UNSUPERVISED DEEP TRANSFER LEARNING APPROACH TO PERSON RE-IDENTIFICATION

[1]*Yu-Jhe Li,* [1]*Yu-Chiang Frank Wang,* [2]*Chiou-Sheng Fuh*

[1]Graduate Institute of Communication Engineering, National Taiwan University
[2]Department of Computer Science and Information Engineering
{r06942074, ycwang}@ntu.edu.tw, fuh@csie.ntu.edu.tw

ABSTRACT

Person re-identification (Re-ID) aims at recognizing the same person from images taken across different cameras. To address this task, one typically requires a large amount labeled data for training an effective Re-ID model, which might not be practical for real-world applications. To alleviate this limitation, we choose to exploit a sufficient amount of pre-existing labeled data from a different (auxiliary) dataset. By jointly considering such an auxiliary dataset and the dataset of interest (but without label information), our proposed adaptation and re-identification network (ARN) performs unsupervised domain adaptation, which leverages information across datasets and derives domain-invariant features for Re-ID purposes. In our experiments, we verify that our network performs favorably against state-of-the-art unsupervised Re-ID approaches, and even outperforms a number of baseline Re-ID methods which require fully supervised data for training.

Index Terms— Re-Id, Domain Adaptation, Transfer Learning, Computer Vision

1. INTRODUCTION

Person re-identification (Re-ID) [1] has become popular research topic due to its application to smart city and large-scale surveillance system. As depicted in Figure 1, given a person-of-interest (query) image, Re-ID aims at associating the same pedestrian from multiple cameras, matching people across non-overlapping camera views. Yet, current Re-ID models are still struggling to handle the problems with intensive changes in appearance and environment. With recent advances in deep neural networks, several works have been proposed to tackle the above challenges in supervised [2, 3, 4, 5, 6, 7] and unsupervised manners [8, 9, 10, 11].

However, the aforementioned methods are not able to achieve satisfactory performances if the appearance or camera settings of query images are very different from the training ones. This is known as the problem of *domain shift* (or *domain/dataset bias*) and requires domain adaptation [12] techniques to address this challenging yet practical problem. Thus, several works [6, 13] have been proposed to generalize the discriminative ability across different datasets by increasing the cross-domain training samples with style transfer methods. Zhong *et al.* [6] smooth style disparities across the cameras with style transfer model and label smooth regularization. Similarly, Deng *et al.* [13] further add similarity constraints to enhance the performance on cross-domain Re-ID task. However, the adaptation models based on style transfer are not necessary to preserve the identity during the image translation procedure, and this results in unsatisfactory performance when no corresponding identities appear in both domains/datasets.

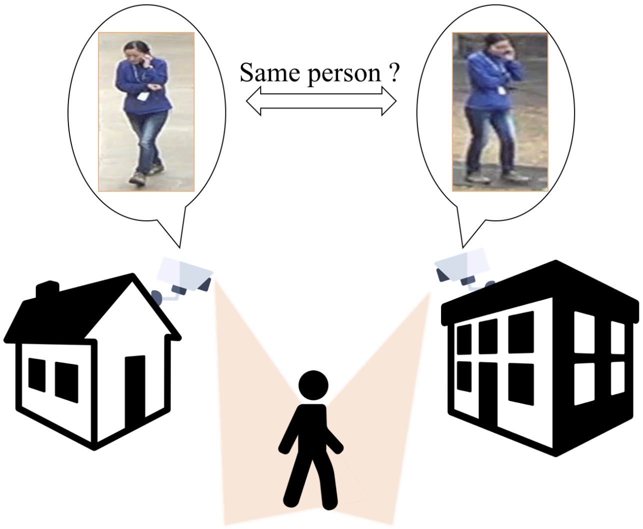

Fig. 1. Person re-identification (Re-ID) is a task which provides critical insights whether a person re-appears in another camera after being spotted in one camera. That is, given a person-of-interest (query) image, Re-ID aims at associating the same pedestrian from the datasets collected from multiple cameras.

