Graph Consistency Based Mean-Teaching for Unsupervised Domain Adaptive Person Re-Identification

Xiaobin Liu, Shiliang Zhang
Department of Computer Science, School of EECS, Peking University
{xbliu.vmc, slzhang.jdl}@pku.edu.cn

Abstract

Recent works show that mean-teaching is an effective framework for unsupervised domain adaptive person re-identification. However, existing methods perform contrastive learning on selected samples between teacher and student networks, which is sensitive to noises in pseudo labels and neglects the relationship among most samples. Moreover, these methods are not effective in cooperation of different teacher networks. To handle these issues, this paper proposes a Graph Consistency based Mean-Teaching (GCMT) method with constructing the Graph Consistency Constraint (GCC) between teacher and student networks. Specifically, given unlabeled training images, we apply teacher networks to extract corresponding features and further construct a teacher graph for each teacher network to describe the similarity relationships among training images. To boost the representation learning, different teacher graphs are fused to provide the supervise signal for optimizing student networks. GCMT fuses similarity relationships predicted by different teacher networks as supervision and effectively optimizes student networks with more sample relationships involved. Experiments on three datasets, *i.e.*, *Market-1501*, *DukeMTMC-reID*, and *MSMT17*, show that proposed GCMT outperforms state-of-the-art methods by clear margin. Specially, GCMT even outperforms the previous method that uses a deeper backbone. Experimental results also show that GCMT can effectively boost the performance with multiple teacher and student networks. Our code is available at https://github.com/liu-xb/GCMT .

1 Introduction

Person Re-Identification (ReID) aims to match a query person image in a gallery set [Ge *et al.*, 2020a; Zhai *et al.*, 2020]. Supervised person ReID has been widely studied from different aspects. However, supervised models suffer from the expensive data annotation and substantial performance drop when applied on different target domains. To address these issues, recent works focus on unsupervised domain adaptive person ReID and exhibit promising performance. More details of recent works are summarized in Sec. 2.

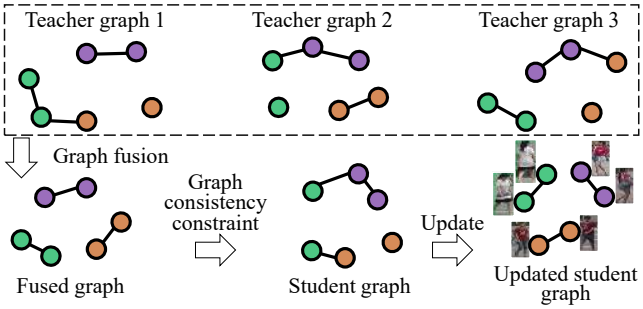

Figure 1: Illustration of Graph Consistency Constraint (GCC) in proposed GCMT method between three teacher graphs and one student graph. Relationships of samples in the student graph are updated to approach relationships in the graph fused by multiple teacher graphs. Dots in different colors denote different identities.

Despite the significant success, there still remain several open issues unexplored. Firstly, previous works commonly perform training on both labeled source data and unlabeled target data [Liu and Zhang, 2020]. In real-world applications, however, it is always impractical to access the labeled source data due to privacy issue. Moreover, labeled datasets are also more expensive to store and process than pre-trained models. This hinders the training of these methods. Secondly, person ReID model optimized by mean-teaching framework requires an effective student network training strategy that should be robust to noisy labels and also effective in feature learning. Existing methods use soft triplet loss on selected samples based on pseudo labels [Zhai *et al.*, 2020]. However, it is sensitive to noisy pseudo labels and ineffective in feature learning with few selected sample relationships involved. Thirdly, models pre-trained on different labeled datasets are public available and using them in training has potential to boost the performance on target datasets. However, current methods with soft triplet loss are not effective in collaborative learning with multiple models.

This paper is motivated to study an effective training method for unsupervised domain adaptive person ReID. To release the dependence on labeled source data, only unlabeled target data is used in unsupervised training. To enhance the effectiveness in feature learning of student networks and co-

operation of teacher networks, we propose a Graph Consistency based Mean-Teaching (GCMT) method with a Graph Consistency Constraint (GCC) between teacher and student networks as shown in Fig. 1. GCMT merges sample relationships predicted by different teacher networks as supervision and optimizes more relationships among different samples in student networks, thus provides effective supervision and optimization for feature learning on student networks.

