IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier xxxx xxxx

Graph Neural Network Based Attribute Auxiliary Structured Grouping for Person Re-Identification

GEYU TANG[1,2], XINGYU GAO[1], (MEMBER, IEEE), ZHENYU CHEN[1], (MEMBER, IEEE), AND HUICAI ZHONG[1].

[1]Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China.
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Xingyu Gao (e-mail: gxy9910@gmail.com).

This work was supported by National Natural Science Foundation of China under Grant 61702491.

ABSTRACT Recently, person re-identification (re-ID) with weakly labeled or unlabeled data has drawn much attention in open-set and cross-domain re-ID systems especially for the attribute auxiliary weakly supervised person re-ID. Although state-of-the-art clustering-based methods have achieved good performance, the pseudo labels generated through clustering are often low-quality and noisy. To address this problem, we propose a graph neural network based Attribute Auxiliary structured Grouping (A^2G) to improve the confidence of the pseudo labels. Different from the existing clustering-based approaches that only utilize the similarity in feature space, we learn the feature representation from the similarities in both attribute space and feature space by graph learning on the pedestrian attribute graph. Specifically, we first utilize the pair-wise attribute similarity to represent the relation between two pedestrians to construct a pedestrian attribute graph. Furthermore, we aggregate the features from their neighborhood on a pedestrian attribute graph by the graph neural network, which would make the attribute similar pairs closer and simultaneously take the dissimilar pairs further in the feature space. Finally, to avoid the over-confidence of the *hard* pseudo labels, we regularize the learning of the embedding model with the smoothed pseudo label (SPL) in the optimization of the feature embedding network. We conduct extensive experiments on several person re-ID datasets to validate the efficacy of our proposed method. The results demonstrate that our technique is effective and promising for person re-ID tasks.

INDEX TERMS Unsupervised person re-identification, attribute-auxiliary structured grouping, graph neural network.

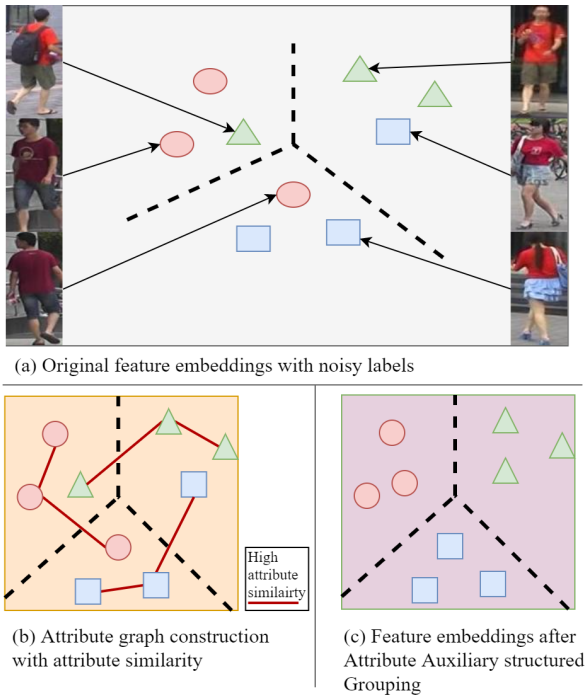
I. INTRODUCTION

PERSON re-identification (person re-ID) aims at matching the same pedestrian's image from a database across different cameras [1]. Learning a discriminative feature embedding network that is invariant to pose, illumination, and camera style is the key for person re-ID system. In the past few years, supervised person re-ID approaches based on deep learning have achieved great success [2]–[5]. However, the performance drops evidently when we change to a new camera system, and the recollection of a large-scale annotated pedestrian dataset is time-consuming and expensive. Hence, unsupervised domain adaptive (UDA) person re-ID has draw much attention recently, which aims to transfer the knowledge of labeled source domain to unlabeled target domain.

Clustering-based self-training UDA approaches [6]–[8] explore the hidden intra-identity similarity and inter-identity dissimilarity in feature space through dynamic assigned pseudo labels. They group the embedding features of unlabeled target-domain images and assign the cluster ID as the pseudo labels. Then, the feature embedding network is optimized in a supervised learning manner. Although the quality of pseudo labels improves iteratively with the increasing discriminative-ability of feature embedding network, the low-confidence and noisy pseudos substantially hinder the learning capacity of the embedding network. In this paper, we focus on improving the quality of pseudo label and suppressing the over-confidence of the *hard* pseudo labels.

We propose a graph neural network based *attribute auxiliary structured grouping* (A^2G) algorithm to refine the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3069915, IEEE Access



(a) Original feature embeddings with noisy labels

(b) Attribute graph construction with attribute similarity

(c) Feature embeddings after Attribute Auxiliary structured Grouping

FIGURE 1: The motivation of graph neural network based Attribute Auxiliary structured Grouping (A^2G). (a) The unlabeled target-domain samples in feature space. Each shape denotes one identity. The appearance-similar samples are misclustered into false clusters. (b) We construct pedestrian attribute graph, where pedestrians are considered as nodes, attribute similarities as edges. The red line represents the high attribute-similarity relation, and low similarity pairs are disconnected. (c) We aggregate the features from their neighborhood on the pedestrian attribute graph via a graph neural network and re-cluster the refined features. Therefore, the former mislabeled samples group correctly with the help of the similarities in attribute space.

pseudo labels with the consideration of the pedestrian similarity in attribute space. Pedestrian attributes are semantic-level annotations for person re-ID task and have been widely studied in supervised training setting [9]–[11]. However, in attribute auxiliary weakly supervised person re-ID, attribute annotations are not been fully explored, especially for clustering based approach. The proposed A^2G is motivated from graph embedding algorithms on graph data. Graph embedding algorithm aims at generating similar embeddings for the inter-connected nodes on the graph [12]. As to attribute auxiliary UDA person re-ID, we hope to propagate the features of unlabeled pedestrian image from their high attribute-similarity neighborhood. By conducting attribute auxiliary feature aggregation on pedestrian attribute graph, we consider the similarity in attribute space and improve the quality of pseudo labels. As shown in Figure 1, the unlabeled target-domain appearance-similar samples cluster incorrectly and are assigned with false pseudo labels. A^2G constructs pedestrian attribute graph and learns node embedding through graph neural network. By grouping on the attribute auxiliary refined features, A^2G provides more robust and high-confidence pseudo labels.

The over-confidence of pseudo label assignment also effects the learning ability of embedding network, we regularize the learning of the embeddig model with smoothed pseudo label (SPL) in the optimization of embedding network. Label smoothing improves the generalization and convergence speed of a multi-classes neural network [13]. Although the label smoothing has been successfully used in many tasks, its efficacy in the optimization of embedding network with *hard* pseudo labels for person re-ID is not been fully explored. In the experiments, the introduce of SPL boosts the performance evidently, which provides a simple but effective technique in reducing the over-confidence of *hard* pseudo labels in clustering-based approaches. Our contributions can be summarized in four-fold:

- We propose a graph neural network based Attribute Auxiliary structured Grouping algorithm to improve the confidence of pseudo labels for attribute auxiliary weakly supervised person re-ID. By aggregating the features from the high attribute similarity neighborhoods in the pedestrian attribute graph, our proposed method could explore the pedestrian similarity in attribute space.
- We formulate the relation between image pairs with the pair-wise attribute similarity and unify the similarities in feature space and attribute space into a pedestrian attribute graph. With the representation learning on the pedestrian attribute graph, our technique could effectively improve the feature embeddings in the pseudo label generation stage.
- We regularize the learning of embedding model with smoothed pseudo label (SPL) in the optimization of feature embedding network, which would relief the over-confidence of the *hard* pseudo labels for learning discriminative embeddings.
- We conduct extensive experiments on person re-ID datasets, including Market-1501, DukeMTMC-reID, and MSMT17, and the encouraging results validate the efficacy of our proposed method in person re-identification.

