Graph-based Attribute-aware Unsupervised Person Re-identification with Contrastive learning

1st Ge Cao, 2nd Qing Tang, 3rd Kanghyun Jo*
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Republic of Korea
caoge9706@gmail.com; zucchini.tang@hotmail.com; acejo@ulsan.ac.kr

2022 International Workshop on Intelligent Systems (IWIS) | 979-8-3503-9893-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWIS56333.2022.9920894

***Abstract*—This paper is employed on the unsupervised person re-identification (Re-ID) task which does not leverage any annotation provided by the target dataset and other datasets. Previous works have investigated the effectiveness of applying self-supervised contrastive learning, which adopts the cluster-based method to generate the pseudo label and split each cluster into multiple proxies by camera ID. This paper applies the Attribute Enhancement Module (AEM), which utilizes Graph Convolutional Network to integrate the correlations between attributes, human body parts features, and the extracted discriminative feature. And the experiments are implemented to demonstrate the great performance of the proposed Attribute Enhancement Contrastive Learning (AECL) in camera-agnostic version and camera-aware version on two large-scale datasets, including Market-1501 and DukeMTMC-ReID. Compared with the baseline and the state-of-the-art, the proposed framework achieves competitive results.**

***Index Terms*—Contrastive learning, graph convolutional network, attribute learning**

I. Introduction

A challenging and important technique named person re-identification seeks to retrieve pedestrians of the target individual for a non-overlapping camera system. Based on the increasing demand and vital application in video surveillance, person re-ID has drawn the attention of many researchers. In Re-ID systems, given an image containing the retrievable identity (query set) and a set of images, the model is streamlined to extract fore discriminative representations from the query and gallery images. Supervised re-ID methods [1] learn the representation by using the human-annotation labels, which is time-consuming and hard to implement in real-world deployments. Towards this, unsupervised algorithms are increasingly researched for omitting the cumbersome human annotation.

Unsupervised person re-ID directly learns instance representations from unlabeled images. Most of the previous unsupervised re-ID methods are proposed based on unsupervised domain adaptation (UDA) [2]. In the case of UDA methods, they leverage annotated labels from a source dataset to fine-tune the model to adapt to the target domain. Although many UDA methods [3] achieve great results on some large-scale source datasets, the performance is impacted by the quality and scale of the source dataset. Otherwise, in real-world cases, facing various situations, the dataset with detailed annotation is not available all the time. Under this kind of situation, fully unsupervised person re-ID methods [4] are more flexible and low-cost and do not require annotation or leverage of other dataset.

As the content mentioned above, this work is employed for the fully unsupervised re-ID task. Previous related methods mainly adopted clustering, k-nearest neighbors, or graph-based techniques to generate pseudo labels for learning. Recently, many existing methods apply the clustering-based method with contrastive learning as the loss function to train the unsupervised person re-ID framework and gain excellent performance. Those clustering-based methods generate pseudo labels for the samples which are similar enough to be clustered as a cluster. The contrastive learning controls the relative distance among the samples in the same cluster and different clusters to improve the quality of the extracted feature. In one cluster, the samples included could be taken from the same identity or not, it's decided by the discriminative ability of the extracted feature. So the clustering process would harm the performance of the re-ID model when it cannot get robust and accurate clustering results. Self-paced Contrastive Learning (SpCL) [5] generates the centroid of the multiple positive samples to solve the above problem, which includes the feature of each positive sample. Then through the contrastive learning loss function, the samples would converge to its centroid at a uniform pace. Although SpCL gained impressive results, it ignores the invariance for inter-camera samples, which caused lower converging speed and performance. Camera-aware Proxies (CAP) [6] alleviates the problem by splitting every single cluster into each proxy using the camera label. Each proxy represents the centroid of samples captured from the same camera. CAP considers not only the intra-camera case but also the inter-camera case and achieves state-of-the-art performance in unsupervised person re-ID tasks.