As illustrated in Fig. 1, the proposed Graph Consistency Constraint (GCC) in GCMT method is performed between teacher graphs and student graphs. Specifically, given public pre-trained models, GCMT initializes a student network and its temporal mean teacher network from each model. Features extracted by a teacher network are formed into a teacher graph with using features as nodes. Connections between nodes in teacher graphs are determined by KNN strategy, *i.e.*, a node is only connected with its KNN nodes. Weights of connections are computed as feature similarities. Features extracted by a student network are formed into a student graph with weights of connections also computed as feature similarities. Teacher graphs are fused to guide the connection weight leaning in student graphs by GCC via encouraging the connection weights in student networks to approach the weights in the fused graph. Thus, GCC guides student graphs to describe the same sample relationship with the fused teacher graph. Compared with previous methods, proposed GCC avoids the dependency on pseudo labels. It involves relationships among more samples in feature learning and also can effectively fuse sample relationships predicted by multiple teacher networks, hence is more effective in feature learning and also robust against noises in pseudo labels. To the best of our knowledge, it is the first work to use GCC in mean-teaching framework for unsupervised training.

We test our method on three large-scale datasets, *i.e.*, *Market-1501*, *DukeMTMC-reID* , and *MSMT17*. Comparison with recent works shows that our method outperforms state-of-the-art methods by a clear margin. For instance, training from model pre-trained on *ImageNet*, GCMT achieves mAP accuracy of 63.6% on *DukeMTMC-reID*, outperforming recent UTAL [Li *et al.*, 2019a] by 19%. By using model pre-trained on *DukeMTMC-reID*, our GCMT method achieves mAP accuracy of 77.1% on *Market-1501*, significantly outperforming recent MEB-Net [Zhai *et al.*, 2020] and MMT [Ge *et al.*, 2020a] by 4.9% and 5.9%, respectively. Experiments also show that GCMT boosts the performance with multiple pre-trained models, *e.g.*, the mAP accuracy is boosted to 79.7% on *Market-1501* with models pre-trained on *MSMT17*, *CUHK03*, and *DukeMTMC-reID*.

2 Related Work

Domain-invariant feature learning. Some works design GAN based models to transfer labeled images to target domains [Liu *et al.*, 2019]. Several works map labeled and unlabeled images to a shared feature space to bridge the domain gap [Liu and Zhang, 2020]. These methods require labeled source data for training, which is hardly available due to privacy issue in real-world applications. Compared with them, this work only uses pre-trained models instead of labeled data, thus is more suitable for real-world applications.

Self-supervised learning. Some works locally predict pseudo labels for unlabeled samples [Zhong *et al.*, 2019; Yu *et al.*, 2019]. To acquire reliable pseudo labels, some researchers try to refine unsupervised clustering [Ding *et al.*, 2019; Lin *et al.*, 2019]. For unsupervised optimization, current works use pseudo labels as supervision and adopt triplet loss [Fu *et al.*, 2019; Zhang *et al.*, 2019] or contrastive loss [Liu and Zhang, 2020; Zhong *et al.*, 2019] for training. However, these methods fail to consider the noise in pseudo labels, which substantially hinders the model training. Compared with them, proposed model adopts fused graph of temporal mean teacher networks to supervise feature learning, alleviating the effect of noisy labels and improves the performance. Ge *et al.* [Ge *et al.*, 2020b] propose a self-paced learning method to boost the ReID performance. However, they have different motivations and propose different algorithms compared with this paper. They aim to make full use of labeled data by caching features in a hybrid memory bank. While, this paper aims to study an effective mean-teaching method on unlabeled datasets without labeled data by proposing a graph consistency constraint. Experiments show our method achieves better performance under the same setting.

Teacher-student training. Teacher-student framework is widely studied in semi-supervised learning [Tarvainen and Valpola, 2017; Han *et al.*, 2018].However, these methods hold the assumption that labeled and unlabeled images are of same categories, making them not suitable for person ReID task. Recently, some works adopt mean-teaching method in unsupervised domain adaptive person ReID. MMT [Ge *et al.*, 2020a] uses two pairs of teacher and student networks for pseudo label refinery. MEB-Net [Zhai *et al.*, 2020] further uses three pairs of teacher and student networks of three different structures for training. MMT and MEB-Net use triplet loss in model training and soft triplet loss between teacher and student networks, which are sensitive to noisy label and ineffective in feature learning. Moreover, training with multiple teacher networks is still ineffective and the improvements by using three teacher networks is limited in [Zhai *et al.*, 2020]. Compared with them, proposed GCMT is effective in both student network training and teacher network cooperation. Experimental results show that GCMT achieves better performance compared with these methods.