To address the domain shifts between datasets, we propose a deep architecture to perform cross-domain Person

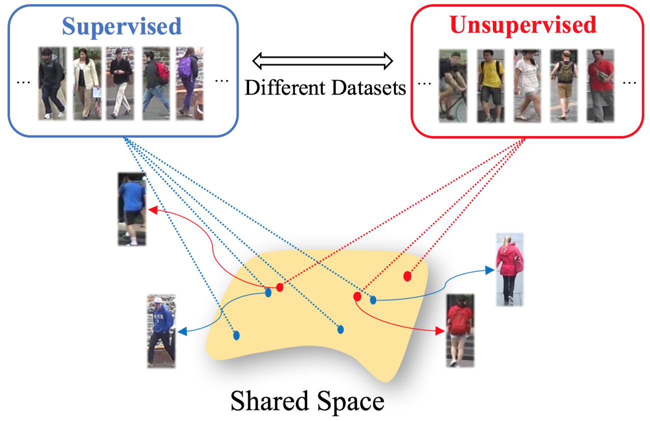

Fig. 2. Illustration of cross-dataset person re-identification (Re-ID). While Re-ID of images in the target-domain dataset is of interest, no labeled data is available for training. Our idea is to leverage information from auxiliary labeled images in a distinct and irrelevant source domain (i.e., dataset not of interest). With such an unsupervised domain adaptation setting for learning domain-invariant features, Re-ID in the target domain can be performed accordingly.

Re-identification with the only supervision from a single dataset/domain as shown in Figure 2. Toward this end, with the labeled data, our model derives the discriminative property to distinguish the images between different classes. To perform such property on alternative domain without annotation, our model learns to adapt the discriminative property from supervised (i.e., source) to unsupervised (i.e., target) domain. This is achieved by decomposing the cross-domain feature into domain-invariant and domain-specific one. Once the domain-invariant feature is learned, our model can perform cross-domain Re-ID by matching the query image and gallery images in the shared latent space. To further enhance the discriminative property of our proposed model, we aim at increasing the margin between the classes with our proposed contrastive objective, which is later verified in the experiment.

The contributions of our paper can be summarized as follows:

- We address unsupervised person Re-ID by exploiting and adapting information learned from an auxiliary labeled dataset, which can be viewed as a unsupervised domain adaptation approach.
- Our proposed Adaptation and Re-ID Network (ARN) aims at learning domain-invariant features for matching images of the same person, while no label information is required for the data domain of interest.
- Our ARN not only performs favorably against state-of-the-art Re-ID approaches in the unsupervised setting, it also outperforms baseline supervised Re-ID methods.

2. RELATED WORKS

2.1. Person Re-Identification (Re-ID)

Supervised Learning for Re-ID: Most existing Re-ID models are learned in a supervised setting. That is, given a sufficient number of labeled images across cameras, techniques based on metric learning [3] or representation learning [4] can be applied to train the associated models. Cheng *et al.* [3] propose a multi-channel part-based convolutional network for Re-ID, which is formulated via an improved triplet framework. Lin *et al.* [4] present an attribute-person recognition network which performs discriminative embedding for Re-ID and is able to make a prediction for particular attributes. While promising performances have been reported on recent datasets (e.g., Market-1501 [14], DukeMTMC-ReID [15]), it might not be practical since collecting a large amount of annotated training data is typically computationally prohibitive.

Unsupervised Learning for Re-ID: To alleviate the above limitation, researchers also focus on person Re-ID using unlabeled training data [9, 8]. For example, Fan *et al.* [9] apply techniques of data clustering, instance selection, and fine-tuning methods to obtain pseudo labels for the unlabeled data; this allows the training of the associated feature extractor with discriminative ability. Wang *et al.* [8] propose a kernel-based model to learn cross-view identity discriminative information from unlabeled data. Nevertheless, due to the lack of label information for images across cameras, unsupervised learning based methods typically cannot achieve comparable results as the supervised approaches do.