Recently, attribute-based methods [35] proved their effectiveness by providing semantic features in deep learning. Inspired by GPS [7], which proposed the Graph-based Person Signature to combine information of attribute embedding and body part embedding, we combined the GPS part with an unsupervised person re-ID baseline. Then we employ the GCN to model the correlation between attributes and body parts features. The proposed method is named Attribute Enhancement Contrastive Learning (AECL) which utilizes the GCN and attribute label to improve the extracting ability of the backbone

network. The architecture of the proposed framework is shown in Fig. 1. We conduct an extensive comparison with the state-of-the-art methods in Table. III.

II. Related Works

A. Unsupervised Person Re-identification

The existing unsupervised person can be divided into three categories. Firstly, some methods applied the traditional unsupervised algorithm [10], [23], [24] to solve the challenging task. Most of them are proposed before deep learning using CNNs sprang up.

They did great research on feature representation and metric learning, which solve the various illumination and viewpoints changes and learn more samples to get the discriminative distance metric, respectively. And the complex environmental condition impacted the framework so much compared with the other two categories. The methods included in the second category employed the CNN models as the backbone to extract deep features and leveraged the clustering-based method to generate pseudo labels for the training samples. A typical method is proposed by Yang *et al*. [25], which divided the input samples into upper body parts and bottom body parts then exploited the similarity between whole samples feature and body part feature. It's helpful for generating clusters in inter-camera cases. Different from some methods that directly leveraged the clustering-base algorithm, Lin *et al*. [14] innovatively proposed a bottom-up clustering framework that treats each sample as a single cluster and integrates similar clusters iteratively. The third category utilizes the annotated label from another person re-ID dataset to train the backbone network on the target dataset. ECN [18] mainly focuses on reducing the impact of intra-domain variations. The proposed exemplar memory which is applied for storing features of samples is first used in person re-ID task and achieved great performance. Also, there are some algorithms different from the above categories, The GCL [17] utilized the pre-produced human skeleton mesh and GAN method to generate the multiple-view images of a sample, which improve the diversity for contrastive training.

B. Graph Models

Graph Convolutional Networks (GCN) [31], [32] is proposed to learn the relation between graph nodes. Then many different GNN variations are proposed and applied in many computer vision applications, such action recognition [27], anomaly detection [28], recommendation system [29] and person re-ID [7], [30]. Among them, [7] constructed graph using body-part feature and attribute feature to improve the discriminative extracting ability of backbone network, [30] employed with multiple source dataset and leveraged GNN model to minimize the domain gap by integrating features of different domains.

III. Methodology

A. Overview

In this section, we introduce the preliminary of unsupervised person re-ID baseline in Sec III.B, which shown the strategy of camera-agnostic person re-ID pipeline for most paper. Sec III.C demonstrate the methodology of camera-aware case, which includes the definition of proxy and camera-aware loss function. Sec III.D shows the definition of Graph-based Person Signature followed [7]. Then after constructing the graph, Sec III.E applies the GCN for processing the integration feature and the details of GCN. Finally, the loss function for guaranteeing the effectiveness the attribute enhancement branch.

B. Problem formulation and Overview

The target of fully unsupervised person re-ID is to train robust model for extracting discriminative features on the unlabeled target dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, where the signal N means the number of pedestrians included. For the whole re-ID task, the $\mathcal{X}$ is split into a training set for training the model and a testing set for the inference process, where the training set can be denoted as $\mathcal{T}$, and the testing set should be divided into a query set $\mathcal{Q}$ and a gallery set $\mathcal{G}$. In the inference process, when giving samples from $\mathcal{Q}$, we want to retrieve all the samples of the same pedestrian from $\mathcal{G}$.