II. RELATED WORK

A. DISCRIMINATIVE LEARNING WITHOUT IDENTITY LABELS FOR PERSON RE-ID

Discriminative learning without identity labels is a challenging and practical task because it relieves the heavy cost to acquire manual annotations. Unsupervised learning approaches have gained popularity and developed rapidly [14]. ENC [15] explored the target-domain invariance from the following aspects: exemplar, camera, and neighborhood. SSL [10] learned the discriminative features from a softened label. BUC [16] conducted a bottom-up clustering and assigned pseudo labels for unlabeled images with their cluster center.

In [10], a clustering approach with camera-level style transfer was proposed to minimize the cross-camera variance in the learning process. AE [17] learned a non-parameter classification model, where the selection of neighborhoods for each person image was conducted with an adaptive and balanced strategy. As to video person re-ID, DGM [18] estimated the pseudo labels via a dynamic graph co-matching schema, where the quality of labels was improved with strategies including iterative graph structure updating, label re-weighting, and co-matching. In RACE [19], a robust anchor embedding approach was proposed to estimate the labels by mining the underlying similarity between the unlabeled sequences and anchors with affine hull regularization.

Unsupervised and unsupervised domain adaptive (UDA) methods relieve the requirement of penalty of annotated labels by searching transferable knowledge in the labeled source domain. UDA methods in person re-ID are categorized into three groups: domain-transfer based methods, memory-auxiliary contrast learning based methods, and clustering based methods. SPGAN [20] transferred the source-domain images to target-domain while preserving the identity similarity. PTGAN [21] preserved the pedestrian-involved part while transferring the background part. HHL [22] proposed a camera-aware adaptation framework through camera style image generation and domain-separate learning. However, the quality of generated images highly effect the retrieval performance of domain-transfer based methods, and they ignore the relative similarity in the unlabeled target-domain images. MAR [23] learned a soft multi-label from an auxiliary domain to learn identity-discriminative features. MMCL [24] predicted the image label by pair-wise similarity and conducted multi-label classification for feature learning. UDAP [25] proposed a self-training framework and provided a theoretical analysis on UDA re-ID. SSG [7] explored the local and global similarities simultaneously under self-training framework. PAST [8] progressively improved the model performance through triplet-loss-based conservative stage and classification based promoting stage. MMT [6] proposed a soft label assignment through mutual teaching to refine the *hard* pseudo label.

Weakly supervised approaches also learn a discriminative model without labels, but they involve auxiliary annotations of pedestrians, such as attribute annotations and untrimmed video-level labels. TJ-AIDL [26] jointly learned an attribute-semantic and identity-discriminative space, which was transferable to target space. Deep CV-MIML [27] proposed a multi-instance multi-label learning model for discriminative learning with video-level annotations.

Our proposed A^2G is based on the self-training UDA framework, where pseudo labels assignment and model training are conducted alternatively. Different from the above methods, we propose a graph neural network based attribute auxiliary feature aggregation algorithm to improve the quality of pseudo label. Compared with PurifyNet [28] which handles the label "noise" by regularizing the model's output logits, A^2G explores the similarity in attribute space to increase the "confidence" of the pseudo labels. As far as we know, this is an early work on Attribute Auxiliary structured Grouping for weakly supervised re-ID.

B. ATTRIBUTE AUXILIARY PERSON RE-ID

Learning pedestrian embedding with the auxiliary of attribute has draw much attention recently. APR [9] proposed a multi-task learning framework to learn visual representation and attribute representation separately. APDR [29] detected the attribute relevant parts with attribute detection network and extracted the feature of image patch where pedestrian exists. AANET [11] proposed a "soft" attention module to focus on the pedestrian-involved region. TJ-AIDL [26] proposed to transfer the domain knowledge with the domain-shared attribute for UDA re-ID. Compared with other semantic-based person retrieval tasks, such as language-query-based person retrieval [30], A^2G has the potential to explore the similarity in the language queries by formulating them into word embeddings. Different from above methods, we explore the similarity in the attribute space to increase the confidence of pseudo labels in attribute auxiliary weakly supervised re-ID, which has not been fully explored yet.

C. GRAPH REPRESENTATION LEARNING ON COMPUTER VISION TASKS

Graph neural network generalizes the input data of neural network from Euclidean data to graph data (non-Euclidean). Among all types of graph neural networks, convolutional graph neural network is becoming popular because of its computational efficiency and easy adaption with other neural networks. In 2016, Kipf et al. [31] proposed a first-order approximation for graph convolutional operator, which achieved better performance with high computational efficiency in node classification task. Since then, convolutional graph neural network has developed rapidly. Instead of taking all neighborhoods of the node in the forward propagation, GraphSage [12] randomly samples a fixed size of the neighborhood, which is efficient for large-scale graph applications, such as e-commerce and social network. GAT [32] utilizes self-attention mechanisms to learn the relative importance of neighboring nodes to their center node, which is a powerful and efficient architecture. FastGCN [33] aims at reducing the sampling variance by importance sampling strategy, which increases the learning efficiency without accuracy loss.

With the development of deep learning, graph neural network has achieved fabulous performance in many tasks. GCT [34] explored the spatial-temporal structure of historical target exemplar in visual tracking through graph neural network (GCN). ML-GCN [35] modeled the relation of multiple labels through GCN for multi-label image classification. ST-GCN [36] explored both spatial and temporal information via GCN for skeleton-based action recognition. Compared with DDAG [37] which utilizes a graph-based attention module to aggregate the cross-modality part-level feature, A^2G refines the feature representation via a graph neural network on the pedestrian attribute graph to further explore the similarity

of attribute space. In this paper, we construct a pedestrian attribute graph and aggregate the features through graph embedding algorithm, which is a novel approach for attribute-auxiliary person re-ID.

D. LABEL SMOOTHING ON COMPUTER VISION TASKS

Label smoothing was firstly proposed to boost the performance of image classification with cross-entropy loss [38] in supervised learning setting, which transforms the *hard* one-hot label into "soft" label with smoothing regularization. And it is also widely utilized in speech recognition [39], and speech recognition [40]. Although label smoothing is presented in above work, its efficacy on reducing the over-confidence of pseudo labels is ignored in previous clustering-based re-ID methods.

III. GRAPH NEURAL NETWORK BASED ATTRIBUTE AUXILIARY STRUCTURED GROUPING

In this paper, we focus on attribute auxiliary weakly supervised person re-ID. We aim to learn the discriminative ability for unlabeled data by simultaneously exploring the similarity in feature space and attribute space. In the evaluation stage, we extract the features of query and gallery datasets with the learned embedding model. The retrieval results are obtained by calculating the Euclidean distances between the feature embeddings of query and gallery images.

Our proposed graph neural network based Attribute Auxiliary structured Grouping (A^2G) framework aims to address the problems of low-quality and noisy pseudo labels assignment and the "over-confidence" of the *hard* pseudo labels in clustering-based approaches. Our key idea is to formulate the relation between two images by the pair-wise attribute similarity in order to construct a pedestrian attribute graph and propagate the attribute similarity on pedestrian attribute graph via a graph neural network. Through the representation learning on pedestrian attribute graph, we refine the feature embedding with the similarity in attribute space. In addition, discriminative learning with *hard* pseudo labels may amplify the errors in model learning, we smooth the *hard* labels to avoid the problem of "over-confidence". The overall framework of A^2G is described in Figure 2. We firstly transform the images into feature space via a feature embedding network. Then, we formulate the pair-wise relation in feature space with attribute similarity and unify the similarity in feature space and attribute space into a pedestrian attribute graph. Next, we refine the feature with a graph neural network. Finally, we conduct discriminative learning for the feature embedding network.