2.2. Cross-Domain Re-ID

Recently, some transfer learning algorithms [16, 10] are proposed to leverage the Re-ID models pre-trained in source datasets to improve the performance on target dataset. Geng *et al.* [16] transfer representations learned from large image classification datasets to Re-ID datasets using a deep neural network which combines classification loss with verification loss. Peng *et al.* [10] propose a multi-task dictionary learning model to transfer a view-invariant representation from a labeled source dataset to an unlabeled target dataset.

Besides, domain adaption and image–to–image translation approaches have been applied to Re-ID tasks increasingly, Deng *et al.* [13] combine CycleGAN [17] with similarity constraint for domain adaptation which improve performance in cross-dataset setting. Zhong *et al.* [6] introduce camera style transfer approach to address image style variation across multiple views and learn a camera-invariant descriptor subspace.

2.3. Domain-Invariant Feature Learning

We deal with the cross-domain Re-ID by learning domain-invariant feature. Here we review the recent works [18, 19, 20, 21] on learning domain-invariant feature. In order to achieve cross-domain classification tasks, Tzeng *et al.* [18] present domain confusion loss to learn domain-

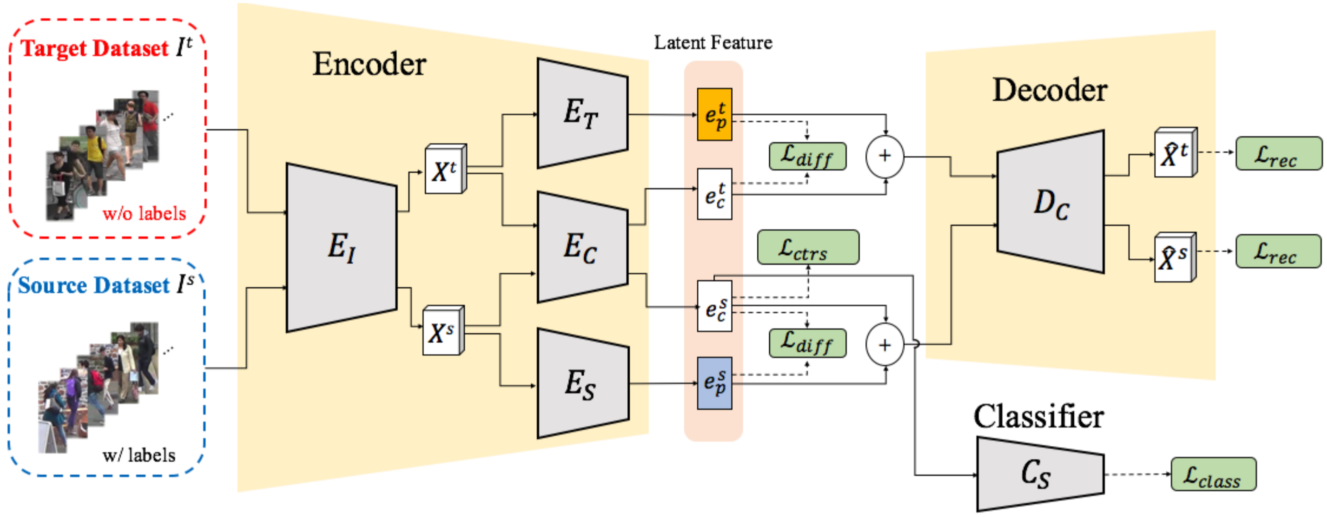

Fig. 3. The architecture of our Adaptation and Re-Identification Network (ARN). Note that the Encoder contains the two shared modules (E_I, E_C), and two private modules (E_T, E_S). E_I aims to retrieve visual feature maps (X^t, X^s), which are fed into E_C, E_S, and E_T for learning domain-invariant (shared) and specific (private) features. With the private (e_p^t, e_p^s) and shared (e_c^t, e_c^s) latent features observed, the Decoder D_C performs feature reconstruction for both target and source-domain images. Finally, the classifier C_S is designed to perform supervised learning from source-domain data.