3 Problem Formulation

Given an unlabeled dataset $\mathcal{D}$, unsupervised domain adaptive person ReID aims to learn the ReID model on $\mathcal{D}$ without annotation. $\mathcal{D}$ can be denoted as $\{x_i|i = 1...N\}$. x_i and N denote the i-th image and the number of images in $\mathcal{D}$, respectively. In this paper, training starts from public models pre-trained on different datasets, while labeled source data is not used. This setting is more practical in real-world scenario compared with several previous works as discussed in Sec. 1. Public pre-trained models are denoted as $\{M^1, M^2, ..., M^m\}$ where m denotes the number of models.

We adopt mean-teaching framework for unsupervised training on $\mathcal{D}$ as shown in Fig. 2. m student networks are initialized by m public pre-trained model, respectively. The pa-

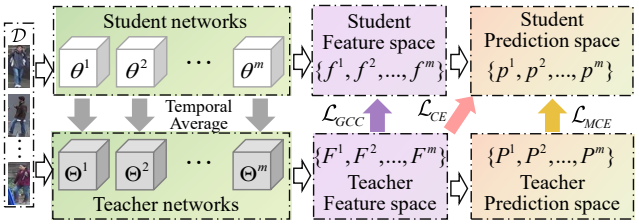

Figure 2: Framework of proposed GCMT method. m pairs of student and teacher networks are initialized by m public pre-trained models. Teacher network Θ^j is updated by temporal average strategy with corresponding student network θ^j. Features extracted by student networks are supervised by features extracted by teacher networks via graph consistency constraint $\mathcal{L}_{GCC}$. Class predictions by student networks are supervised by pseudo labels via cross entropy loss $\mathcal{L}_{CE}$ and predictions by teacher networks via mutual cross entropy loss $\mathcal{L}_{MCE}$. Pseudo labels are generated by off-line clustering on features extracted by teacher networks.

rameters of student networks are denoted as $\{\theta^1, \theta^2, ..., \theta^m\}$ with θ^j denoting the parameters of the j-th student network. We use temporal average network as the teacher network for each student network as in [Tarvainen and Valpola, 2017]. Parameters in teacher networks are denoted as $\{\Theta^1, \Theta^2, ..., \Theta^m\}$, respectively. In the beginning of training, θ^j and Θ^j are both initialized by M^j. In each iteration of the training, θ^j is updated by objective functions and Θ^j is then updated by temporal average strategy with updated θ^j as:

$$\Theta^j \leftarrow 0.999\Theta^j + 0.001\theta^j. \tag{1}$$

Coefficients in Eqn. (1) are empirical values following [Ge *et al.*, 2020a; Zhai *et al.*, 2020].

For unlabeled images x_i, the extracted feature after L2-normalization and class prediction after softmax normalization by the j-th student network θ^j are denoted as f_i^j and p_i^j, respectively. And the extracted feature and prediction by the j-th teacher network Θ^j for x_i are denoted as F_i^j and P_i^j, respectively. Note that we use the superscript and subscript to denote the index of networks and images, respectively.

The target of training on $\mathcal{D}$ is to make extracted features discriminative for person ReID task. In this paper, we propose a Graph Consistency based Mean Teaching (GCMT) method to achieve this goal via constraints in two spaces: 1) class prediction space, and 2) feature space. The framework of proposed method is illustrated in Fig. 2.

In class prediction space, we first generate pseudo ID labels on $\mathcal{D}$ by unsupervised clustering on averaged features extracted by teacher networks before each epoch. Based on pseudo labels, cross entropy loss $\mathcal{L}_{CE}$ is computed for pseudo ID classification on each student network. We use a new fully connected layer for label prediction after each clustering step following [Ge *et al.*, 2020a; Zhai *et al.*, 2020]. After clustering features into C clusters, a mean feature for each cluster is computed by averaging all features in this cluster. This results in C mean features corresponding to C clusters. Then, a new fully connected layer with C-way outputs is used for label prediction and parameters are initialized with corresponding mean features, *i.e.*, the parameters for the c-th class is initialized by the mean feature of the c-th cluster.

To eliminate the negative effects of noises in pseudo label inside each training batch, mutual cross entropy loss $\mathcal{L}_{MCE}$ is used to encourage student networks to predict same class probabilities with teacher networks for training samples. Specifically, the probabilities of x_i by different teacher networks are averaged as the training target for each student network, which is denoted as $\hat{P}_i = \sum_j^m P_i^j$. The objective function of $\mathcal{L}_{MCE}$ with m teacher networks is formulated as:

$$\mathcal{L}_{MCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\hat{P}_i(c)\sum_{j=1}^{m}\log p_i^j(c), \tag{2}$$