A. BASELINE: PRE-TRAINING AND SELF-TRAINING

Source domain pre-training: Let $\mathbb{D}^s = \{(x_i^s, y_i^s)|_{i=1}^{N_s}\}$ be the source training dataset with $\mathcal{M}^s$ pedestrian identities (classes), where x_i^s and y_i^s are the i-th image and its manual annotation, respectively. For person re-ID task, we aim to learn a discriminative embedding network $E(x_i|\boldsymbol{\theta})$ that is utilized to generate image feature representation $\mathbf{f}_i \in \mathbb{R}^m$ for the image x_i. The parameters $\boldsymbol{\theta}$ of the embedding network are optimized by an identity classification loss $\mathcal{L}_{id}^s$ and a triplet loss $\mathcal{L}_{tri}^s$ as follows,

$$\mathcal{L}_{id}^s = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{L}_{ce}\left(C^s\left(E\left(x_i^s|\boldsymbol{\theta}\right)\right), y_i^s\right), \tag{1}$$

$$\mathcal{L}_{tri}^s = \frac{1}{N_s}\sum_{i=1}^{N_s}\max\left(0, D(x_i^s, x_p^s) - D(x_i^n, x_n^s) + m\right), \tag{2}$$

where $C^s : \mathbb{R}^m \rightarrow \{1, \cdots, M^s\}$ is the identity classifier, $\mathcal{L}_{ce}$ is the cross-entropy loss, $D(\cdot,\cdot)$ is the distance function between features, m is the triplet distance margin, and x_p^s and x_n^s denote the hardest positive and negative sample in the mini-batch, respectively.

Target domain self-training: After the source domain pre-training, we adapt the learned embedding model to target domain through a self-training process. Let $\mathbb{D}^t = \{x_i^t|_{i=1}^{N_t}\}$ be the target training dataset, where N_t is the number of images in target domain. The self-training on target domain conducts the following procedures alternatively: 1) Dynamic pseudo label assignment. 2) Model learning with pseudo label. In the pseudo label assignment stage, we firstly extract the target-domain feature embeddings $\{\mathbf{f}_i^t|_{i=1}^{N_t}\}$ with feature extraction neural network $E(\cdot|\boldsymbol{\theta})$. Then, we assign pseudo labels $\{\tilde{y}_i^t|_{i=1}^{N_t}\}$ by clustering the target domain features $\{\mathbf{f}_i^t|_{i=1}^{N_t}\}$ into $\mathcal{M}^t$ classes. In the model learning stage, we aim at improving the discriminative ability of embedding model $E(\cdot|\boldsymbol{\theta})$. Let $\tilde{\mathbb{D}}_t = \{(x_i^t, \tilde{y}_i^t)|_{i=1}^{N_t}\}$ be the target domain dataset with pseudo labels. We optimize the parameters $\boldsymbol{\theta}$ of feature embedding network $E(\cdot|\boldsymbol{\theta})$ according to Eq. (1) and Eq. (2) on this new generated target domain dataset. This self-training process with dynamic label assignment and model optimization is conducted alternatively until the model converges.

B. ATTRIBUTE AUXILIARY STRUCTURED GROUPING FOR PSEUDO LABEL GENERATION

The performance of clustering-based self-training framework highly depends on the quality of pseudo labels. Existing approaches only utilize the similarity in feature space. However, the domain variances of illumination, pose, and camera may introduce false positive samples in pseudo labels, which is harmful to the discriminative learning of the embedding model. To overcome the above issues, we take the similarity in attribute space into consideration and propose a graph neural network based attribute-auxiliary feature aggregation algorithm to explore the similarity in attribute space.

Let $\mathbb{D}_a^t = \{(x_i^t, \mathbf{a}_i^t)|_{i=1}^{N_t}\}$ be the target domain pedestrian attribute dataset, where $\mathbf{a}_i^t \in \mathbb{R}^A$ are the attribute annotations of the i-th pedestrian and A is the number of attribute classes. Note that the attribute annotations are binary vector, where "1" represents the pedestrian has the corresponding attribute, otherwise has not. To capture the similarity in attribute space, we formulate the relation between two pedestrian through pair-wise attribute similarity,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3069915, IEEE Access

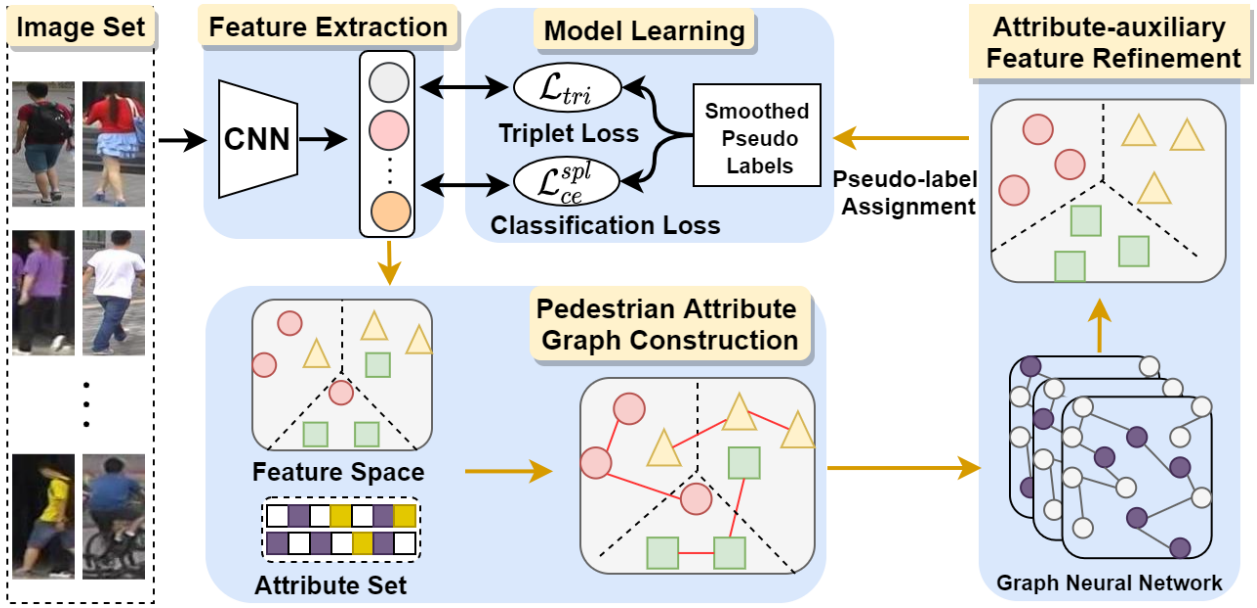


FIGURE 2: The flowchart of our proposed A^{2}G. A^{2}G consists of three key components including pedestrian attribute graph construction, attribute-auxiliary feature refinement, and model learning. Pedestrian attribute graph construction aims to formulate pedestrian relation with attribute similarity and unify the similarity in feature space and attribute space into a pedestrian attribute graph. Attribute-auxiliary feature refinement propagates the attribute similarity of pedestrians with their neighborhood in pedestrian attribute graph to generate the refined features. Model learning improves the discriminative ability of embedding network with the regularization of smoothed pseudo labels (SPL) to relief the "over-confidence" problem of *hard* pseudo labels.

so that the image pairs that are highly similar in attribute space are inter-connected in the pedestrian attribute graph. According to the pair-wise relations and feature embeddings, we could construct a pedestrian attribute graph. To propagate the pair-wise attribute-similarity, we aggregate the features with their neighborhood on pedestrian attribute graph via a graph neural network. After the graph representation learning on pedestrian attribute graph, we obtain the refined features that unify the similarities in feature space and attribute space.

