Chapter 5

Learning a Domain-Invariant Embedding for Unsupervised Person Re-identification

In Chapters 2, 3 and 4, we have proposed several methods to solve the challenges of the lifelong learning theme. Unlike the core problem in lifelong learning that is caused by the nature of the stream data, we move our research focus to a more general situation where models are supposed to learn on the data lacking annotations, e.g., unsupervised domain adaptation (UDA). In this chapter, therefore, we aim to study what information the unlabeled data can provide in UDA ReID (**RQ 4**).

Although recent ReID works have achieved human-level accuracy on several ReID benchmarks, their successes heavily depend on large pre-labeled datasets for deep model training. These methods are not always suitable for real-world applications since practical scenarios often lack labeled data. In order to tackle this drawback, we propose a novel domain-invariant embedding network (DIEN) to learn a domain-invariant embedding (DIE) feature by introducing a multi-loss joint learning with recurrent top-down attention (RTDA) mechanism. Furthermore, we propose an improved triplet loss to enable the model to utilize both source-domain (labeled) data and target-domain (unlabeled) data. We compare our method with recent competitive algorithms and also evaluate the effectiveness of the proposed modules.

This chapter is based on the following publication:

- **Pu, N**, Georgiou T., Bakker, E. M., and Lew, M. S. "Learning a Domain-Invariant Embedding for Unsupervised Person Re-identification.". International Joint Conference on Neural Networks, 2019.

5.1 Introduction

In recent years, rerson re-identification (ReID) in large-scale surveillance systems has been one of the most challenging and hottest topics of computer vision. The ReID technology helps us to match pedestrian images that include the same person and are captured by different cameras at different locations or the same cameras at different time.

In order to overcome various changes in appearance and environment, current deep learning based person ReID models focus on learning robust features automatically (e.g., end-to-end learning) instead of handcrafted features. Most existing ReID works focus on supervised methods. They utilize deep CNNs [141, 142, 143, 144] to extract robust feature representations. Nevertheless, these methods achieve significant performance improvements only when a large amount of labeled training data is available. In real ReID scenarios, by using mature pedestrian detection technology, we can conveniently obtain very large ReID datasets but without labels [145]. Labeling data is expensive and time-consuming. So, if we can transfer the ReID capability of a deep neural network that trained on a fully labeled dataset, to perform ReID on another unlabeled dataset, we may make accurate ReID more tractable. Usually, related works treat two different datasets as a source domain (fully labeled) and target domain (without label). It is well-known that ReID models trained on one domain often fail to generalize well to another [146]. Some researchers handle this problem by utilizing Unsupervised Domain Adaption (UDA) method [146, 147, 148, 149, 150, 151]. And other works treat it as a transfer learning problem [152]. Both approaches need to make use of the unlabeled data to alleviate this drawback. In general, we can regard this problem as a domain shift or dataset shift [153] and thus apply a domain adaptation method to solve it. However, in unsupervised domain adaptation person ReID the two datasets do not share class labels (person identity), which is different from in traditional domain adaptation method. The challenge lies in how to obtain semantically meaningful domain-invariant features with good robustness for each identity. That is the main goal of our work presented in this paper.

To address above mentioned problems, we proposed a Domain-Invariant Embedding Network (DIEN), taking advantage of both source-domain (labeled) data and target-domain (unlabeled) data by using a novel proposed centering constrained cross-domain triplet loss (CCCDTL) function, to learn a domain-invariant embedding (DIE) feature for cross-domain Person ReID. Due to the supervision of the source domain and the auxiliary information of the target domain, the DIE feature is not only very discriminative, but is also robust under domain shift.

To further improve the discriminative power of DIE feature and the supervised information propagation, we introduce a new Recurrent Top-Down Attention (RTDA) module to recurrently find the region of interest on feature maps and re-weight each

channel of the feature maps to enable knowledge distillation. This is achieved by multi-loss joint learning and iteratively updating the parameters of the attention module. After finishing DIE feature learning, our model can perform cross-domain ReID by directly retrieving DIE features of the query image and the gallery images.