invariant representation. Bousmalis *et al.* [19] propose to extract the domain-invariant feature to improve the performance of cross-domain classification task. On the other hand, to tackle the problem of image style translation, Coupled GAN [20] also learn to synthesize cross-domain images from a domain-invariant feature. UNIT [21] further learn a domain-invariant feature to translate the image across domains. It is worth noting that, inspired by the above methods, we address the cross-domain Re-ID task by learning the domain-invariant feature for describing the human identity across distinct domains.

3. PROPOSED METHOD

Given a set of image-label pairs $\{I_i^s, y_i^s\}_{i=1}^{N_s}$ and another set of images $\{I_i^t\}_{i=1}^{N_t}$, where N_s and N_t denote the total images of source and target dataset respectively, the goal of our model is to perform cross-dataset Re-ID by adapting the discriminative ability learned from source dataset to unlabeled target dataset.

We present our Person Re-ID model trained in a supervised manner in section 3.1. To address the cross-dataset Person Re-ID, our model leverages information from supervised data and adapts it to unsupervised dataset in section 3.2. Later in section 3.3, we demonstrate the learning and evaluation of our proposed method.

3.1. Supervised Learning for Person Re-ID

To perform person re-identification, our model aims to learn the image feature with discriminative property to distinguish between classes. With labeled data, such feature property can be learned from image classification task. To achieve this, we introduce encoder $\{E_I, E_C\}$ and classifier C_S to extract the image feature e_c^s from source dataset image I^s and obtain its category prediction $\hat{y}^s$ respectively. Specifically, to reduce the training burden, pre-trained model (e.g., ResNet) can be used for feature extractor module E_I. Thus, we define the classification loss $\mathcal{L}_{class}$ to minimize the negative log-likelihood of the ground truth label y^s for source dataset image I^s:

$$\mathcal{L}_{class} = -\sum_{i=1}^{N_s} y_i^s \cdot \log \hat{y}_i^s \tag{1}$$

To further enhance the discriminative property of our learned feature, we consider contrastive loss $\mathcal{L}_{ctrs}$ [22] as an additional objective of our model:

$$\mathcal{L}_{ctrs} = \sum_{i,j} \lambda(e_{c,i}^s - e_{c,j}^s)^2 + (1-\lambda)[max(0, m - (e_{c,i}^s - e_{c,j}^s))]^2 \tag{2}$$

where $\lambda = 1$ if $\{e_{c,i}^s, e_{c,j}^s\}$ belong to same category, and $\lambda = 0$ if $\{e_{c,i}^s, e_{c,j}^s\}$ belong to different categories. Note that $m > 0$ is a margin, which is regarded a radius around $E_c(x_i)$. Dissimilar pairs contribute to the loss function only if their distance is within this radius.

However, the above supervised model cannot be directly applied to alternative dataset without label annotation. Thus, we further consider the adaption technique to generalize the discriminative ability to dataset without any label annotation.

3.2. Unsupervised Domain Adaptation for Cross-Dataset Re-ID

Here we regard cross-dataset Re-ID as adaptation of discriminative ability from supervised source dataset to unsupervised target dataset. To this end, our model aims to eliminate dataset shift in the procedure of inferencing discriminative feature. Thus, as depicted in Figure 3, our model first introduces E_S/E_T to decompose visual feature maps X^s/X^t into dataset-invariant feature e_c^s/e_c^t and dataset-specific feature e_p^s/e_p^t. Our model acquires discriminative ability by applying the dataset-invariant feature e_c^s to predict its corresponding category. Once such feature is learned, even without supervision in target dataset, we can transfer discriminative knowledge from supervised to unsupervised dataset.