Pedestrian attribute graph construction. Let $G = (V, E)$ be the pedestrian attribute graph, where V ($|V| = N_t$) and E are the sets of nodes (pedestrian) and edges (pair-wise attribute similarity), respectively. Let $\mathbf{X} \in \mathbb{R}^{N_t \times m}$ be a matrix containing the features of nodes, and its row vector is initialed as $\mathbf{X}_i = \mathbf{f}_i^t$. We define the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N_t \times N_t}$ of G as follow,

$$\mathbf{A}_{ij} = \begin{cases} 1 & \mathsf{sim}(\mathbf{a}_i, \mathbf{a}_j) > \tau \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathbf{a}_i$ and $\mathbf{a}_j$ are the attributes of the i-th and j-th pedestrian respectively, $\mathsf{sim}(\cdot,\cdot)$ is the cosine similarity function, and τ is the similarity threshold. Note that the attribute annotations are binary vector, hence the cosine similarity could properly represent the pair-wise attribute annotations similarity. Specifically, we present the $\mathsf{sim}(\cdot,\cdot)$ as follow,

$$\mathsf{sim}(\mathbf{a}_i, \mathbf{a}_j) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{\|\mathbf{a}_j\| \cdot \|\mathbf{a}_j\|}, \tag{4}$$

where $\|\cdot\|$ is l_2 norm. Then, we normalize the adjacency matrix as, $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is the degree matrix (diagonal matrix) of $\mathbf{A}$ with $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, $\sum_j$ is the column-wise sum.

Attribute auxiliary feature refinement. To propagate the similarity in attribute space, we conduct feature aggregation on the pedestrian attribute graph. Recently, Graph Neural Networks (GNN) have achieved superb performance in learning representative embedding on graph. GraphSage [12] is an inductive multi-layer neural network that operates on a graph, which learns feature representation of nodes based on a aggregation of their neighborhoods. In this paper, original target domain features $\{\mathbf{f}_i^t|_{i=1}^{N_t}\}$ are updated from their local neighborhood by an attribute auxiliary feature aggregation through an unsupervised node embedding algorithm.

Suppose that the number of layer in graph neural network is K, for each node (pedestrian) $v \in V$, we have the aggregated hidden feature at k-th layer from its local neighborhood as follow,

$$\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow f_{aggregate}^k \left(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\}\right), \tag{5}$$

where $\mathcal{N}(\cdot)$ and $f_{aggregate}^k(\cdot)$ are node neighborhood sample function and feature aggregation function at k-th layer. After obtain the aggregated features from their neighborhood, we have the updated hidden feature of v as follows,

$$\mathbf{h}_v^k \leftarrow \sigma(\mathbf{W}^k \cdot (\mathbf{h}_v^{k-1} \| \mathbf{h}_{\mathcal{N}(v)}^k)), \tag{6}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3069915, IEEE Access

where $\mathbf{W}^k$ is the trainable parameter at k-th layer of GNN, σ is the activation function, $\|$ is the concatenation operator. Note that we initialize the node embedding $\mathbf{h}_v^0$ with the original target domain feature $\mathbf{f}_i^t$, as $\mathbf{h}_v^0 = \mathbf{X}_i = \mathbf{f}_i^t$. After this forward propagation, we obtain the final refined feature as $\mathbf{z}_v = \mathbf{h}_v^K$, and the learning details is presented in Algorithm 1.

Optimization of the graph neural network. In order to learn representative embeddings that reveal the similarity in attribute space, we apply an unsupervised graph-based loss to optimize the parameters $\boldsymbol{\theta}_{gnn} = \{\mathbf{W}_i|_{i=1}^K\}$ in neural network as follows,

$$\mathcal{L}_{att} = -\log\left(\sigma\left(\mathbf{z}_v^\top \mathbf{z}_u\right)\right) - Q \cdot \mathbb{E}_{u_n \sim P_n(u)} \log\left(\sigma\left(-\mathbf{z}_v^\top \mathbf{z}_{u_n}\right)\right), \tag{7}$$

where v and u are co-occurs nodes on fixed-length random walk, $P_n(u)$ is a negative sample distribution that consists of the disconnected edges of u, Q is the negative penalty parameter. Through this learning process, the embeddings of those pedestrian with high attribute similarity became closer, which exploits the similarity in attribute space.

After obtaining the attribute auxiliary features $\{\mathbf{z}_i^t|_{i=1}^{N_t}\}$, we calculate the pair-wise distance matrix $\mathbf{D}$, where $\mathbf{D}_{ij}$ denotes the distance between i-th and j-th image embeddings. To perverse local similarity, we adopt the k-reciprocal distance [41], which is calculated by the Jaccard distance between the nearest neighborhood sets. We adopt the a density-based clustering algorithm DBSCAN [42] to grouping the images into different clusters, where we assign the pseudo labels of target-domain images with their cluster IDs. In addition, we discard the image samples that not belong to any clusters to reduce the noise in pseudo labels. In the end, we obtain the new generated target domain dataset $\tilde{\mathbb{D}}_t = \{(x_i^t, \tilde{y}_i^t)|_{i=1}^{N_t}\}$ with pseudo labels $\{\tilde{y}_i^t|_{i=1}^{N_t}\}$.

C. MODEL LEARNING WITH SMOOTHED PSEUDO LABELS

In model learning stage, we aim at improving the discriminative ability of embedding network $E(\cdot|\boldsymbol{\theta})$ on target domain data. Given the target-domain dataset with attribute auxiliary pseudo labels $\tilde{\mathbb{D}}_t = \{(x_i^t, \tilde{y}_i^t)|_{i=1}^{N_t}\}$, we consider to optimize the embedding network $E(\cdot|\boldsymbol{\theta})$ with *hard* pseudo labels according to Equation (1) and (2). However, dynamic labels assignment through clustering are often noisy: 1) Lack of human annotations may introduce more false positive samples, i.e., same pseudo labels are assigned to the images with different identity labels, due to the poses and cameras variances. In addition, appearance similar image pairs may in-large the false positive samples, since CNN usually generates similar feature embeddings for visual similar image pairs. 2) The false positive samples may amplify the errors in the learning of feature embedding network, which would increase the noisy in pseudo labels.

To overcome the over-confidence of the *hard* pseudo labels, we propose to regularize the learning of feature embedding network with the smoothed pseudo labels (SPL). We denote the cross-entropy loss with *hard* pseudo labels as $\mathcal{L}_{ce}$ in the form of,

$$\mathcal{L}_{ce} = \sum_{k=1}^{\mathcal{M}_t} -\tilde{\mathbf{y}}_k log(p_k) \tag{8}$$

where $\tilde{\mathbf{y}}_k$ equals "1" for the pseudo class label and "0" for the rest, p_k is the predicted probability of the k-th class, which is obtained by applying softmax operation on the output of classifier. We regularize the optimization of embedding network via smoothed pseudo label (SPL) with smoothing parameter α, and the smoothed pseudo label distribution $\tilde{\mathbf{y}}_k^{spl}$ is presented as follow,

$$\tilde{\mathbf{y}}_k^{spl} = \begin{cases} \frac{\alpha}{\mathcal{M}_t} & k \neq \tilde{y} \\ 1 - \alpha + \frac{\alpha}{\mathcal{M}_t} & k = \tilde{y} \end{cases} \tag{9}$$

We denote the identities classification loss with smoothed pseudo label (SPL) regularization as $\mathcal{L}_{ce}^{spl} = \sum_{k=1}^{\mathcal{M}_t} -\tilde{\mathbf{y}}_k^{spl} log(p_k)$. Hence, the overall loss function for optimization feature embedding network $E(\cdot|\boldsymbol{\theta})$ is presented as follows,

$$\mathcal{L}_{all} = \mathcal{L}_{ce}^{spl} + \lambda \mathcal{L}_{tri} \tag{10}$$

where the λ is the parameter weighting the two losses, $\mathcal{L}_{tri}$ is the triplet loss presented in Equation (2).