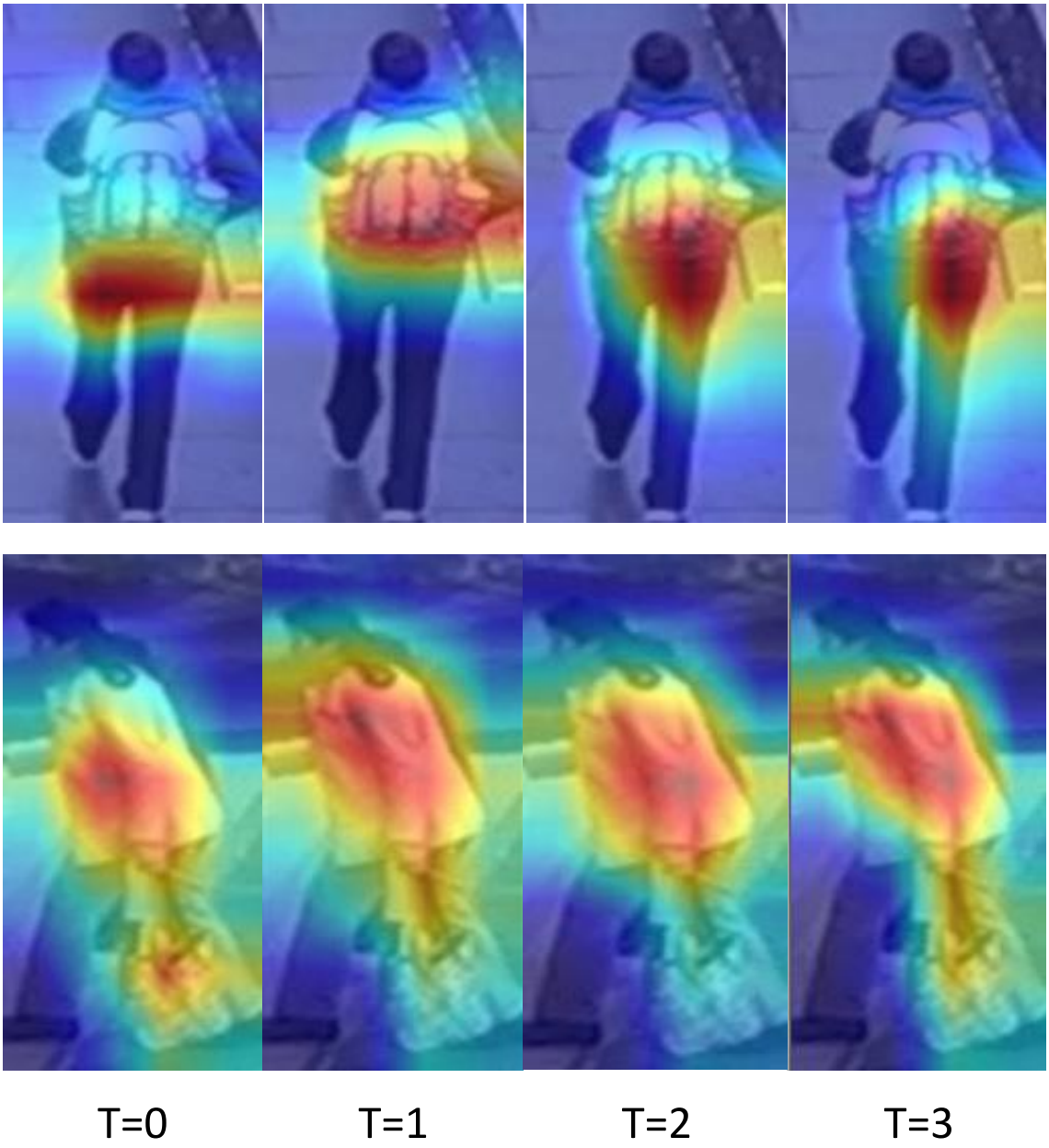

Figure 5.1: Grad-CAM [154] visualization results for different Ts.