In every training epoch, the backbone network [19] extracts the feature vector $\mathcal{M} = \{m_1, m_2, ..., m_N\}$ when giving the training set. Then for generating the pseudo label for each sample, we employ the algorithm DBSCAN [21] to do the clustering process, and gain the pseudo labels $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$, where $y_i \in \{-1, 0, 1, 2, ..., N_c\}$. If the value of $y_i = -1$, then the sample is the outlier in this training epoch, and if $y_i > -1$ means the sample is an inlier. This paper only chooses the inlier samples for contrastive training. The centroid of the a-th cluster is computed by,

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

where the signal N_a denotes the number of samples in the a-th cluster. In this section, we don't discuss the condition with camera label, so it is the camera-agnostic case. And the contrastive loss for camera-agnostic case is formulated by,

$$\mathcal{L}_{agnostic} = \mathbb{E}\left[-log\frac{exp(f_a \cdot p_a/\tau_a)}{\sum_{i=1}^{|p|} exp(f_a \cdot p_i/\tau_a)}\right] \tag{2}$$

where τ_a is the temperature factor for expanding the gap among the values and $|p|$ is the number of clusters in the current training epoch. The product of feature and centroid feature on the numerator is the positive set while the products on the denominator is the negative sets.

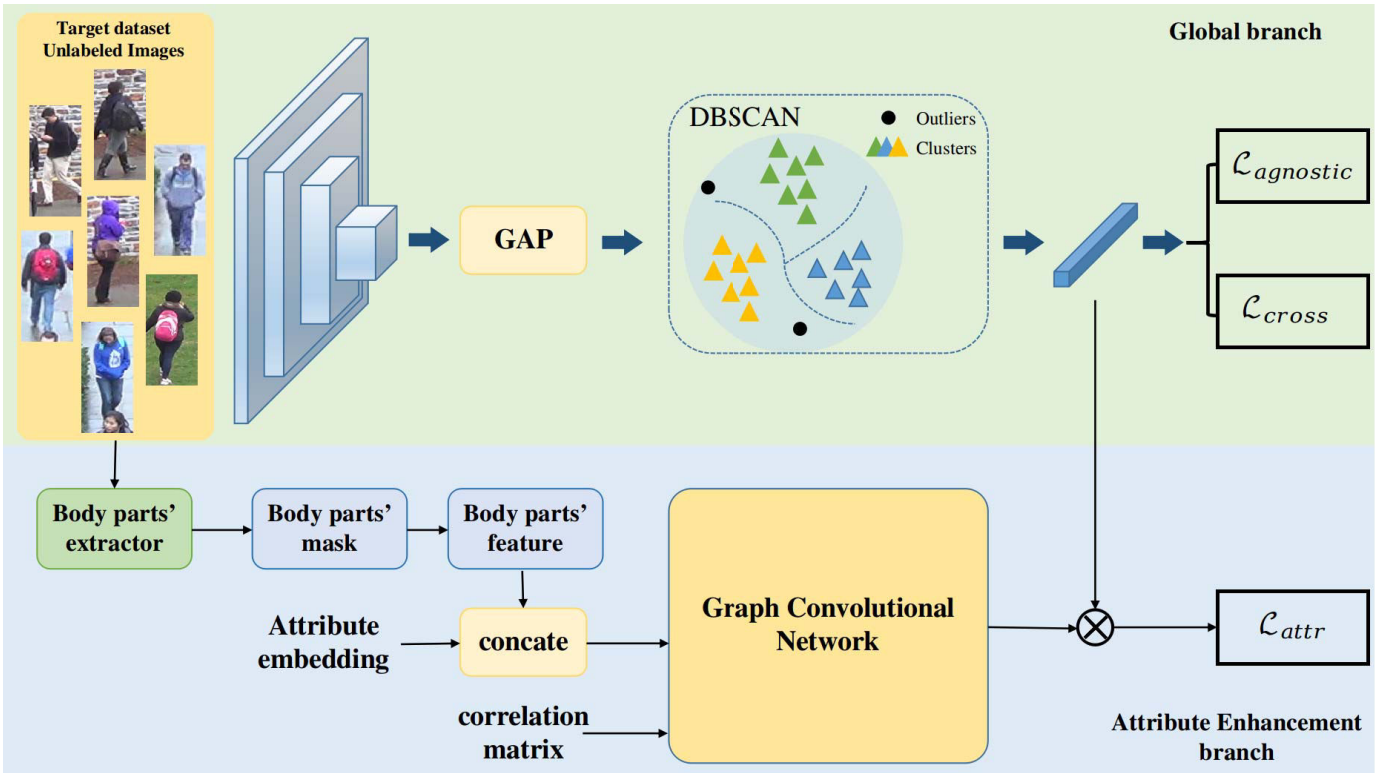

Fig. 1. The overview of the proposed model for unsupervised person re-ID. The upper part of the framework employ the backbone network, clustering method and two versions of loss function, which named global branch in this paper. The bottom part shows the structure about integrating human body parts feature and attribute embedding, which called attribute enhancement branch.