D. ALTERNATIVELY TRAINING

We progressively explore the similarity in feature and attribute space and improve the quality of pseudo labels via an alternatively training schema. Different from existing approaches that only consider the similarity in the feature space, the graph neural network based attribute-auxiliary feature aggregation algorithm takes the pair-wise similarity in attribute space into consideration and increases the confidence of pseudo labels. Considering the "over-confidence" of inaccurate label may be harmful to the discriminative learning, we regularize the learning of the embeddig model with smoothed pseudo labels (SPL) when training with cross-entropy loss for identity classification. With the improvements in pseudo label generation and model learning, we could reduce the noise in pseudo labels and conduct the discriminative learning more effectively.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION METRICS

Market-1501 [43] contains 32,668 images of 1,501 identities captured from 6 cameras. The train set contains 12,936 images with 751 identities. The test set is split into query and gallery sets. The query and gallery contain 3,368 and 15,913 images, respectively. 27 human-annotated attributes are presented for each image [9]. We abbreviate Market-1501 as Market in this paper.

DukeMTMC-reID is a subset of the DukeMTMC [44] dataset. It has 702 identities with $16,522$ images for training and 702 identities with 19,889 images for testing. Each image has 23 human annotated attributes [9]. We will abbreviate DukeMTMC-reID as Duke.

MSMT17 is the largest person re-ID dataset with 126,441 images of 4,101 identities under 15 cameras [21]. The training set contains 32,021 images of 1,041 identities. During testing, it consists of 11,659 and 82,161 images of 3,060 identities for query and gallery set, respectively.

Evaluation Metrics. Following these works [6], [7], [10], we utilize Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) to evaluate the person re-ID task. In the comparison with state-of-the-art approaches, we report Rank@1, Rank@5, and Rank@10 accuracies.

B. IMPLEMENTATION DETAILS

For feature embedding network $E(\cdot|\boldsymbol{\theta}_{cnn})$, we adopt ResNet-50 [58] and IBN-ResNet-50 [59] as backbone networks in our experiments. The dimension of feature embedding is 2048. We normalize all images and resize them into 256×128. In the UDA training setting, we firstly initialize the backbone network with ImageNet [60] pre-trained weights. Then, we pre-train the model in source domain with triplet (Eq. (2)) loss and identity-classification loss (Eq. (1)). We take ADAM as model optimizer with a weight decay of 0.0005. The initial learning rate is 0.00035 with learning rate decay, which is decreased to its 1/10 at 40-th and 70-th epoch in the total 80 epochs. In the unsupervised training setting, we directly initialize the model with ImageNet pre-trained weights without source-domain pre-training.

For target-domain self-training, we set the learning epochs to 40 for alternative training. We optimize the embedding network $E(\cdot|\boldsymbol{\theta})$ with Eq. (10) and keep the learning rate and optimizer the same with source-domain pre-training stage. In attribute-auxiliary feature aggregation, we set the number of layers in graph neural network to 2 and the dimension of hidden feature to 2048. We optimize the graph neural network with Eq. (7) by ADAM optimizer. The learning rate is set to 0.0003. The threshold τ in pedestrian attribute graph construction is set to 0.95, balance parameter λ in $\mathcal{L}_{all}$ (Eq. (10)) is set to 1.0, smoothing parameter α in Eq. (9) is set to 0.1, The aggregation function $f_{aggregate}$ (Eq. (5)) is GCN [31].

C. COMPARISON WITH THE STATE-OF-THE-ARTS

We compare our proposed A^{2}G framework with unsupervised, unsupervised domain adaptation, and attribute auxiliary weakly-supervised methods on four cross-dataset person re-ID tasks: Duke-to-Market, Market-to-Duke, MSMT-to-Duke, and MSMT-to-Market. We compare three types of approaches, including unsupervised learning methods: PUAL [45], BUC [16], SSL [10], HCT [46], D-MMD [49], CSE [10], and TAUDL [47], transfer learning based methods: SPGAN [20], HHL [22], CFSM [48], ENC [15], UDATP [25], UCDA-CCE [50], PDA-Net [51], PCB-PAST [8], SSG [7], MMCL [24], DG-NET++ [52], B-SNR+GDS-H [53], DGNET [3], OG-Net [54], AE [17], and AD-Cluster [55], and attribute auxiliary weakly supervised method: TJ-AIDL [26].

As shown in Table 1, on Duke-to-Market, compared with the state-of-the-art method DG-NET++ [52], we achieve 9.9% and 5.3% gains on mAP and Rank@1, respectively. In addition, evident 2.9% and 6.7% gains in mAP (2.5% and 4.5% in Rank@1) are achieved on MSMT-to-Duke and MSMT-to-Market, respectively. The above results validate the effectiveness of our proposed A^{2}G framework. A^{2}G explores the similarity in attribute space through an effective attribute auxiliary feature aggregation to improve the quality of pseudo labels, which is ignored in previous methods. The experimental results also prove the necessity of regularizing the learning of the embeddig model with smoothed pseudo labels (SPL), which has boosted the performances by large margins.

Our proposed A^{2}G could also be extended to unsupervised learning, and we compare A^{2}G with state-of-the-art unsupervised approaches. In unsupervised setting, we initialize the embedding network with ImageNet pre-trained weights instead of pre-training on source-domain. A^{2}G surpasses all compared unsupervised approaches by a considerable margin of 15.2% mAP on and Duke-to-Market tasks. We also observe a significant gain and 12.1% mAP between the "baseline" and A^{2}G in unsupervised setting on Duke-to-Market. The above experimental results validate the effectiveness of A^{2}G under different training settings.

Compared with attribute auxiliary weakly-supervised approach TJ-AIDL [26], A^{2}G achieves significant improvement on retrieval accuracy, which validates the effectiveness of the graph representation for feature similarity and attribute similarity. Conducting the graph leaning on the pedestrian attribute graph would effectively refine the feature representation with the similarity of attribute space.

D. FURTHER EVALUATIONS

Evaluation of Key Components. We have three key components in our A^{2}G framework: attribute auxiliary feature refinement, model regularization with smoothed pseudo labels (SPL), identity classification loss $\mathcal{L}_{ce}^{spl}$, triplet loss $\mathcal{L}_{tri}$. To analysis the effectiveness of these parts, we conduct component-wise evaluation and present the results in Table 2.

When comparing A^{2}G with "baseline", we observe the improvements of 5.7% and 2.1% mAP respectively on ResNet-50 and IBN-ResNet-50 backbones. The attribute auxiliary feature refinement simultaneously explores the similarities in feature space and attribute space, which removes the false positive samples caused by the variances of appearance, pose, and illumination in the feature-subspace clusters. We validate the necessity of regularization with smoothed pseudo labels (SPL) by removing the smoothing penalty, i.e., $\alpha = 0$ in Eq. (9). Such experiments are represented by "A^{2}G (w/o SPL)". 4.4% and 3.6% mAP drops are observed on these two backbones. Because the *hard* pseudo labels are dynamically assigned in each iteration, regularizing the learning of the embeddig model with SPL could alleviate the problem of over-confidence. The increases achieved by the smoothed pseudo labels demonstrate the effectiveness of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3069915, IEEE Access

TABLE 1: Comparison of retrieval accuracy with state-of-the-arts on Market-1501 [43], DukeMTMC-reID [44], and MSMT17 [21]. In the "setting" column, "Unsupervised" denotes training only with the unlabeled target-domain; "UDA" denotes the unsupervised domain adaptation methods, in which labeled source-domain images are utilized for training. "Weakly" denotes training with attribute annotations. "Baseline" denotes the clustering-based self-training approach with DBSCAN [42]. "SPL" denotes the discriminative model learning with smoothed pseudo labels. The "Auxiliary INFO" column shows the auxiliary information used in model learning.