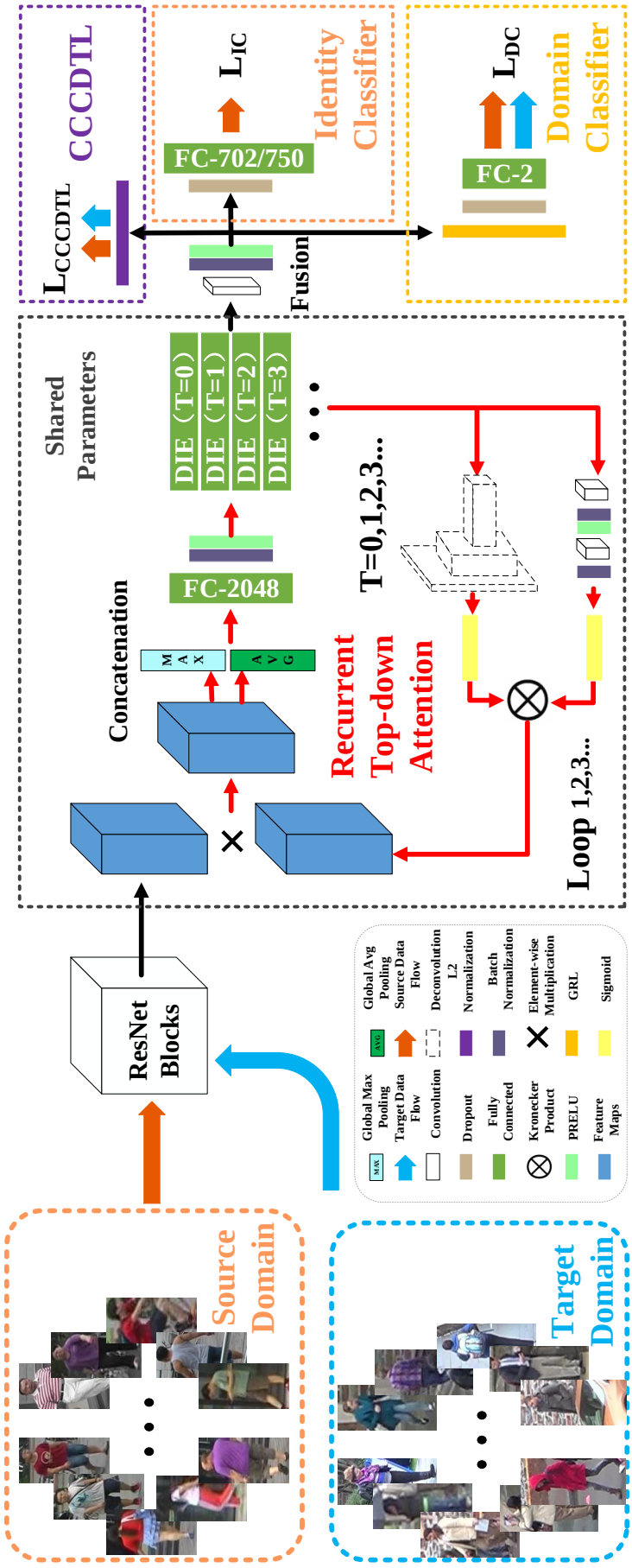

Figure 5.2: Illustration of the deep domain-invariant embedding neural network (DIEN). The proposed model consists of a backbone network, an identity classifier (IC), a centering constrained cross-domain triplet loss (CCCDTL) function, a domain classifier (DC) and a Recurrent Top-Down Attention (RTDA) module.

The main novel contributions of our paper can be summarized as follows:

- We propose a novel centering constrained cross-domain triplet loss (CCCDTL) function to achieve cross-domain learning. By using this loss function, our model can make full use of labeled and unlabeled data simultaneously.

- Our proposed DIE Network (DIEN) is a new end-to-end deep domain adaptation model. It is capable of learning domain-invariant embedding (DIE) features and recurrently refine learned feature by the Recurrent Top-Down Attention (RTDA) module proposed in this paper.

5.2 Related Work

5.2.1 Person Re-Identification (ReID)

Supervised Learning for ReID: Most existing ReID models [141, 142, 143] are trained using supervised learning strategies. For example, in order to handle body parts misalignment, Suh et al. [141] propose a two-stream network to learn a part-aligned representation for person ReID by using a bilinear-pooling layer. Further, He et al. [142] present a Deep Spatial feature Reconstruction (DSR) method to address the partial person ReID problem. Recently, Conditional Random Fields (CRFs) are exploited to mine second-order relationships of mini-batch training data in [143], which dramatically improves the performance of deep neural networks for ReID. Although those methods achieve a significant increasing performance on recent datasets, namely Market-1501 [155] and DukeMTMC-ReID [156], these methods may not be practical since collecting a large amount of annotated training data depends on lots of manpower and time.

Unsupervised Learning for ReID: To alleviate the above limitation, researchers also focus on person ReID using unlabeled training data [149, 157]. As an example, Li et al. [157] take full advantage of the information of cameras in the target domain, treating multiple one-person images from different cameras as a tracklet. Another typical work introduces a progressive unsupervised learning (PUL) method [149], which utilized a clustering method to select representative samples to modify the pre-trained model. PUL aimed at transferring pre-trained deep representations to an unseen domain by a Self-paced Learning. Nevertheless, due to the lack of label information for images across different cameras, unsupervised learning based methods typically can not perform as well as the supervised methods do.

5.2.2 Unsupervised Domain Adaptation for ReID

Unsupervised domain adaptation (UDA) has been studied widely in various computer vision tasks [150, 153, 158] and recently faces new challenges in person ReID [148].