C. Proxy Centroid Contrastive Baseline

Given the camera label $\mathcal{C} = \{c_1, c_2, ...\}$ of each training samples, the proxy p_{ab} is defined as the contorid of the samples including in the a-th cluster and captured by camera c_b:

$$p_{ab} = \frac{1}{N_{ab}} \sum_{m_i \in y_a \cap m_i \in c_b} m_i \tag{3}$$

where the signal N_{ab} denote the number of samples including in the a-th cluster and captured by camera c_b.

The sample including in the a-th cluster and captured by camera c_b is denoting as f_{ab}. With the camera label, we set the camera-aware contrastive loss function as follows,

$$\mathcal{L}_{cross} = \mathbb{E}\left[-\frac{1}{|P|} log \sum_{i \neq b \cap i \in C} \frac{exp(< f_{ab} \cdot p_{ai} > /\tau_c)}{\sum_{j=1}^{N_{neg}+1} exp(< f_{ab} \cdot p_j > /\tau_c)} \right] \tag{4}$$

where $< \cdot >$ means the cosine similarity, and the signal τ_c is the temperature hyper-parameter of camera-aware contrastive loss, $|P|$ is the number of positive proxies in a training epoch. The whole contrastive loss is calculated by,

$$\mathcal{L}_{proxy} = \mathcal{L}_{agnostic} + 0.5\mathcal{L}_{cross} \tag{5}$$

D. Graph-based Person Signature

Followed GPS [7] setting, construct the graph contain the relationship of human body parts and associated attribute $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the signal $\mathcal{V} = \{v_1, v_2, ..., v_N\}$ denotes the nodes, where $N_G = N_A + N_P$ denotes the number of nodes, which is equal to the summation of the number of attributes N_A and the number of human body parts N_P.

The nodes is initialized with a learnable feature embedding. And the adjacency matrix $M \in \mathbb{R}^{N_G \times N_G}$ is representing by,

$$M = \begin{bmatrix} AA & AP \\ PA & PP \end{bmatrix} \tag{6}$$

where $AA \in \mathbb{R}^{N_A \times N_A}$ is the attribute-attribute correlation matrix, $PP \in \mathbb{R}^{N_P \times N_P}$ is the parts-parts correlation matrix, $PA \in \mathbb{R}^{N_P \times N_A}$ is the parts-attributes correlation matrix, and $AP \in \mathbb{R}^{N_A \times N_P}$ is the attribute-parts correlation matrix.

1) The attributes-attributes matrix: The elements AA_{ij} denotes the situation when attribute i and j occurs in the same sample. So it's computed as the co-occurrence number L_{ij} divide by the occurrence time of attribute i.

$$AA_{ij} = \frac{L_{ij}}{K_i} \tag{7}$$

2) The parts-parts matrix: The value of the elements in PP is set to 1 because the human body parts almost could be tested.

3) The parts-attributes matrix: The value of the elements in PA denote the situation that attribute i occurs in the associate body part. The human body part and the related attribute is shown in Table. I. So if attribute occurs, the value is set to 1,; otherwise, it is 0.

TABLE I
HUMAN BODY PARTS AND THEIR ASSOCIATED ATTRIBUTES IN MARKET-1501 DATASET [9].

Body Part	Detailed Attribute
Foreground	gender, young, teenager, adult, old
Head	hair, hat
Upper body	backpack, upper clothing's type, up-black, up-white, up-red, up-purple, up-yellow, up-gray, up-blue, up-green
Lower body	lower clothing's type, lower clothing's length, down-black, down-pink, down-purple, down-yellow, down-gray, down-blue, down-green, down-brown, down-white
Arm	sleeve length, bag, handbag

4) The attributes-parts matrix: The AP and PA are diagonally symmetric.