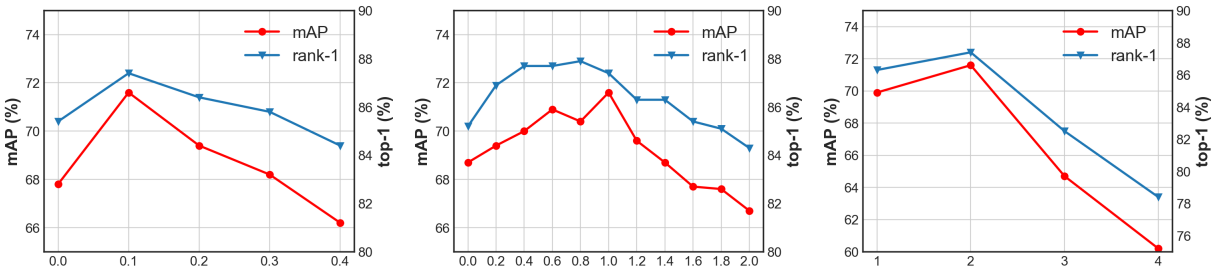
Methods	Setting	Auxiliary INFO	Market-to-Duke				Duke-to-Market			
			mAP	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10
PUL [45]	Unsupervised	None	16.4	30.0	43.4	48.5	20.5	45.5	60.7	66.7
BUC [16]	Unsupervised	None	27.5	47.4	62.6	68.4	38.3	66.2	79.6	84.5
SSL [10]	Unsupervised	Camera ID	28.6	52.5	63.5	68.9	37.8	71.7	83.8	87.4
HCT [46]	Unsupervised	None	50.7	69.6	83.4	87.4	56.4	80.0	91.6	95.2
TAUDL [47]	Unsupervised	Camera ID	43.5	61.7	-	-	41.2	63.7	-	-
CSE [10]	Unsupervised	Camera ID	30.6	56.1	66.7	71.5	38.0	73.7	84.0	87.9
SPGAN [20]	UDA	None	22.3	41.1	56.6	63.0	22.8	51.5	70.1	76.8
HHL [22]	UDA	Camera ID	27.2	46.9	61.0	66.7	31.4	62.2	78.8	84.0
ENC [15]	UDA	Camera ID	40.4	63.3	75.8	80.4	43.0	75.1	87.6	91.6
CFSM [48]	UDA	None	27.3	49.8	-	-	28.3	61.2	-	-
D-MMD [49]	UDA	None	46.0	63.5	78.8	83.9	48.8	70.6	87.0	91.5
UDATP [25]	UDA	None	49.0	68.4	80.1	83.5	53.7	75.8	89.5	93.2
UCDA-CCE [50]	UDA	Camera ID	31.0	47.7	-	-	30.9	60.4	-	-
PDA-Net [51]	UDA	None	45.1	63.2	77.0	82.5	47.6	75.2	86.3	90.2
PCB-PAST [8]	UDA	None	54.3	72.4	-	-	54.6	78.4	-	-
SSG [7]	UDA	None	53.4	73.0	80.6	83.2	58.3	80.0	90.0	92.4
MMCL [24]	UDA	None	51.4	72.4	82.9	85.0	60.4	84.4	92.8	95.0
DG-NET++ [52]	UDA	None	**63.8**	**78.9**	87.8	90.4	61.7	82.1	90.2	92.7
B-SNR+GDS-H [53]	UDA	None	55.1	73.1	-	-	61.2	81.1	-	-
OG-Net [54]	UDA	None	16.3	31.3	-	-	17.2	41.4	-	-
AE [17]	UDA	None	46.7	67.9	79.2	83.6	58.0	81.6	91.9	94.6
DGNet [3]	UDA	None	24.3	42.6	58.6	64.6	26.8	56.1	72.2	78.1
AD-Cluster [55]	UDA	None	54.1	72.6	82.5	85.5	68.3	86.7	94.4	96.5
TJ-AIDL [26]	Weakly	Attribute	23.0	44.3	59.6	65.0	26.5	58.2	74.8	81.1
Baseline	Unsupervised	None	48.6	65.9	79.2	83.8	51.8	73.7	87.6	91.6
A^2G	Weakly	Attribute	51.0	69.0	82.3	86.4	63.9	83.0	92.9	95.4
Baseline	UDA	None	54.9	70.8	82.5	86.3	65.9	84.5	94.2	95.9
A^2G (w/o SPL)	Weakly	Attribute	58.7	75.9	86.0	89.7	67.2	84.9	94.1	96.6
A^2G	Weakly	Attribute	61.2	77.1	**88.2**	**91.2**	**71.6**	**87.4**	**95.2**	**97.2**

Methods	Setting	Auxiliary INFO	MSMT17-to-Duke				MSMT17-to-Market			
			mAP	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10
D-MMD [49]	Unsupervised	None	51.6	68.8	82.6	87.3	50.8	72.8	88.1	92.3
MAR	UDA	None	48.0	67.1	79.8	-	40.0	67.7	81.9	-
PAUL [56]	UDA	None	53.2	72.0	82.7	86.0	40.1	68.5	82.4	87.4
CASCL [57]	UDA	Camera ID	37.8	59.3	73.2	77.8	35.5	65.4	86.2	35.5
DG-NET++ [52]	UDA	None	58.2	75.2	73.6	86.9	64.6	83.1	91.5	94.3
OG-Net [54]	UDA	None	25.9	44.9	-	-	21.4	47.6	-	-
Baseline	Unsupervised	None	51.6	68.5	81.2	86.2	40.5	63.6	79.1	85.2
A^2G	Weakly	Attribute	54.8	71.8	84.2	88.2	61.0	80.4	92.6	95.4
Baseline	UDA	None	54.2	70.5	82.7	86.9	67.6	85.1	94.6	96.5
A^2G (w/o SPL)	Weakly	Attribute	58.9	74.6	85.3	88.6	71.3	87.0	**95.4**	**97.1**
A^2G	Weakly	Attribute	**61.1**	**77.7**	**87.7**	**90.8**	**71.3**	**87.6**	95.2	97.0

label smoothing. We also verify the effectiveness of identity classification loss by removing the $\mathcal{L}_{ce}^{spl}$ in Eq. (10)), and the experiments are denoted as "A^2G (w/o $\mathcal{L}_{ce}^{spl}$)". We observe significant decreases of 30.9% and 32.9% mAP on these two backbones, which validates the effectiveness of learning with identity classification loss $\mathcal{L}_{ce}^{spl}$. By applying cross-entropy loss, we simultaneously minimize the distance of intra-class samples and maximize the inter-class and generate more representative feature embeddings. Furthermore, we evaluate the effectiveness of triplet loss by removing it in train process, *i.e.*, $\lambda = 0$ in Eq. (10), and the experiments are presented as "A^2G (w/o $\mathcal{L}_{tri}$)". Considerable decreases of 2.9% and 0.8% mAP are shown on ResNet-50 and IBN-ResNet-50 backbone for Duke-to-Market. The introduce of triplet loss captures the relative similarity in training sample and enhances the representation ability of embedding network. The increase of experimental results demonstrate the effectiveness of triplet loss.