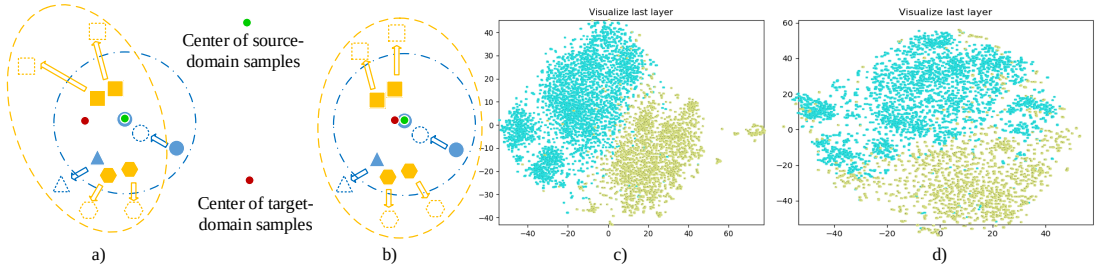

Figure 5.3: Graph a) and b) reprents different effect of Triple loss in [10] and our CCCDTL. The t-SNE of the pre-train features and learned DIE features are drawn in scatter c) and d) respectively (blue and green point donates target and source data respectivel).

From a UDA perspective, most related works are concetrated around Domain-Invariant Feature Learning [150]. Some recent works leverage an auto-encoder to achieve knowledge distillation [151] so as to learn a domain-invariant representation with significant generalization. In order to increase reasonable cues for person ReID and decrease the influence of camera variance, Zhong et al. utilize CamStyle [147] to generate camera-style images [148] as extra training data. In [159], Wang et al. employed both the attribute and identity labels to encode an embedding feature to promote unsupervised cross-dataset or cross-domain ReID. Due to the success of generation models, now many cross-domain tasks are dominated by GAN-based methods such as Similarity Preserving Generative Adversarial Network (SPGAN) [146] and Person Transfer Generative Adversarial Network (PTGAN) [152]. Both of them showed that using data augmentation methods strengthens the ReID ability of a deep neural network on the target domain thus improving the performance and closing the domain gap.

In this paper, we follow the general setting of unsupervised domain adaptation as used in [148]. Specifically, we provide labeled source training images and unlabeled target training images as training data and evaluate the performance of the proposed model on the target testing database.

5.2.3 Attention Mechanism

It is well known that attention plays an important role in human perception [160]. Recently, there have been several attempts [144, 161] to incorporate attention processing, to improve the performance of CNNs in Person ReID tasks. For example, Li et al. [161] designed a new two-stream model to learn both local and global features by hard and soft attention interactive learning. By refining the feature maps, their network not only performs well but is also robust to noisy inputs.

Unlike the attention-based methods in[144, 161], our proposed Recurrent Top-Down Attention (RTDA) leads deep neural network to recurrently update parameters and

take the high-level feedback signal into feature extraction instead of extracting feature vectors based on one-pass of the data through the network forward.

5.3 Proposed Method

5.3.1 Network Architecture Overview

In our deep domain-invariant embedding neural network, we deploy five blocks of ResNet-50 [162] as a primary feature extractor and follow the training strategy in [148] which fine-tunes on the ImageNet pre-trained model. We construct our backbone network by replacing Global Average Pooling (GAP) layer and the last 1,000-dim fully connected (FC) layer with two pooling layers and a 2,048-dim FC layer followed by batch normalization (BN) [163] and PReLU [164], as shown in Fig.5.2.

Inspired by CBAM [165], we use both average-pooled and max-pooled features to keep the distinctive object clues gathered by max-pooling. Specifically, we concatenate the two outputs of the global max pooling and global average pooling and feed them to the next FC layer. The output of this FC layer is a 2,048-dim feature vector, which we call the "domain-invariant embedding" (DIE).

For the purposes of strengthening the information flow and distilling the DIE feature, Recurrent Top-Down Attention (RTDA) is exploited to recurrently re-weight the channel and spatial position of feature maps simultaneously. The RTDA module is implemented by multiple deconvolution and convolution layers whose details will be described in the Section 5.3.4. Through T (T $= 0, 1, 2, 3...$) loops, we employ an 1×1 convolution to fuse the output of each loop and obtain the final DIE feature vector. Subsequently, the DIE features are fed into the centering constrained cross-domain triplet loss (CCCDTL) function after L2-normalization. At the same time DIE features are forwarded to both identity classifier (IC) and domain classifier (DC) module.

The IC module consists of an FC layer and a Dropout layer[166]. This is a general multi-class classifier trained using standard cross-entropy loss function. This loss function is formulated as,

$$\begin{aligned} \mathcal{L}_{IC}(\mathbb{I}^s) = -\frac{1}{|\mathbb{I}^s|}\sum_{I\in\mathbb{I}^s}(y_i \log \mathbb{P}(I) \\ + (1-y_i)\log(1-\mathbb{P}(I))) \\ with \quad \mathbb{I}^s \cup \mathbb{I}^t = \mathbb{I} \end{aligned} \tag{5.1}$$

where $\mathbb{I}$ represents images in a training mini-batch. $\mathbb{I}^s$ denotes images from the source (labeled) domain and $\mathbb{I}^t$ represents images from the target (unlabeled) domain. $\mathbb{P}(I)$

is the predicted probability of image I belonging to class y_i and $|\cdot|$ denotes the number of samples in set "$\cdot$" .