With the goal of reducing information loss in the above procedure for compressing the visual feature maps, here we consider a decoder D_C to reconstruct visual feature maps X^t/X^s from the compact domain-invariant and specific features $(e_c^t, e_p^t)/(e_c^s, e_p^s)$. Thus, we define reconstruction loss $\mathcal{L}_{rec}$ as:

$$\mathcal{L}_{rec} = \sum_{i=1}^{N_s} \|X_i^s - \hat{X}_i^s\|_2^2 + \sum_{i=1}^{N_t} \|X_i^t - \hat{X}_i^t\|_2^2 \qquad (3)$$

where X_i^s/X_i^t and $\hat{X}_i^s/\hat{X}_i^t$ denote encoded and reconstructed the visual feature maps for source/target dataset respectively.

Note that the above learning objectives cannot ensure that the dataset-invariant and specific feature are mutual exclusive and independent, we therefore introduce a difference loss $\mathcal{L}_{diff}$ to encourage the orthogonality between these two features:

$$\mathcal{L}_{diff} = \|{\mathbf{H}_c^s}^\top \mathbf{H}_p^s\|_F^2 + \|{\mathbf{H}_c^t}^\top \mathbf{H}_p^t\|_F^2 \qquad (4)$$

where $\mathbf{H}_c^s$ and $\mathbf{H}_c^t$ be matrices whose rows are the latent *shared* representations $e_c^s = E_C(X^s)$ and $e_c^t = E_C(X^t)$. $\mathbf{H}_p^s$ and $\mathbf{H}_p^t$ are obtained in a similar manner. Note that $\|\cdot\|_F^2$ is the square Frobenius norm.

3.3. Learning and Performing Re-ID

In sum, the total training objective $\mathcal{L}_{total}$ for our ARN can be written as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{class} + \alpha \cdot \mathcal{L}_{ctrs} + \beta \cdot \mathcal{L}_{rec} + \gamma \cdot \mathcal{L}_{diff} \qquad (5)$$

where α, β, and γ are hyper-parameters that control the interaction of the total loss. We train our model by the minimizing $\mathcal{L}_{total}$ in an end-to-end manner. Once the model is learned, as shown in Figure 4, our model performs Re-ID by measuring the cosine similarity of features of query and gallery images.

Note that our ARN is able to perform Re-ID task in the unsupervised dataset by adapting discriminative ability from source to target domain.

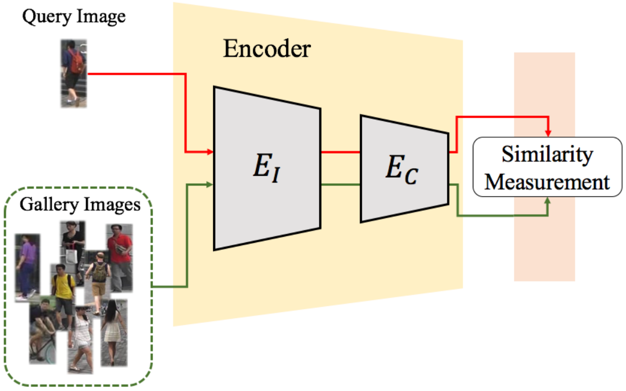

Fig. 4. When performing Re-ID using our ARN, only E_I and E_c in the latent encoder are required. That is, we match the latent feature e_c^q of person-of-interest (in the dataset) with the latent feature e_c^t of the test image by calculating the similarity ranking.

4. EXPERIMENTS

We now evaluate the performance of our proposed network, which is applied to perform cross-domain Re-ID tasks. To verify the work of each component in ARN, we provide ablation studies in Section 4.3. Furthermore, in Section 4.4, we compare the performance of our ARN with several supervised and unsupervised methods.

4.1. Datasets

To evaluate our proposed method, we conduct experiments on Market-1501 [14] and DukeMTMC-reID [15, 23], because both datasets are large-scale and commonly used. The details of the number of training samples under each camera are shown in Table. 1.