Impact of Aggregation Function in GNN. The aggregation function is the key element of graph neural network based attribute-auxiliary feature aggregation. To study the impact of different aggregation function, we present the performance of three commonly utilized functions proposed in [12]: "mean", "pool", and "GCN [31]". Briefly speaking, "mean" and "pool" conduct "average" and "max" operation before multiplying the weight matrix in graph neural network, respectively. As to "GCN", it directly multiplies the weight matrix. More design details are presented in [12].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3069915, IEEE Access



(a) The smoothing parameter α. (b) The balance weight λ. (c) The number of layers K in GNN.

FIGURE 3: Retrieval accuracy curves with model parameters on Duke-to-Market. Parameter α is the smoothing penalty of smoothed pseudo label (SPL) in Equation (9) for $\mathcal{L}_{ce}^{spl}$. Parameter λ is the balance weight in Equation (10) for $\mathcal{L}_{all}$. GNN is the model for attribute-auxiliary feature aggregation, and the number of layers in GNN controls the model complexity.

TABLE 2: Component-wise evaluation of A^{2}G with ResNet-50 and IBN-ResNet-50 backbones on Duke-to-Market. "Direct Transfer" denotes the source pre-trained model. "Baseline" are the clustering-based self-training model with DBSCAN [42]. "SPL" denotes model leaning with smoothed pseudo labels. "$\mathcal{L}_{tri}$ (Eq. (2))" and $\mathcal{L}_{ce}^{spl}$ (Eq. (9)) are the two terms in $\mathcal{L}_{all}$ (Eq. (10)).

Methods	ResNet-50		IBN-ResNet-50	
	mAP(%)	Rank@1(%)	mAP(%)	Rank@1(%)
Direct Transfer	31.8	61.9	35.6	65.3
Baseline	65.9	85.4	74.7	89.9
A^{2}G (w/o $\mathcal{L}_{tri}$)	68.7	85.2	76.0	90.4
A^{2}G (w/o $\mathcal{L}_{ce}^{spl}$)	40.7	63.2	43.9	67.2
A^{2}G (w/o SPL)	67.2	87.4	73.2	89.2
A^{2}G	71.6	87.4	76.8	90.6

TABLE 3: The comparison of different aggregation function $f_{aggregate}$ (Eq. (5)) in GNN on Duke-to-Market.

Aggregation	mAP	Rank@1	Rank@5	Rank@10
pool	67.4	84.5	93.8	96.1
mean	69.6	87.3	94.9	97.0
GCN	**71.6**	**87.5**	**95.9**	**97.3**

As shown in Table 3, we obtain the best result with mAP $= 71.6\%$ when using GCN as aggregation function. With "mean" and "pool" aggregation function, we obtain slightly inferior mAP accuracies of 69.6% and 67.4%, respectively. From the above observations, we may conclude that aggregation function with addition sampling operation, such as "mean" and "pool", corrupts the representation of individual image and the structure of original cluster. As to GCN, which consists of graph convolution operator without sampling, we could preserve the independence of cluster and explore the similarity in attribute space.

Impact of the Threshold τ. The parameter τ controls whether the image pairs are connected in the pedestrian attribute graph or not. As shown in Table 4, the retrieval accuracy reaches a plateau when τ is larger than 0.85 and less than 1.0, which indicates that the parameter τ is not sensitive in this range. From the above experiments, we may conclude that the value of τ in the range ($0.85 \leq \tau \leq 1.0$) would be general to other datasets.

TABLE 4: Retrieval accuracy with different values of the threshold τ (Eq. (3)) in pedestrian attribute graph construction. We also present the number of edges under different thresholds.

Threshold	# of edges	mAP(%)	Rank@1(%)
0.8	6,792,452	69.8	85.7
0.85	3,054,744	71.2	87.4
0.9	1,605,246	71.3	87.3
0.95	602,100	**71.6**	**87.5**

(a) Duke-to-Market.

Threshold	# of edges	mAP(%)	Rank@1(%)
0.8	9,053,746	59.8	75.2
0.85	7,084,622	60.8	76.6
0.9	3,670,594	**61.2**	**77.1**
0.95	3,127,828	60.9	76.8

(b) Market-to-Duke.

E. PARAMETER ANALYSIS

Effect of the Label Smoothing Parameter α. Parameter α controls the confidence of pseudo labels in discriminative learning with cross-entropy loss. We vary α from 0 to 0.4 and present the mAP and Rank@1 accuracies on Duke-to-Market in Figure 3(a). As α increases from 0 to 0.1, the improvement is increasingly significant, which validates the necessity of smoothing the label to avoid the over-confidence of *hard* pseudo labels. If we set α to be larger than 0.1, the over-smoothing of pseudo label leads to ineffective discriminative learning for embedding network.

Effect of Weight Balance Parameter λ. Parameter λ is the balance weight of $\mathcal{L}_{tri}$ and effects the hardest triplet relative learning for embedding features. We vary λ from 0 to 2.0 and present retrieval accuracies in Figure 3(b). As λ increases for 0 to 1.0, the retrieval accuracies increase steadily, which When λ goes large, $\mathcal{L}_{tri}$ dominates $\mathcal{L}_{all}$, so that the discriminative learning with cross-entropy loss is weakened and cannot provide enough supervision for feature representation learning.

Effect of the Number of Layers K. Parameter K is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3069915, IEEE Access

number of layers in graph neural network and effects the model capacity in learning attribute-auxiliary feature embedding. We vary K from 1 to 4 and show the mAP and Rank@1 accuracies in Figure 3(c). The performances of model are improved when K increases from 1 to 2. The increase of K improves the model capacity of graph neural network based attribute auxiliary feature aggregation, which boosts the retrieval accuracy. When K is larger than 2, the retrieval accuracy drops significantly because of model overfitting.

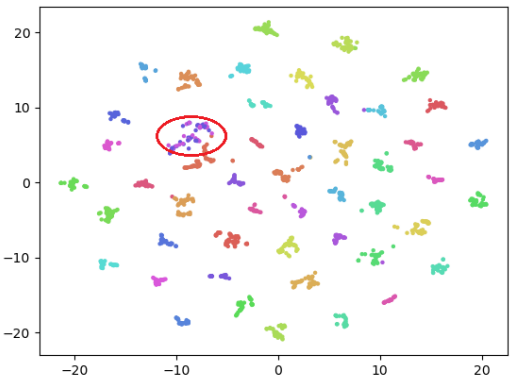
F. QUALITATIVE ANALYSIS

To further demonstrate the improvements of A^{2}G in feature representation, we visualize the feature embeddings with and without attribute-auxiliary feature aggregation. As shown in Figure 4, the embeddings of the same identity with attribute-auxiliary feature aggregation gather tighter than the original features, which demonstrates that A^{2}G increases the similarity of the intra-identity. We also observe that A^{2}G removes the false positive samples and increases the distance of inter-identity in the red circle of Figure 4(a). From the above observations, we conclude that A^{2}G could improve the feature representation of images and the quality of the pseudo labels.

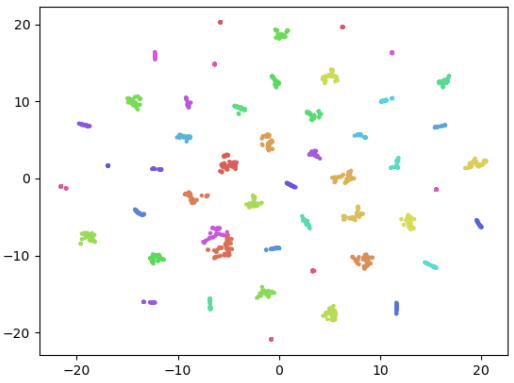
Furthermore, we present the retrieval results of direct transfer model, baseline model, and our proposed A^{2}G, as shown in Figure 5. We observe that A^{2}G is robust to the variance of illumination (first retrieval sample), the existence of occlusion (second retrieval sample), and the distraction of visual similar pedestrian (third retrieval sample). The above results validate that our proposed A^{2}G could improve the feature representation of images through generating high-quality pseudo label with the consideration of similarities in attribute space and feature space.