5.3.2 Centering Constrained Cross-domain Triplet Loss

As triplet loss (TL) benefits from hard mining and metric learning, TL is a very common loss function in supervised ReID. In [148], Zhong et al. treat each two images from the target domain and the source domain as a negative pair, which enables TL to be used for cross-domain training. Following the assumption in [148], each image in target domain is assumed to have a different identities, since labels of the target domain are not available. The aforementioned strategy leads to a mistake when applied to cross-domain ReID. Even if two images from the target domain belong to the same person, they will be pushed away if TL is used in this way, which is demonstrated in Fig.5.3 a).

So, in order to alleviate this issue, we introduce a correction. More specifically, we mine hard positive pair only in the source domain and hard nagetive pair in both the source and the target domain. Meanwhile, we also introduce the Maximum Mean Discrepancy (MMD) distance to constrain target images which are pushed far away from the position where they should be. Finally, we propose a centering constrained cross-domain triplet loss (CCCDTL) function to further improve discernment of embedding features by closing the farthest intra-class distance, pushing closest inter-class distance and minimizing the distance between the source and target distributions simultaneously, which is shown in Fig.5.3 b) and is formulated as,

$$\begin{aligned}\mathcal{L}_{CCCDTL}(\mathbb{I}) = &\sum_{I_a, I_p \in \mathbb{I}^s, I_n \in \mathbb{I}} max\{D(\phi(I_a), \phi(I_p)) \\ &- D(\phi(I_a), \phi(I_n)) + m, 0\} \\ &+ \lambda \times D(\frac{1}{|\mathbb{I}^s|}\sum_{I \in \mathbb{I}^s}\phi(I), \frac{1}{|\mathbb{I}^t|}\sum_{I \in \mathbb{I}^t}\phi(I)) \\ &with \qquad \mathbb{I}^s \cup \mathbb{I}^t = \mathbb{I},\end{aligned} \tag{5.2}$$

where λ is a hyperparameter to balance the importance of two terms. I_a is an anchor point. I_p is the hardest (farthest) sample in the same class with I_a, and I_n is the hardest (closest) sample with a different class for I_a. m is a margin parameter and D is the Euclidean distance between two embedding feature vectors.

5.3.3 Domain-invariant Embedding by Gradient Reversal Layer

Inspired by conventional unsupervised domain adaptation methods, we utilize the gradient reversal layer (GRL) in [158] to construct a domain classifier (DC) module to further improve the domain-invariant capability of DIE features.

Based on the covariate shift assumption [167], we assume that there exist two distributions $S(I;L)$ and $T(I;L)$, where I and L donate images and labels, respectively. There are referred as the source distribution and the target distribution. Both distributions are assumed to be very complex and unknown, and furthermore similar but different. In order to obtain a similar ReID performance on both target domain without labels and source domain with labels, we should constrain two distributions (i.e., S and T) to be similar. Unfortunately the distributions are unknown and can be very complex, which makes this problem difficult to solve. So, we reversely consider this problem that making two distributions as different as possible is equivalent to classifying them. With the help of gradient reversal layer (GRL), we can transfer the classified supervised signal to an indiscriminate (domain-invariant) supervised signal, which is formulated as,

$$\begin{aligned} \mathcal{L}_{DC}(\mathbb{I}) = -\frac{1}{|\mathbb{I}|}\sum_{I\in\mathbb{I}}(\Gamma_I \log \mathbb{P}(I) \\ + (1-\Gamma_I)\log(1-\mathbb{P}(I))) \\ \Gamma_I \begin{cases} 1 & , I \in \mathbb{I}^t \\ 0 & , I \in \mathbb{I}^s \end{cases}, \end{aligned} \tag{5.3}$$

where Γ_I is an indicator function to index which domain image I belongs to. During the backpropagation processing, GRL makes the gradient negative and feeds it beck to next layer, which is formulated as,

$$\theta \leftarrow \theta - \alpha(\frac{\partial \mathcal{L}_O}{\partial \theta} - \frac{\partial \mathcal{L}_{DC}}{\partial \theta}), \tag{5.4}$$

where $\mathcal{L}_O$ represents the loss functions other than $\mathcal{L}_{DC}$, θ are all the parameters of the whole nueral network and α is the step size of SGD.

