training with SNCE Loss

The performance of pseudo-label-based FU person Re-ID methods depends on the accuracy of the clustering result. However, clustering algorithms will inevitably introduce label noise. Though SLR can rectify the noisy pseudo labels to some extent, a gap still exists between the noisy pseudo labels and the ground-truth labels. To this end, a noise-tolerant loss function is proposed in this paper to resist noisy labels. First of all, NCE is introduced [12], which is formulated as,

$$\mathcal{L}_{nce}(f) = -\log \frac{\exp(\tau \cdot \langle f, \mathcal{K}[j]\rangle)}{\sum_{k=1}^{Z} \exp(\tau \cdot \langle f, \mathcal{K}[k]\rangle)} \tag{4}$$

where f is a query feature vector; $\mathcal{K}[j]$ indicates the positive class prototype corresponding to f; the definition of temperature τ is the same as that in Eq. (2); $\langle\cdot,\cdot\rangle$ denotes the inner product between two feature vectors to measure their similarity; Z is the cluster number. NCE pulls an instance close to the centroid of its class and pushes it away from the centroids of all other classes. NCE is similar to cross-entropy loss, which uses hard labels and may lead the model overfitting to noisy labels. Besides, noisy pseudo labels do not represent the true class distribution, but the soft label confidence can reflect the true distribution to a certain extent [18]. Inspired by the symmetric idea from KL-divergence, this paper designs SNCE, which acts as a noise-robust counterpart of NCE.

$$\mathcal{L}_{snce}(f) = -\sum_{t=1}^{Z} \frac{\exp(\tau \cdot \langle f, \mathcal{K}[t]\rangle)}{\sum_{j=1}^{Z} \exp(\tau \cdot \langle f, \mathcal{K}[j]\rangle)} \log(softmax(y)[t]) \tag{5}$$

where y is the label corresponding to f. Softmax normalization is used to avoid the calculation of zero values inside the logarithm caused by a dimension of zero in labels. Meanwhile, triplet loss is commonly used to achieve optimal performance in person Re-ID tasks. In this paper, the soft softmax-triplet loss [6] is adopted so that the network can benefit from softly refined labels.

$$\begin{aligned} \mathcal{L}_{stri}(f) &= \mathcal{L}_{bce}\left(\mathcal{T}\left(f, f^+, f^-\right), \mathcal{T}\left(y, y^+, y^-\right)\right) \\ \mathcal{T}(\mathrm{a},\mathrm{p},\mathrm{n}) &= \frac{\exp(\|a-n\|)}{\exp(\|a-p\|)+\exp(\|a-n\|)} \end{aligned} \tag{6}$$

where $\mathcal{L}_{bce}$ denotes the binary cross-entropy loss; f^+ and f^- respectively denote the features of hardest positive and negative samples corresponding to f in the mini-batch; $\|\cdot\|$ is the L^2-norm distance; y, y^+, and y^- are labels corresponding to f, f^+ and f^-, respectively. Since the proposed network model has two branches, i.e., the global branch and the local branch, our loss function is formulated as follows

$$\begin{aligned} \mathcal{L}_{local} &= \mathcal{L}_{nce}\left(f_{up}, Y_{up}\right) + \mathcal{L}_{nce}\left(f_{low}, Y_{low}\right) \\ \mathcal{L}_{global} &= \mathcal{L}_{nce}\left(f_g, Y_g\right) + \mathcal{L}_{snce}\left(f_g, \tilde{Y}_g\right) + \mathcal{L}_{stri}\left(f_g, \tilde{Y}_g\right) \\ \mathcal{L}_{all} &= \mathcal{L}_{glocal} + \lambda \mathcal{L}_{local} \end{aligned} \tag{7}$$

where λ is a parameter to balance the two terms of $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$, and it is set to 0.5 in this paper. The model parameters are updated by gradient descent during back-propagation, while the three memory banks are updated by

$$\mathcal{K}[j] = \mu \mathcal{K}[j] - (1-\mu) f^{hard} \tag{8}$$

where $\mathcal{K}[j]$ is the j-th entry of the memory, which stores the updated feature centroid of class j; f^{hard} is the feature of the batch-hard instance with pseudo label j and the minimum similarity to the cluster feature $\mathcal{K}[j]$; $\mu \in [0,1]$ is an updating rate.