V. CONCLUSIONS

In this paper, we study the attribute auxiliary weakly supervised person re-ID and focus on improving the quality of pseudo labels and reducing the over-confidence of the pseudo label in discriminative learning for embedding network. Clustering-based UDA approaches in person re-ID highly rely on the quality of pseudo labels. Due to the existence of domain variances, such as illuminations, cameras, and viewpoints, assigning pseudo labels according to the cluster in the feature space may generate more false-positive samples during training. Besides, learning with these inaccurate *hard* labels may damage the discriminative learning for embedding networks. To address the above problems, we propose a graph neural network based attribute auxiliary structured grouping to explore the similarity in attribute space. Different from existing clustering-based approaches that only utilize the similarity in the feature space, we also consider pedestrian attributes. A graph neural network based attribute auxiliary feature aggregation is presented to refine the embedding features with the similarity in attribute space. Besides, we regularize the learning of the embeddig model with smoothed pseudo labels to avoid the "over-confidence"



(a) Feature embeddings without attribute-auxiliary feature aggregation.



(b) Feature embeddings with attribute-auxiliary feature aggregation.

FIGURE 4: T-SNE visualization of the features embeddings with and without attribute-auxiliary feature aggregation on a part of Market-1501 training set (50 identities). Points with the same color represent the images of the same identities.

in discriminative learning. By comparing various state-of-the-art algorithms, the encouraging results demonstrate that the proposed A^{2}G is effective and promising for person re-identification.

REFERENCES

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193, 2020.

[2] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu. A strong baseline and batch normalization neck for deep person re-identification. IEEE Transactions on Multimedia, pages 1–1, 2019.

[3] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[4] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). pages 480–496, 2018.

[5] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned

FIGURE 5: The pedestrian retrieval results (3 query images) with directly transfer model, baseline model, and our proposed A^2G (from top to bottom) on Market1501 test set. The green ✓ denotes the correct match image, the red ✗ denotes the mismatch one.

cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(1):1–20, 2017.

[6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In International Conference on Learning Representations, 2020.

[7] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 6112–6121, 2019.

[8] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 8222–8231, 2019.

[9] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. Pattern Recognition, 95:151–161, 2019.

[10] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened similarity learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3390–3399, 2020.

[11] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7134–7143, 2019.

[12] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Advances in neural information processing systems, pages 1024–1034, 2017.

[13] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In Advances in Neural Information Processing Systems, pages 4694–4703, 2019.

[14] Mang Ye, Jianbing Shen, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Augmentation invariant and instance spreading feature for softmax embedding. IEEE transactions on pattern analysis and machine intelligence, 2020.

[15] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 598–607, 2019.

[16] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 8738–8745, 2019.

[17] Yuhang Ding, Hehe Fan, Mingliang Xu, and Yi Yang. Adaptive exploration for unsupervised person re-identification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(1):1–19, 2020.

[18] Mang Ye, Jiawei Li, Andy J Ma, Liang Zheng, and Pong C Yuen. Dynamic graph co-matching for unsupervised video-based person re-identification. IEEE Transactions on Image Processing, 28(6):2976–2990, 2019.

[19] Mang Ye, Xiangyuan Lan, and Pong C Yuen. Robust anchor embedding for unsupervised video person re-identification in the wild. In Proceedings of the European Conference on Computer Vision (ECCV), pages 170–186, 2018.

[20] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 994–1003, 2018.

[21] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[22] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In Proceedings of the European Conference on Computer Vision (ECCV), pages 172–188, 2018.

[23] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jianhuang Lai. Unsupervised person re-identification by soft multilabel learning. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[24] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10981–10990, 2020.

[25] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. Pattern Recognition, 102:107173, 2020.

[26] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2275–2284, 2018.

[27] Jingke Meng, Sheng Wu, and Wei-Shi Zheng. Weakly supervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 760–769, 2019.

[28] Mang Ye and Pong C Yuen. Purifynet: A robust person re-identification model with noisy labels. IEEE Transactions on Information Forensics and Security, 15:2655–2666, 2020.

[29] Shuzhao Li, Huimin Yu, and Roland Hu. Attributes-aided part detection and refinement for person re-identification. Pattern Recognition, 97:107016, 2020.

[30] Hehe Fan and Yi Yang. Person tube retrieval via language description. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 10754–10761, 2020.

[31] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR), 2017.

[32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. International Conference on Learning Representations, 2018. accepted as poster.

[33] Jie Chen, Tengfei Ma, and Cao Xiao. Fastgcn: Fast learning with graph convolutional networks via importance sampling. In International Conference on Learning Representations, 2018.

[34] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4649–4659, 2019.

[35] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5177–5186, 2019.

[36] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[37] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In European Conference on Computer Vision (ECCV), 2020.

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[39] Jan Chorowski and Navdeep Jaitly. Towards better decoding and language model integration in sequence to sequence models. arXiv preprint arXiv:1612.02695, 2016.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[41] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1318–1327, 2017.

[42] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, volume 96, pages 226–231, 1996.

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision, pages 1116–1124, 2015.

[44] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European Conference on Computer Vision, pages 17–35. Springer, 2016.

[45] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(4):1–18, 2018.

[46] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13657–13665, 2020.

[47] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In Proceedings of the European conference on computer vision (ECCV), pages 737–753, 2018.

[48] Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M Hospedales. Disjoint label space transfer learning with common factorised space. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3288–3295, 2019.

[49] Djebril Mekhazni, Amran Bhuiyan, George Ekladious, and Eric Granger. Unsupervised domain adaptation in the dissimilarity space for person re-identification. arXiv preprint arXiv:2007.13890, 2020.

[50] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 8080–8089, 2019.

[51] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In Proceedings of the IEEE International Conference on Computer Vision, pages 7919–7929, 2019.

[52] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. arXiv preprint arXiv:2007.10315, 2020.

[53] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Global distance-distributions separation for unsupervised person re-identification. In European Conference on Computer Vision, pages 735–751. Springer, 2020.

[54] Yi Yang Zhedong Zheng. Parameter-efficient person re-identification in the 3d space. arXiv 2006.04569, 2020.

[55] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9021–9030, 2020.

[56] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3633–3642, 2019.

[57] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 6922–6931, 2019.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[59] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In Proceedings of the European Conference on Computer Vision (ECCV), pages 464–479, 2018.

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015.

GEYU TANG received his B.E. degree in Electronic and Information Engineering from Xidian University, Shaanxi, China in 2016. He is a Ph.D candidate in Electronic and Information Engineering at the Institute of Microelectronics of the Chinese Academy of Sciences, Beijing, China. His interested fields include: image retrieval, person re-identification, graph represent learning.

XINGYU GAO (M'18) received the Ph.D. degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is currently a Full Professor with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing. He was with Nanyang Technological University, Singapore, and Singapore Management University, Singapore, as a Visiting Scholar. His current research interests include machine learning, multimedia information retrieval, and ubiquitous computing.

ZHENYU CHEN (M'18) received the Ph.D. degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is currently a Senior Engineer with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing. He was with the Computer Science Department, Dartmouth College, Hanover, NH, USA, as a Visiting Scholar. His current research interests include machine learning, pervasive computing, and mobile sensing.

HUICAI ZHONG received his PhD in microelectronics from the Department of Electronic Engineering, North Carolina State University. He has been engaged in research and product development of high performance chips such as CPU and Flash in AMD, IBM, SanDisk and other companies. Currently, he is a professor with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China.