Eventually, the backbone network learns to generate domain-invariant representation. Note that the use of GRL during the several initial epochs of training is not stable. This is because the backbone network is struggling to find the optimal path in the beginning, due to the entangled supervision signal yielded by GRL. After model has acquired robust representative capability, the GRL guides the representation towards better generalization.

Table 5.1: Ablation studies of Domain-Invariant Embedding Network under different configurations.

configurations	Duke==>Market				Market==>Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Base	45.1	62.7	70.1	20.8	32.9	49.7	54.9	17.0
Base+TL[148]	49.9	67.7	74.8	23.9	37.0	52.4	59.3	21.1
Base+CCCDTL	51.7	68.2	75.0	24.9	38.3	54.1	60.9	22.1
Base+DC	51.2	67.7	73.7	24.2	37.9	53.5	60.1	21.8
Base+CCCDTL+DC	53.6	69.9	76.1	26.3	39.8	55.3	62.5	22.3
Base+CCCDTL+DC+RTDA (T = 1)	58.1	74.1	80.9	24.8	44.4	60.1	66.1	24.2
Base+CCCDTL+DC+RTDA (T = 2)	58.5	74.9	81.4	26.9	45.8	61.6	67.8	26.1
Base+CCCDTL+DC+RTDA (T = 3)	**58.7**	**75.4**	**81.9**	**27.1**	**46.7**	**62.5**	**68.3**	**26.4**
Base+CCCDTL+DC+RTDA (T = 4)	58.6	75.2	81.5	27.0	46.4	62.3	68.0	26.2
Base+CCCDTL+DC+RTDA (T = 5)	57.9	73.9	80.4	26.3	44.1	61.9	67.8	24.0

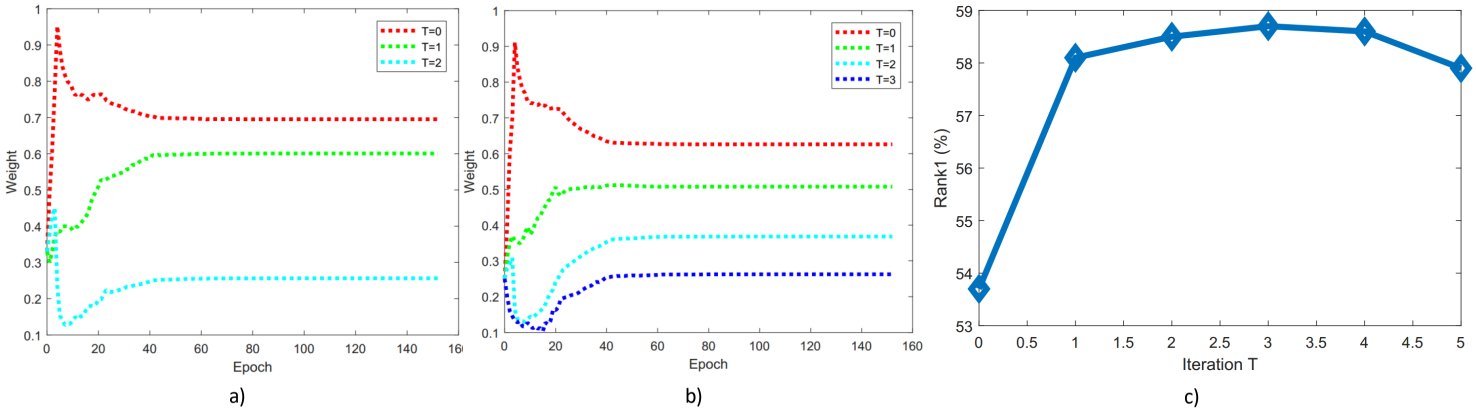

Figure 5.4: a) The learned weight of fusing DIE feature (T = 2). b) The learned weight of fusing DIE feature (T = 3). c) Rank-1 accuracy of our model trained on the Duke dataset and tested on the Market dataset with different Ts.

5.3.4 Recurrent Top-down Attention

Most attention-based ReID models implement attention mechanism by utilizing extra neural network modules to predict where the model should focus on. These modules usually consist of Multi Layer Perceptrons (MLPs) and rely on the outputs of lower layers. The feature extraction and the attention prediction in the current layer work independently and at the same time. Nevertheless, it is well known that when an object catches our attention, our brain makes a decision to focus on the discriminative region, which should be a top-down (from brain to visual system) procedure. Similarly, our models also need to use the high-level feedback signal to guide feature extraction instead of using low-level features. Thus, we mimic the human visual attention process when addressing the ReID problem from in a complicated image, taking a first glimpse and then rethink several times, to optimize attention.