III. Experiment

A. Datasets and Evaluation Metrics

In this paper, our proposed method is evaluated on three widely-used person Re-ID datasets: Market-1501 [24], DukeMTMC-reID [13], and MSMT17 [19]. Market-1501 contains 32,668 images of 1,501 identities, and the images are captured from 6 cameras at Tsinghua University. DukeMTMC-reID is composed of 36,411 images of 1,812 identities, and the images are captured by 8 cameras at Duke University. MSMT17 is the most challenging person re-ID dataset, which contains 126,411 person images of 4,101 identities. The images of this dataset are collected by 15 cameras at Peking University over four days. The performance of our method is evaluated by Mean average precision (mAP) and rank-1 accuracy. In the experiments, ground-truth IDs are not available for training, and no post-processing like re-ranking [25] is applied.

TABLE I
COMPARISON OF THE STATE-OF-THE-ARTS METHODS INCLUDING FU AND UDA

	Methods	Market-1501		DukeMTMC-reID		MSMT17	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
UDA	ECN++[26]	63.8	84.1	54.4	74.0	-	-
	SSG++[5]	68.7	86.2	60.3	76.0	18.3	41.6
	MMT[6]	71.2	87.7	65.1	78.0	23.3	50.1
	IDM[2]	82.8	93.2	70.5	83.6	35.4	63.6
FU	BUC[11]	38.3	66.2	27.5	47.4	-	-
	MMCL[16]	45.5	80.3	40.2	65.2	11.2	35.4
	HCT[22]	56.4	80.0	50.7	69.6	-	-
	SPCL[7]	73.1	88.1	-	-	19.1	42.3
	IICS[20]	72.9	89.5	64.4	80.0	26.9	56.4
	ClCo[3]	80.9	91.7	71.2	83.7	31.0	59.8
ours	**SLRNet**	**82.7**	**92.8**	**72.1**	**84.2**	**33.3**	59.3

B. Implementation Details

The input image is resized to 256×128. Meanwhile, several data augmentation strategies are applied to all training images, including horizontal flipping, random erasing, and random cropping. The total number of training epochs is set to 50. Each mini-batch includes 64 images of 16 pseudo identities. At the beginning of each epoch, DBSCAN [4] is used for clustering to generate pseudo labels. During the training process, Adam optimizer is adopted with an initial learning rate of 3.5×10^{-4} and a weight decay of 5×10^{-4}. The memory updating rate μ is empirically set to 0.2. Besides, two hyper-parameters required by SLRNet are set. The initial value of α in Eq. (3) is set to 0.2, and it follows the poly policy for decay; the temperature τ in Eq. (2) is set to 10.

C. Comparison with State-of-the-arts

In this section, SLRNet is compared with state-of-the-art FU person Re-ID methods and UDA person Re-ID methods on the three datasets. It can be seen from Table I that SLRNet significantly outperforms most of the comparison methods. The six most recent FU person Re-ID methods taken for comparison are based on pseudo labels, and they aim to generate more reliable clusters. By contrast, SLRNet focuses on rectifying noisy pseudo labels via part homogeneity and training with a noise-tolerant loss function. Thus, our method obtains a performance improvement. Some unsupervised person Re-ID works focus on UDA techniques that exploit external labeled data for performance improvement. Surprisingly, without using any labeled information, our method achieves similar performance to these works on three benchmarks. Compared with the FU person Re-ID method that ranks the second in performance, SLRNet gains a performance improvement of 1.8%, 0.9% and 2.3% in mAP on Market, Duke, and MSMT17.

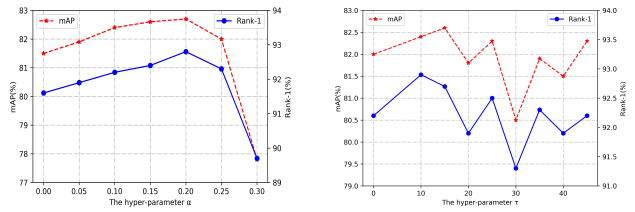
D. Ablation Studies

Ablation studies are conducted to investigate the effectiveness of the components of the proposed method. For reference, the results of the baseline model are presented. In the baseline model, DBSCAN is used to generate multi-granularity pseudo labels. Meanwhile, the SLR scheme and SNCE are not applied.