where C denotes the number of identities generated by clustering. $\hat{P}_i(c)$ and $p_i^j(c)$ denote the probability of x_i being classified to c-th class in $\hat{P}_i$ and p_i^j, respectively. The objective function in label prediction space $\mathcal{L}_{LP}$ is the combination of $\mathcal{L}_{CE}$ and $\mathcal{L}_{MCE}$ as:

$$\mathcal{L}_{LP} = \mathcal{L}_{CE} + \mathcal{L}_{MCE}. \tag{3}$$

Classification probabilities describe the relationship between samples and classes. Thus, $\mathcal{L}_{LP}$ focuses on classifying training samples to corresponding pseudo labels. For ReID task, however, we also need to optimize the model in feature space, as identities in testing set are different with training set and ReID is performed as retrieval by extracted features. Thus, we propose Graph Consistency Constraint (GCC) between teacher and student networks in feature space.

As shown in Fig. 2, GCC is proposed to train student networks in feature space. GCC computes sample similarity relationship for each network with corresponding extracted features, respectively. Sample similarity relationships predicted by different teacher networks are fused as the target relationship. Student networks are encouraged by GCC to output features that have the same sample similarity relationship as the target relationship as shown in Fig. 1. The objective function of GCC is denoted as $\mathcal{L}_{GCC}$ and details of the computation on $\mathcal{L}_{GCC}$ will be given in Sec. 4.

$\mathcal{L}_{LP}$ is performed in class prediction space and focuses on relationship between samples and pseudo IDs. While $\mathcal{L}_{GCC}$ is performed in feature space and focuses on similarity relationship among samples. Therefore, $\mathcal{L}_{LP}$ and $\mathcal{L}_{GCC}$ are complementary to each other. We hence denote the final objective function as the combination of $\mathcal{L}_{LP}$ and $\mathcal{L}_{GCC}$:

$$\mathcal{L} = \mathcal{L}_{LP} + \lambda_{GCC}\mathcal{L}_{GCC}, \tag{4}$$

where λ_{GCC} is the loss weight for $\mathcal{L}_{GCC}$.

4 Graph Consistency Constraint

In feature learning, we aim to fuse sample similarity relationships computed by different teacher networks to provide training supervision for student networks. Features extracted by different teacher networks are of different distributions. Thus, directly averaging similarities computed by different teacher networks is not reasonable for relationship fusion. Moreover, when only relationship between several selected samples are involved, *e.g.*, soft triplet loss with hard sample mining in [Zhai *et al.*, 2020], noise in pseudo labels will mislead both the sample selection and optimization.

To chase a reasonable relationship fusion and an effective student network optimization, we propose the Graph Consistency Constraint (GCC) that describes sample similarity relationship as weights of edges in graphs. Then, GCC performs relationship fusion and network optimization based on graphs. By fusing relationship predicted by different teacher networks and involving more samples in training student networks, GCC eliminates the effect of noises in pseudo labels and provides effective update for student networks.

4.1 Teacher Graph Construction

Similarity relationship predicted by each teacher network is described as a graph. Specifically, a K-nearest neighbour graph $G(F^j, W^j)$ is constructed for features extracted by the j-th teacher network Θ^j. $F^j = \{F_i^j | i = 1...N\}$ and $W^j = \{W_{i,k}^j | i, k = 1...N\}$ denote the set of features of nodes and the set of weights of edges between nodes, respectively. Weights of edges describe the distance relationships among different features in the graph, which are computed based on K-NN relationship. Specifically, F_k^j is connected to F_i^j if F_k^j is among the K-nearest neighbours of F_i^j. The weight of edge from F_i^j to F_k^j, *i.e.*, $W_{i,k}^j$, is computed as $W_{i,k}^j = (F_i^j)^T F_k^j$. For pairs of nodes that are not connected, corresponding weights of edges are assigned as 0.

Graphs constructed by different teacher networks are fused to provide training target for student networks training. Weights in each graph are first normalized by softmax and then averaged to obtain the fused weights as follows:

$$\hat{W}_{i,k} = \frac{1}{m} \sum_{j=1}^{m} \frac{\exp(W_{i,k}^j)}{\sum_{h=1}^{N, h \neq i} \exp(W_{i,h}^j)}. \tag{5}$$

The set of fused weights $\hat{W}_{i,k}$ is denoted as $\hat{W}$.

4.2 Graph Consistency Constraint Computation

Student networks are supervised by GCC to extract features that have the same similarity relationships between different samples described by $\hat{W}$. The computation of GCC should satisfy three conditions. 1): GCC should simultaneously update every relationship between different training samples for student networks, making optimization effective and precise. 2): GCC should be aware of the density of neighbours. Some samples have dense neighbours whose K-nearest neighbours should reach rather high similarities with them. While some other samples have sparse neighbours, and restriction on their neighbours can be relaxed. Thus, GCC should optimize relative similarity between samples, instead of absolute similarity. 3): GCC should be aware of the hardness of K-nearest neighbours of each sample, *i.e.*, gradients of far neighbours should be larger than gradients of near neighbours.