When a mini-batch of input images pass through all the layers, instead of immediately generating DIE feature vectors, a feedback module is deployed to propagate the supervised information to the bottom layers and update the network. On the one hand, intuitively, when two images with different identities have similar DIE features, they are not easy to be distinguished. Instead of outputting the feature vector directly, a better way is to recurrently guide the previous layers based on the primary DIE feature (when T equals to 0), such that the bottom layers can be strengthened or weakened to produce more discriminative features specifically for those identities that are difficult to distinguish. Furthermore, through aforementioned loss function, the DIE feature from the top layer will be more domain-invariant which is often come from high-level information. Thus we propose a new Recurrent Top-Down Attention (RTDA) module and allow DIE network to use the high-level feedback information for feature extraction.

More specifically, we utilize the primary DIE feature during the first "glimpse" the image to predict the spatial positions of interest (spatial attention) and the weights of channels (channel attention) on the feature maps, and then make the network refocus on those regions and rethink emphasized or suppressed channels. In detail, spatial attention is implemented by three deconvolutional layers followed by a sigmoid layer, and channel attention consists of two convolutional layers and a sigmoid layer, as shown in Fig.5.2. We employ the Kronecker product to combine spatial- and channel-attention, which generates a spatial-channel mask with the same dimensions as the feature maps. After that these feature maps are updated by element-wise multiplication. After T times recurrent forward propagating, we fuse the DIE features from each loop by a weighted sum where the weights are learned by an 1×1 convolution and are initialized to $\frac{1}{T+1}$. Notably, our experiments on benchmark datasets clearly demonstrate the advantage of the RTDA algorithm in cross-domain ReID, which is reported in Section 5.4.2.

Table 5.2: Comparison with State-of-the-art Methods.

Methods	Duke==>Market				Market==>Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
UMDL [168]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PUL [149]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
SPGAN [150]	57.7	**75.8**	**82.4**	26.7	46.4	62.3	68.0	26.2
TJ-AIDL[159]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
ours (T = 3)	**58.7**	75.4	81.9	**27.1**	**46.7**	**62.5**	**68.3**	**26.4**

5.3.5 Multi-loss Joint Learning

In order to confirm that all modules work harmoniously and allow the proposed neural network to be trained in an end-to-end manner, we sum the three loss functions to form the final loss, which is written as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{IC} + \beta_1 \mathcal{L}_{CCCDTL} + \beta_2 \mathcal{L}_{DC} \tag{5.5}$$

where β_1 and β_2 are hyper parameters to balance the importance of the three terms. Through cross validation, we set β_1, β_2 and λ to 1, 0.1 and 1 respectively.

In addition, we adopt the stochastic gradient descent (SGD) method to update the parameters of the network while different learning rates are applied on different layers. More specifically, the weights of the pre-trained primary feature extractor should not be updated as fast as the other modules because we should keep the useful information acquired by training on ImageNet. Hence, we set the learning rate for the backbone network to a relatively smalll value, more specifically to 10^{-4}. For the other modules, IC, CCCDTL, DC and RTDA, the learning rate are 10^{-1}, 10^{-1}, 10^{-1} and 10^{-2} respectively.

5.4 Experiments and Analysis

5.4.1 Datasets

The performance of our proposed method on the task of ReID is evaluated on two popular benchmark datasets: Market-1501 [155] and DukeMTMC-reID [156].

Market-1501 consists of 32,668 labeled images of 1,501 identities collected from 6 camera views. All of the identities are divided into two parts: 12,936 images from 751 identities for training and 19,732 images from 750 identities for testing. During testing, 3368 query images from 750 identities are treated as probe for matching persons in the gallery.

DukeMTMC-reID is also a large-scale ReID dataset. It is collected from 8 cameras and contains 36,411 labeled images belonging to 1,404 identities. It consists of 16,522 training images from 702 identities, 2,228 query images from the other 702 identities, and 17,661 gallery images.

We use rank-1 accuracy and mean average precision (mAP) for our evaluation on both datasets. In the experiments, there are two source-target settings:

1. Target: Market-1501 / Source: DukeMTMC-reID.

2. Target: DukeMTMC-reID / Source: Market-1501.

5.4.2 Ablation Study

In order to analyze the effectiveness of the proposed Domain-Invariant Embedding Network, we compare the baseline model with ten different configurations. For our model we use DIE features from different the Ts after L2-normalization as retrieved vectors in testing. The result of each experiment is reported in the each row of Table 5.1.