5) Body parts representation: The state-of-the-art algorithm SCHP [33] which pre-trained on LIP dataset [34] is employed to generate the human body parts' mask beforehand. The mask is resized to the same size of global feature of the final stage of the backbone and then L1 normalization to get $h_i^k \in \mathbb{R}^{W\times H\times D}$. Then multiply with the global feature, which denoted as $f_{part}^k \in \mathbb{R}^{N_P\times D}$,

$$f_{part}^k = \sum_{N_P}^{i=1} h_i^{(k)} f_i \tag{8}$$

where $h_i^{(k)}$ is the scalar value value at the location i of H_k'. The $f_{part}^{(k)}$ is projected to a D_w-dim vector.

E. Graph-based Person Signature: Process GCN

In this paper, two layer of GCN is applied for integrating and extracting the input features with the adjacency matrix. Precisely, each layer in GCN is formulated as a function $f(X, M)$ which giving the nodes $X \in \mathcal{R}^{N_G\times D_w}$ as the input and updating its weight by propagating. Denoting $H^{(k)}$ is the feature matrix after passing the input nodes X to k-th GCN layers. We follow GCN formulation proposed in [31], which takes node features $H^{(k)} \in \mathcal{R}^{N_G\times d}$ and the corresponding correlation matrix M as inputs and pass through a GCN layer to transform to $H^{(k+1)} \in \mathcal{R}^{N_G\times d'}$. According to [31], every GCN layer can be represented as

$$H^{(k+1)} = LeakyReLU(\hat{M}H^{(k)}\theta^{(k)}), \tag{9}$$

where $\theta^k \in \mathcal{R}^{d\times d'}$ is a layer-specific trainable weight matrix and $\hat{M}$ is the normalized version of correlation matrix M. Formally, $\hat{M}$ is defined as:

$$\hat{M} = (I + D)^{-\frac{1}{2}}(M + I)(I + D)^{-\frac{1}{2}} \tag{10}$$

where D is the diagonal degree matrix of M, the identity matrix $I \in \mathcal{R}^{N_G\times N_G}$ is added for forcing the self-loop in G. We aim to learn a set of parameters $\theta = \{\theta^1, \theta^2, ..., \theta^k\}$ that maps X to a set of inter-dependent classifier for person multi-attribute recognition.

F. Attributes Recognition Loss

For guaranteeing the effectiveness of the attribute enhancement branch, the attribute recognition loss is applied to control the accuracy when predicting the attribute class. Denote the the feature extracted by global branch as f_{global} and the attribute prediction extracted by attribute enhancement branch as $\hat{y}$, which is from the GCN part. And the ground truth of the attribute of image x_i is denoted as $\mathbf{y} \in \{0, 1\}^{N_A}$, then the attribute loss function is computed by,

$$\mathcal{L}_{attr}^{(i)} = -\frac{1}{N_A}\sum_{c=1}^{N_A} y_c^{(i)} log(\sigma(\hat{y_c}^{(i)})) + (1-y_c^{(i)}) log(1-\sigma(\hat{y}_c^{(i)})), \tag{11}$$

where $\sigma(.)$ is the sigmoid function to control the distribution of data, y_c indicates the result to judge if attribute c occurs. So the attribute recognition loss is computed as

$$\mathcal{L}_{attr} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{attr}^{(i)}, \tag{12}$$

where N is the number of samples in the training set.

The entire loss for model learning is

$$\mathcal{L} = \mathcal{L}_{cross} + \alpha\mathcal{L}_{attr} \tag{13}$$

where α is a parameter to balance the two terms.

IV. EXPERIMENTS

A. Dataset and evaluation metrics

Market-1501 [9], *DukeMTMC-reID* [8] datasets are two iconic and widely used public person re-ID datsets. The detail of the two dataset is shown in Table. II. DukeMTMC-reID is harder to train a great model due to the more occlusions in the images. And the inference results are evaluated by Mean average precision (mAP) and the Cumulative Matching Characteristic (CMC) Rank-1/5/10 matching accuracy.