For computing GCC, student graph $G(f^j, w^j)$ is constructed for features extracted by the j-th student network θ^j, where $f^j = \{f_i^j | i = 1...N\}$ and $w^j = \{w_{i,k}^j | i, k = 1...N\}$ denote the set of features of nodes and the set of weights of edges between nodes, respectively. Weights of edges from f_i^j to f_k^j in $G(f^j, w^j)$ is computed as:

$$w_{i,k}^j = \frac{\exp((f_i^j)^T f_k^j / \beta)}{\sum_{h=1}^{N, h \neq i} \exp((f_i^j)^T f_h^j / \beta)}, \tag{6}$$

where β is a hyper-parameter to adjust the scale of distribution. GCC is computed to encourage $\{w^j | j = 1...m\}$ to approach $\hat{W}$ via cross entropy which is formulated as:

$$\mathcal{L}_{GCC} = -\frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{N, k \neq i} \hat{W}_{i,k} \sum_{j=1}^{m} \log w_{i,k}^j. \tag{7}$$

Proposed objective function in Eqn. (7) satisfies aforementioned three conditions. 1): Every edge in student graphs is updated by $\mathcal{L}_{GCC}$ based on $\hat{W}$. 2): Due to the softmax normalization in Eqn. (6), $\mathcal{L}_{GCC}$ normalizes loss of samples with different densities of neighbours to the same scale. Thus, $\mathcal{L}_{GCC}$ is aware of the density of neighbours. 3): The gradient of $\mathcal{L}_{GCC}$ relative to $w_{i,k}^j$ when $\hat{W}_{i,k} \neq 0$ is computed as: $\frac{\partial \mathcal{L}_{GCC}}{\partial w_{i,k}^j} = -\frac{\hat{W}_{i,k}}{N w_{i,k}^j}$. The scale of gradient increases with $w_{i,k}^j$ decreasing. This indicates that $\mathcal{L}_{GCC}$ pays more attention on K-nearest neighbours with small similarities for each sample. Thus, $\mathcal{L}_{GCC}$ is also aware of hardness.

4.3 Discussion

GCMT fuses sample similarity relationship computed by different teacher networks as supervise signal instead of pseudo labels. As networks are being more discriminative with training, similarity relationship computed in each iteration has potential to be more precise than pseudo labels computed offline before each epoch. Moreover, $\mathcal{L}_{GCC}$ avoids sample mining algorithm and automatically focuses on hard neighbours.

Compared with soft triplet loss used in [Zhai *et al.*, 2020] and [Ge *et al.*, 2020a] that trains student networks based on pseudo labels, $\mathcal{L}_{GCC}$ optimizes distance relationship in student graphs. This considers more relationship among samples and avoids the negative effect of noises in pseudo labels. Specially, soft triplet loss can be regarded as a special case of GCC that uses sub-graphs with selected samples. Advantages of our method will be shown in Sec. 5.3.

Compared with MEB-Net [Zhai *et al.*, 2020] that fuses relationship predicted by different teacher networks via soft triplet loss based on pseudo labels, GCC performs relationship fusion based on graphs. This avoids effect of noises in pseudo label and involves more relationship in training. Advantages of our method in multiple teacher networks cooperation will be shown in Sec. 5.4.

5 Experiment

5.1 Dataset

Experiments are performed on three datasets, *i.e.*, *DukeMTMC-reID* [Zheng *et al.*, 2017], *Market-1501* [Zheng *et al.*, 2015], and *MSMT17* [Wei *et al.*, 2018].

DukeMTMC-reID contains 36,411 images of 1,812 identities. 16,522 images of 702 identities are used for training. In the rest of images, 3,368 images are selected as query images and remaining 19,732 images are used as gallery images.