Effectiveness of CCCDTL and DC module. As for the first configuration in Table 5.1, our baseline model consists of the backbone network and the IC module, which is trained on the source datasets and directly evaluated on the test set of the target dataset. The second and third experiments aim at comparing the triplet loss function in [148] with our CCCDTL model. The results show that our method addresses the aforementioned mistake and shows a better performance. Furthermore, by adding an independent DC module into the base model, Rank-1 accuracy is increased by 6.1%. From Fig. 5.3 c) and d), we can see that the two distributions of the different domains blend into each other due to the effectivity of the DC module. Furthermore, our experiments show that all of the proposed modules are effectively increasing accuracy. Furthermore, combining with third, fourth and fifth row of Table 5.1, the proposed modules do not conflict to each other, combining all loss functions achieves 53.6% at Rank-1 accuracy.

Effectiveness of RTDA module. Firstly, in order to investigate the influence of different T = 1s, we conducted several experiments using T (T = 1, 2, 3, 4 and 5). From Figure 5.4, it is obvious that increasing the number of recursive steps for information feedback allows the bottom layers to receive richer top-down information. We observe from our experiments that after T > 3 the performance decreases since the model is overfitted. Empirically, we set T = 3 in the training phase of DIEN to compare with the state-of-the-art. Furthermore, combining Fig.5.4 a), b) and Fig.5.1 leads us to think that our model extracts coarse features which contain a large proportion of the information when T = 0. With the a increase in iterations, extracted features are more fine-grained with smaller proportions of information. Benefiting from aggregating multi-step features, the ReID performance of our model significantly increases again. Furthermore, from the Grad-CAM visualization results with different Ts in Figure 5.1, we can observe that domain-invariant features pay more attention on discriminative cues but not on the complete foreground. Finally, our proposed model achieves 58.7% at Rank-1 accuracy on the Market-1501 dataset.

Hence, adding RTDA modules does help perform representation learning and it is cooperating with the DC and CCCDTL modules.

5.4.3 Comparison with State-of-the-art Methods

We compared our method with the state-of-the-art unsupervised learning methods. Table 5.2 presents the comparison when Market-1501 is the source set and Duke is the target set and viceversa. We compared with four unsupervised methods, including UMDL [168], PUL [149], SPGAN [150] and TJ-AIDL[159].

UMDL employed hand-crafted features and a multi-task dictionary learning method to learn cross-dataset feature, PUL is a typical post-processing method by reselecting training samples for fine-tuning a CNN model, SPGAN is a famous GAN-based baseline method for Person ReID, and TJ-AIDL is recently published and achieves the state-of-the-art result. Compared to the PUL method, our method achieves +13.2% higher rank-1 accuracy and a +6.6% improvement for mAP. As for the comparison with SPGAN, our method has +1.0% higher rank-1 accuracy and +0.2% higher mAP, while it is noted that GAN-based methods have significantly greater computational costs and memory consumption than our method and rely heavily on data augmentation. We also compare to TJ-AIDL and our results are slightly better than it, since TJ-AIDL given extra supervised information in the form of attributes of a person in the source dataset such as backpack or handbag et al. Our method without data augmentation has similar or better performance than all selected competitors.

5.5 Chapter Conclusions

In this paper, we proposed an end-to-end deep model, the Domain-invariant Embedding Network (DIEN), for solving cross-domain ReID tasks. Our DIEN utilizes both source-domain (labeled) datasets and target-domain (unlabeled) datasets as training data to explore the common cues of cross-domain ReID by jointly optimizing multiple loss functions. We also introduced a Recurrent Top-Down Attention module to refine the DIE features. Benefiting from the recurrent iteration, the model is able to extract more discriminative low-level features with the guidance from high-level information. With this proposed DIEN, we conducted experiments on the Market-1501 and the DukeMTMC-reID datasets, and evaluated the effectiveness of our model in different configurations. Finally, compared to several recent unsupervised person ReID methods, the proposed DIEN achieved state-of-the-art performance and reduced the gap between supervised and unsupervised methods. Our experimental findings can be summarized as:

(1) Even through we introduce label noise when we regard all the unlabeled identities as different identities, the model still acquires considerable improvements. This

phenomena implies that the deep model learns transferable knowledge in labeled data and is flexible enough to handle under-investigated data structures.

(2) Through observing the attention maps (in Figure 5.1) generated from the purposed recurrent top-down attention (RTDA), we find that with the growth of recurrent time, the corresponding focus area is more and more fine-graded, which shows a consistency with our human attention mechanism.

(3) By comparing the weights of fusing feature in different time step as illustrated in Figure 5.4, we find that models mainly depends on the first attentive feature to retrieval pedestrians. The subsequent attentive features are likely to serve as supplementaries regarding to key details.

Future work. On the one hand, our experiments provide some promising insights for understanding how the network works in UDA ReID, which is beneficial for not only ReID tasks but also the future explainable machine learning. Thus, we plan to explore more about the interpretation of deep neural network. On the other hand, since the proposed RTDA is a time-consuming feature refinement method, we plan to design a more efficient way to implement feature refinements.