B. Implementation Details

For the backbone network, we utilize the ResNet-50 [19] pre-trained on ImageNet [20]. The final fully connected layer is replaced with a L2 Normalization layer after the global average pooling operation to get the feature vector. All the input samples of the training set are resized into 256×128 with some data augmentation, such as random horizontal flipping, add padding, random cropping, and random erasing. The batch is set as 32, where randomly sample from 8 proxies with 4 samples per proxy.

In every training epoch, the Jaccard distance with k-nearest neighbors is computed for the clustering process. The DBSCAN is employed as the clustering method with a threshold

TABLE II
DETAILED INFORMATION OF DATASET MARKET-1501 AND DUKEMTMC-REID.

Dataset	#ID	#ID detail			#image	#image detail			#cam
		Train	Query	Gallery		Train	Query	Gallery	
Market-1501 [9]	1,501	751	750	751	32,668	12,936	3,368	1,6364	6
DukeMTMC-reID [8]	1,404	702	702	1,110	36,411	16,522	2,228	17,661	8

TABLE III
UNSUPERVISED PERSON RE-ID PERFORMANCE COMPARISON WITH STATE-OF-THE-ART 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 [10]	CVPR15	None	27.2	41.6	49.1	8	None	12.3	21.3	26.6	4.8
BOW [11]	ICCV15	None	35.8	52.4	60.3	14.8	None	17.1	28.8	34.9	8.3
UDML [12]	CVPR16	None	34.5	52.6	59.6	12.4	None	18.5	31.4	37.6	7.2
DECAMEL [13]	TPAMI18	None	60.2	76	81.1	32.4	-	-	-	-	-
BUC [14]	AAAI19	None	66.2	79.6	84.5	38.3	None	47.4	62.6	68.4	27.5
DBC [15]	BMVC19	None	69.2	83	87.8	41.3	None	51.5	64.6	70.1	30
MMCL [16]	CVPR20	None	80.3	89.4	92.3	45.5	None	65.2	75.9	80	40.2
GCL [17]	CVPR21	None	87.3	93.5	95.5	66.8	None	82.9	87.1	88.5	62.8
SpCL [5]	NeurIPS20	None	88.1	95.1	97.0	73.1	None	81.2	90.3	92.2	65.3
CAP [6]	AAAI21	None	91.4	96.3	97.7	79.2	None	81.1	89.3	91.8	67.3
Ours(AECl)	This paper	None	**92.0**	**96.5**	**97.8**	**81.0**	None	**81.9**	**89.6**	**92.0**	**67.7**

TABLE IV
UNSUPERVISED PERSON RE-ID PERFORMANCE COMPARISON WITH DIFFERENT VERSION OF CAP [6] ON MARKET1501, DUKEMTMC-REID DATASETS.

Model	Market-1501				DukeMTMC-reID			
	mAP	R1	R5	R10	mAP	R1	R5	R10
CAP camera-agnostic	62.9	79.7	88.3	91.2	57.5	74.3	82.7	86.0
AECL camera-agnostic	65.8(**+2.9**)	83.5(**+3.8**)	93.0(**+4.7**)	97.3(**+6.1**)	63.2(**+5.7**)	78.1(**+3.8**)	87.8(**+5.1**)	90.9(**+4.9**)
CAP camera-aware	79.2	91.4	96.3	97.7	67.3	81.1	89.3	91.8
AECL camera-aware	81.0(**+1.8**)	92.0(**+0.6**)	96.5(**+0.2**)	97.8(**+0.1**)	67.7(**+0.4**)	81.9(**+0.8**)	89.6(**+0.3**)	92.0(**+0.2**)

of 0.5 and minimum samples of 4. For the AECL camera-aware case training, we would not compute the inter-camera loss in the first five epochs. The α for balancing the cross-camera loss and the attribute recognition loss is set to 0.2. The temperature factor is set as $\tau_a = 0.2$ and $\tau_c = 0.07$.