Method	Market						Duke					
	S. M.	S. D.	mAP	Rank1	Rank5	Rank10	S. M.	S. D.	mAP	Rank1	Rank5	Rank10
Supervised baseline	M^I	*Market*	0.811	0.930	0.974	0.985	M^I	*Duke*	0.704	0.848	0.920	0.943
DBC [Ding *et al.*, 2019]	M^I	None	0.413	0.692	0.830	0.878	M^I	None	0.300	0.515	0.646	0.701
GLO [Liu and Zhang, 2020]	M^I	None	0.457	0.774	0.880	0.901	M^I	None	0.364	0.605	0.722	0.757
UTAL [Li *et al.*, 2019a]	M^I	None	0.462	0.692	-	-	M^I	None	0.446	0.623	-	-
SpCL [Ge *et al.*, 2020b]	M^I	None	0.731	0.881	0.951	0.970	-	-	-	-	-	-
GCMT	M^I	None	**0.739**	**0.897**	**0.965**	**0.976**	M^I	None	**0.636**	**0.782**	**0.886**	**0.913**
ATNet [Liu *et al.*, 2019]	M^I	*Duke*	0.256	0.557	0.732	0.794	M^I	*Market*	0.249	0.451	0.595	0.642
PDA-Net [Li *et al.*, 2019b]	M^I	*Duke*	0.476	0.752	0.863	0.902	M^I	*Market*	0.451	0.632	0.770	0.825
DIM+GLO [Liu and Zhang, 2020]	M^I	*Duke*	0.651	0.883	0.947	0.963	M^I	*Market*	0.583	0.762	0.857	0.885
SSG [Fu *et al.*, 2019]	M^D	None	0.583	0.800	0.900	0.924	M^M	None	0.534	0.730	0.806	0.832
MEB-Net[§] [Zhai *et al.*, 2020]	M^D	None	0.760	0.899	0.960	0.975	M^M	None	0.661	0.796	0.883	0.922
MEB-Net [Zhai *et al.*, 2020]	M^D	None	0.722	-	-	-	-	-	-	-	-	-
Co-teaching [Han *et al.*, 2018]	M^D	None	0.651	0.825	0.918	0.934	M^M	None	0.557	0.719	0.835	0.881
MMT [Ge *et al.*, 2020a]	M^D	None	0.712	0.877	0.949	0.969	M^M	None	0.631	0.768	0.880	0.922
GCMT	M^D	None	**0.771**	**0.906**	**0.963**	**0.977**	M^M	None	**0.678**	**0.811**	**0.911**	**0.939**

Table 1: Comparison with state-of-the-art methods on *Market* and *Duke*. "S. M." and "S. D." denote source model and labeled source data, respectively. [§] denotes DenseNet121 is used as backbone.

Method	S. M.	S. D.	mAP	Rank1
SpCL [Ge *et al.*, 2020b]	M^I	None	0.191	0.423
GCMT	M^I	None	0.237	0.543
GCMT	M^C	None	0.263	0.573
ECN [Zhong *et al.*, 2019]	M^I	*Market*	0.085	0.253
DIM+GLO [Liu and Zhang, 2020]	M^I	*Market*	0.207	0.497
SSG [Fu *et al.*, 2019]	M^M	None	0.132	0.316
MMT [Ge *et al.*, 2020a]	M^M	None	0.229	0.492
GCMT	M^M	None	**0.249**	**0.548**
ECN [Zhong *et al.*, 2019]	M^I	*Duke*	0.102	0.302
DIM+GLO [Liu and Zhang, 2020]	M^I	*Duke*	0.244	0.565
SSG [Fu *et al.*, 2019]	M^D	None	0.133	0.322
MMT [Ge *et al.*, 2020a]	M^D	None	0.233	0.501
GCMT	M^D	None	**0.266**	**0.579**

Table 2: Comparison with state-of-the-art methods on *MSMT*. "S. M." denotes source model. "S. D." denotes labeled source data.

Market-1501 contains 32,668 images of 1,501 identities. 12,936 images of 751 identities are selected for training. In the rest of images, 3,368 images are selected as query images and remaining 19,732 images are used as gallery images.

MSMT17 contains 126,441 images of 4,101 identities. 32,621 images of 1,041 identities are selected for training. In the rest of images, 11,659 images are selected as query images and remaining 82,161 images are used as gallery images.

For short, we denote *DukeMTMC-reID* as *Duke*, *Market-1501* as *Market*, and *MSMT17* as *MSMT* in the rest of this paper. Following previous works [Ge *et al.*, 2020a], Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are used to evaluate the performance. Rank1, Rank5 and Rank10 accuracies in CMC are reported.