TABLE II
ABLATION STUDIES OF THE PROPOSED METHOD

Methods	Market-1501		DukeMTMC-reID		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
B	81.1	91.7	69.3	82.7	27.0	51.8
B+SLR	82.4	92.5	71.0	82.7	31.2	57.2
B+$\mathcal{L}_{snce}$	81.9	91.5	70.8	83.6	27.6	52.6
B+SLR+$\mathcal{L}_{snce}$	82.7	92.8	72.1	84.2	33.3	59.3

'B' denotes the Baseline.



(a) (b)

Fig. 3. Hyper-parameter analysis results of α in Eq. (3) and τ in Eq. (2).

It is worth noting that only global features are adopted for testing in this paper. As shown in the first row of Table II, the baseline model achieves an mAP of 81.1%, 69.3%, and 27.0% on the three datasets, respectively.

Effectiveness of SLR. Pseudo-label noise hinders the model's ability to distinguish identities in detail. With SLR, noisy global pseudo labels can be rectified through part homogeneity. As shown in the second row of Table II (B+SLR), a great performance improvement is achieved compared to the baseline method, and the mAP is improved by 1.3%, 1.7%, and 4.2% on Market, Duke, and MSMT17. This suggests that the SLR scheme can help the model to identity people regardless of cross-view scene variation.

Effectiveness of SNCE. When SNCE loss is adopted for training, a consistent performance improvement can be obtained. It demonstrates the advantages of SNCE loss when training with the noisy labels generated by cluster algorithms.

Hyper-parameter Analysis. Two important hyper-parameters are investigated, and the parameter study is conducted on Market1501. 1) Momentum α in Eq. (3). α is a momentum value for ensembling. As illustrated in Fig. 3(a), different values of α in the ensemble equation are investigated. Through the experiments, it is found that the optimal performance can be achieved when α is around 0.2. 2) Temperature τ in Eq.(2). τ is a temperature hyper-parameter, and it is used to sharpen the class confidence estimated by the encoded features and class prototypes in memory banks. As shown in Fig. 3(b), when $\tau = 10$, the model obtains the optimal performance.

IV. CONCLUSION

This paper proposes SLRNet to address the performance degradation caused by hard noisy pseudo labels. Specifically, a label rectification technique based on part homogeneity is applied, and a noise-tolerant loss function is introduced. The experimental results demonstrate that the proposed network outperforms other state-of-the-art methods on several datasets.

REFERENCES

[1] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving imagenet classification through label progression," *arXiv preprint arXiv:1805.02641*, 2018.

[2] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, and L.-Y. Duan, "IDM: An Intermediate Domain Module for Domain Adaptive Person Re-ID," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 864–11 874.

[3] Z. Dai, G. Wang, S. Zhu, W. Yuan, and P. Tan, "Cluster contrast for unsupervised person re-identification," *arXiv preprint arXiv:2103.11568*, 2021.

[4] M. Ester, H.-P. Kriegel, J. Sander, X. Xu et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, vol. 96, no. 34, 1996, pp. 226–231.

[5] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person reidentification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6111–6120.

[6] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *8th International Conference on Learning Representations, ICLR*, 2020.

[7] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re-ID," in *Advances in Neural Information Processing Systems*, 2020.

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[9] Y. Hou, S. Lian, H. Hu, and D. Chen, "Part-relation-aware feature fusion network for person re-identification," *IEEE Signal Processing Letters*, vol. 28, pp. 743–747, 2021.

[10] S. Laine and T. Aila, "Temporal Ensembling for Semi-Supervised Learning," in *5th International Conference on Learning Representations, ICLR*, 2017.

[11] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person reidentification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8738–8745.

[12] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[13] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," in *Computer Vision – ECCV 2016 Workshops*, vol. 9914, 2016, pp. 17–35.

[14] X. Shu, G. Li, X. Wang, W. Ruan, and Q. Tian, "Semantic-guided pixel sampling for cloth-changing person re-identification," *IEEE Signal Processing Letters*, vol. 28, pp. 1365–1369, 2021.

[15] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)," in *Computer Vision - ECCV 2018*, vol. 11208, 2018, pp. 501–518.

[16] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 978–10 987.

[17] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[18] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 322–330.

[19] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 79–88.

[20] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 921–11 930.

[21] Q. Yin, G. Wang, G. Ding, S. Gong, and Z. Tang, "Multi-view label prediction for unsupervised learning person re-identification," *IEEE Signal Processing Letters*, vol. 28, pp. 1390–1394, 2021.

[22] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 654–13 662.

[23] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3183–3192.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3652–3661.

[26] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2723–2738, 2021.