The ADAM is applied as the optimizer. And the initial learning rate is 0.00055 with a warm-up scheme in the first ten epochs and is divided by 10 after every 20 epochs. The total epoch is 50. The model is implemented by PyTorch Toolbox and trained on 1 Nvidia GTX 1080Ti GPU.

C. Comparison with the state-of-the-arts

We compare the performance of the proposed AECL with some state-of-the-art unsupervised person re-ID works on *Market-1501* [9], *DukeMTMC-reID* [8].

We compare the performance with LOMO [10], BOW [11], UDML [12], DECAMEL [13], BUC [14], DBC [15], MMCL [16], GCL [17], and SpCL [5] amd the baseline CAP [6].

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 innovatively proposed a bottom-up clustering framework which treat each sample as a single cluster and integrate the similar clusters iteratively.. MMCL proposed memory-based multi-label classification loss and SpCL is based on proxy contrastive learning, which are considered respectively as camera-agnostic and camera-aware baselines in our method. GCL utilized the pre-produced human skeleton mesh and GAN method to generate the multiple-view images of a sample, which improve the the diversity for contrastive training.. And the CAP is the baseline of this paper. It is obvious that our proposed model outperforms other works with a large margin. The comparison results are shown in Table. III.

For instance, Table. IV compared the baseline with the state-of-the-arts. Within camera-agnostic mode, our approach obtains 3.8% Rank-1 and 2.9% mAP gain on Market, 3.8% Rank-1 and 5.7% mAP gain on DukeMTMC-reID. And for the camera-aware mode, the proposed work obtains 0.6% Rank-1 and 1.8% mAP gain on Market, 0.8% Rank-1 and 0.4% mAP gain on DukeMTMC-reID.

V. CONCLUSION

The graph-based attribute enhancement contrastive learning (AECL) framework is introduced in this paper to solve the unsupervised person re-ID task without leveraging any annotations. The global branch extracts the discriminative

feature and the attribute enhancement branch makes the backbone network more robust when recognizing the homologous appearance samples. The performance test on Market-1501 and DukeMTMC-reID dataset prove the effectiveness of the proposed idea.

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] H. Chen, B. Lagadec and F. Bremond, "Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification," 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2472-2481, doi: 10.1109/WACV45572.2020.9093541.

[2] 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

[3] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual meanteaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In ICLR, 2020. 1, 2, 3, 7, 8

[4] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In CVPR, 2020. 1, 2, 7, 8

[5] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In NeurIPS, 2020. 1, 2, 3, 7, 8, 10, 11

[6] 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

[7] B. X. Nguyen, B. D. Nguyen, T. Do, E. Tjiputra, Q. D. Tran and A. Nguyen, "Graph-based Person Signature for Person Re-Identifications," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 3487-3496, doi: 10.1109/CVPRW53098.2021.00388.

[8] 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

[9] 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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

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

[17] Chen, H., Wang, Y., Lagadec, B., Dantcheva, A., and Bremond, F., Joint Generative and Contrastive Learning for Unsupervised Person Re-identification, ¡i¿arXiv e-prints¡/i¿, 2020.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

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

[25] 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.

[26] 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.

[27] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pages 74447452, 2018. 3

[28] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 3

[29] Chen Ma, Liheng Ma, Yingxue Zhang, Jianing Sun, Xue Liu, and Mark Coates. Memory augmented graph neural networks for sequential recommendation. In The ThirtyFourth AAAI Conference on Artificial Intelligence, pages 50455052, 2020. 3

[30] Z. Bai, Z. Wang, J. Wang, D. Hu and E. Ding, "Unsupervised Multi-Source Domain Adaptation for Person Re-Identification," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12909-12918, doi: 10.1109/CVPR46437.2021.01272.

[31] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In ICLR, 2017. 3

[32] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst, 32(1):424, 2021. 3

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

[34] 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

[35] Jinghao Luo, Yaohua Liu, Changxin Gao, and Nong Sang. Learning what and where from attributes to improve person re-identification. In ICIP, 2019. 1, 2, 6, 7