5.2 Implementation Detail

Public models pre-trained on *ImageNet* [Deng *et al.*, 2009], *CUHK03* [Li *et al.*, 2014] (14,096 images with 1,467 identities for training), *Duke*, *Market*, and *MSMT17* are obtained following [Ge *et al.*, 2020a] and denoted as M^I, M^C, M^D, M^M, and M^{MS}, respectively. These pre-trained models adopt ResNet50 [He *et al.*, 2016] as backbone. Input images are resized to 256×128. We use random flipping, random cropping, and random erasing [Zhong *et al.*, 2020] for data augmentation. We feed the same image batch to different pairs of teacher and student networks but with separately random augmentation. K-Means method is used for unsupervised clustering. The number of clusters is set to 500 on *Duke* and *Market* and 1,500 on *MSMT* following [Ge *et al.*, 2020a; Zhai *et al.*, 2020]. In each training batch, 16 identities are randomly selected and 4 images for each identity are selected, resulting 64 images. K is set as 12 in teacher graph construction. Loss weight λ_{GCC} is set as 0.6. β is set to 0.05 following [Liu and Zhang, 2020]. Adam optimizer is used for training. Learning rate is initialized as 0.00035 and decayed by 0.1 after 20 epochs. Models are totally trained for 120 epochs with 400 iterations in each epoch. After training, teacher networks are used to extract features for ReID and the best performance among different teacher networks is reported. Models are trained on a server with three GeForce RTX 2080 Ti GPUs and one Tesla V100 GPU.

5.3 Compared with State-of-the-art Methods

In this section, proposed GCMT method is compared with many state-of-the-art methods. For fair comparison, GCMT uses only one pre-trained model in this section. The comparison results are summarized in Table 1 and Table 2. It can be observed that proposed GCMT method outperforms previous methods by a clear margin. It is worth noting that GCMT with M^I, *i.e.*, in the unsupervised setting, even outperforms most recent domain adaptive methods.

On *Market* and *Duke*, proposed GCMT is compared with recent state-of-the-art methods as shown in Table 1. It is clear that GCMT achieves the best perform compared with others. For example, GCMT achieves 77.1% and 67.8% in mAP accuracy on *Market* and *Duke*, respectively. This outperforms MMT [Ge *et al.*, 2020a] by 5.9% and 4.7%, respectively. MEB-Net[§] [Zhai *et al.*, 2020] uses DenseNet121 [Huang *et al.*, 2017] as backbone, while GCMT still outperforms it by 1.1% and 1.7% in mAP accuracy on *Market* and *Duke*, respectively. This indicates that proposed GCMT is powerful to learn discriminative person features without annotation.

β	0.01	0.03	0.05	0.07	0.1
Market	0.721	0.732	**0.739**	0.727	0.691
Duke	0.611	0.621	**0.636**	0.633	0.601

Table 3: Evaluation on β.

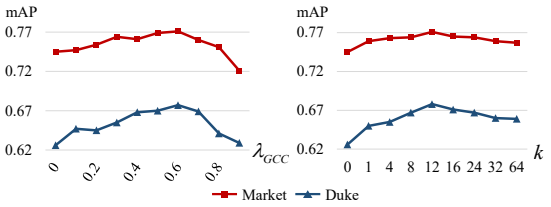

Figure 3: Parameter analysis on λ_{GCC} and k on *Market* and *Duke*.

MSMT is more challenging than *Market* and *Duke* due to more identities and diverse appearance. We compare GCMT with state-of-the-art methods as shown in Table 2. It is clear that GCMT outperforms previous methods by a clear margin. For example, GCMT achieves 54.8% in Rank1 accuracy when using M^M, outperforming DIM+GLO and MMT by 5.1% and 5.6%, respectively. Comparison on *MSMT* further demonstrates the effectiveness of proposed GCMT method on large-scale person ReID task.

5.4 Model Analysis

In this section, we first analysis the effect of β, λ_{GCC} and K respectively by varying the value of one parameter and keeping others fixed to the optimal value. Experiments show that hyper-parameters selected on one dataset can be applied to others. We then evaluate proposed GCC to show its validity. Finally, we show that GCMT could effectively boost the performance with multiple pre-trained models.

Evaluation on β is performed on *Market* and *Duke* using M^I as source model, as shown in Table 3. It is clear that setting β to 0.05 achieves the best performance.

Evaluation on λ_{GCC} is performed on *Market* and *Duke* with pre-trained models M^D and M^M, respectively. The experimental results are shown in Fig. 3. It can be observed that setting λ_{GCC} in the range from 0.1 to 0.7 can always improves the performance and the best performance is achieve when setting λ_{GCC} to 0.6 on both dataset.

Evaluation on K is also performed on *Market* and *Duke* with pre-trained models M^D and M^M, respectively. The evaluation results are shown in Fig. 3. Setting $K = 0$ denotes method without $\mathcal{L}_{GCC}$. It can be observed that setting K larger than 0 always improves the performance compared with $K = 0$ and the best performance is achieve when $K = 12$ on both datasets. This shows the validity of using GCC in GCMT for feature learning.

Evaluation on proposed GCC is performed on *Market* and *Duke* with pre-trained models M^D and M^M, respectively. The comparison results are shown in Table 4. It can be observed that GCC improves the performance on both datasets. For example, GCC improves mAP accuracy by 2.6% and 5.2% on *Market* and *Duke*, respectively. It is also clear that GCC outperforms soft triplet loss used in [Zhai *et al.*, 2020], *e.g.*, by 3.6% in mAP on *Duke*.

Method	Market		Duke	
	mAP	Rank1	mAP	Rank1
w/o $\mathcal{L}_{GCC}$	0.745	0.887	0.626	0.781
Replace $\mathcal{L}_{GCC}$ with soft triplet	0.757	0.891	0.642	0.778
GCMT	**0.771**	**0.906**	**0.678**	**0.811**

Table 4: Evaluation on GCC on *Market* and *Duke*.

Method	Market		Duke	
	S. M.	mAP	S. M.	mAP
GCMT	M^D	0.771	M^M	0.678
GCMT	M^{MS}	0.776	M^{MS}	0.681
GCMT	M^C	0.752	M^C	0.637
GCMT	$M^D + M^{MS}$	0.786	$M^M + M^{MS}$	0.688
GCMT	$M^D + M^C + M^{MS}$	**0.797**	$M^M + M^C + M^{MS}$	**0.691**
MEB-Net*	$M^D + M^C + M^{MS}$	0.731	$M^M + M^C + M^{MS}$	0.636

Table 5: Performance comparison with different source models on *Market* and *Duke*. * denotes our own implementation.

With multiple pre-trained models, multiple pairs of teacher and student networks are used in training. The performance comparison is shown in Table 5. It can be observed that, given multiple pre-trained models, GCMT is able to effectively boost the performance. For example, GCMT improves the mAP accuracy to 78.6% on *Market* by using two pre-trained models M^D and M^{MS}. GCMT further boosts the mAP accuracy to 79.7% on *Market* by additionally using M^C. It is worthy noting that the improvement of using multiple source models is based on high performance by single source model, which approaches the supervised baseline as shown in Table 1. Thus, the improvement is significant though it may be of small scale. We also implement MEB-Net [Zhai *et al.*, 2020] method with three pre-trained models by provided source code. It can be observed that GCMT outperforms MEB-Net by a clear margin, *e.g.*, by 6.6% in mAP on *Market*. This further shows the advantage in teacher networks cooperation of our method.

6 Conclusion

This paper proposes a Graph Consistency based Mean-Teaching (GCMT) method for unsupervised domain adaptive person ReID task. GCMT fuses sample similarity relationships predicted by different teacher networks as supervise signal based on graphs. With fused sample similarity relationships, GCMT uses the proposed Graph Consistency Constraint (GCC) to train student networks with more sample similarity relationships involved. GCMT provides an effective method for teacher networks cooperation and student networks optimization. Extensive experiments show that our GCMT outperforms state-of-the-art methods by a clear margin. Experimental results also show that GCMT effectively improves the performance given multiple pre-trained models.

Acknowledgments

This work is supported in part by Peng Cheng Laboratory, in part by The National Key Research and Development Program of China under Grant No. 2018YFE0118400, in part by Natural Science Foundation of China under Grant No. U20B2052, 61936011, 61620106009, in part by Beijing Natural Science Foundation under Grant No. JQ18012.

References

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Ding *et al.*, 2019] Guodong Ding, Salman Khan, and Zhenmin Tang. Dispersion based clustering for unsupervised person re-identification. In *BMVC*, 2019.

[Fu *et al.*, 2019] 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 *ICCV*, 2019.

[Ge *et al.*, 2020a] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020.

[Ge *et al.*, 2020b] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[Li *et al.*, 2019a] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE T-PAMI*, 2019.

[Li *et al.*, 2019b] 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 *ICCV*, 2019.

[Lin *et al.*, 2019] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019.

[Liu and Zhang, 2020] Xiaobin Liu and Shiliang Zhang. Domain adaptive person re-identification via coupling optimization. In *ACM MM*, 2020.

[Liu *et al.*, 2019] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, 2019.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[Wei *et al.*, 2018] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[Yu *et al.*, 2019] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *CVPR*, 2019.

[Zhai *et al.*, 2020] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, and Rongrong Ji. Multiple expert brainstorming for domain adaptive person re-identification. In *ECCV*, 2020.

[Zhang *et al.*, 2019] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV*, 2019.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[Zheng *et al.*, 2017] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[Zhong *et al.*, 2019] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019.

[Zhong *et al.*, 2020] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020.