Market-1501 [14] is composed of 32,668 labeled images of 1,501 identities collected from 6 camera views. The dataset is split into two non-over-lapping fixed parts: 12,936 images from 751 identities for training and 19,732 images from 750 identities for testing. In testing, 3368 query images from 750 identities are used to retrieve the matching persons in the gallery.

DukeMTMC-reID [15, 23] is also a large-scale Re-ID dataset. It is collected from 8 cameras and contains 36,411 labeled images belonging to 1,404 identities. It also 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 evaluation on both datasets. In the experiments, there are two source-target settings:

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

Market-1501		DukeMTMC-reID	
camera	# of images	camera	# of images
1	2017	1	2809
2	1709	2	3009
3	2707	3	1088
4	920	4	1395
5	2338	5	1685
6	3245	6	3700
		7	1330
		8	1506

Table 1. Numbers of training samples and cameras in Market-1501 and DukeMTMC-reID datasets.

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

4.2. Implementation Details

ARN. Following Section 3, we use ResNet-50 pre-trained on ImageNet as our E_I model in the encoder. In order to perform the latent embedding easily for the modules E_T, E_C, and E_S, we remove the last few layers including average pooling from the pre-trained ResNet-50 model. The input of the E_I will be images with size $224 \times 224 \times 3$, denoting width, height, and channel respectively. In this manner, the output of E_I is the feature-map X with size $7 \times 7 \times 2048$ and will be fed into E_T, E_C, and E_S to obtain the corresponding feature with size $1 \times 1 \times 2048$, which is then flatten to private e_p or sharefji32l4d e_c latent feature with size 2048 as the final output of the encoder. Note that E_T, E_C, and E_S are implemented with fully convolution networks (FCNs) which contains three layers.

The input of our decoder D_c is the concatenated latent feature (e_c, e_p) with size 4096. We also implement the latent decoder D_c with fully convolution network. The output size of the decoder D_c is $7 \times 7 \times 2048$, which is identical to the input of E_T, E_C, and E_S modules. Note that the concatenated vectors in both domains, (e_c^t, e_p^t) and (e_c^s, e_p^s), are fed into the latent decoder simultaneously during the training procedure.

The classifier C_S contains only fully connected layers with dropout mechanism. We only feed the shared latent feature e_c^s into the classifier. The output is the classification result among the identities.

Learning procedure. As mentioned in Section 3, we aim to minimize the total loss $\mathcal{L}_{total}$ in Equation 5 during the training procedure. The parameters α, β, and γ are chosen under the experimental trials. In practice, we set α, β, and γ as 0.01, 2.0, and 1500, respectively. We aim to balance the larger value of $\mathcal{L}_{ctrs}$ and the smaller one of $\mathcal{L}_{diff}$. In addition, we need larger weight to enforce the reconstruction.

While we can directly use the same learning rate for each component to update the whole network, it might result in overfitting issues. We believe that individually setting the customized learning rates for E_I, E_T, E_C, E_S, D_C, and C_S can avoid this problem. For instance, when minimizing $\mathcal{L}_{class}$ and $\mathcal{L}_{ctrs}$, the weights of the pre-trained model E_I should not be updated faster than other modules because we try to keep much useful pre-trained weights ever trained on ImageNet. Hence, we set the learning rate for E_I to a relatively small value, 10^{-7}, and only tune E_I in the first few epochs. In addition, we set the learning rate of E_T, E_C, E_S, D_C to 10^{-3}, and C_S to 2×10^{-3}. We adopt the stochastic gradient descent (SGD) to update the parameters of the network.

Evaluating procedure. At the end of the learning scenario, we proceed to evaluate the performance of our trained network on Re-ID task. We only use E_I and E_c in the encoder for generating the latent features in evaluating scenario. For performance evaluation, we sort the cosine distance between the query and all the gallery features to obtain the final ranking result. The pedestrian retrieval samples are shown in Figure 5 and Figure 6. Note that the cosine distance is equivalent to Euclidean distance when the feature is L2-normalized. Moreover, we employ the standard metrics as in most person Re-ID literature, namely the cumulative matching curve (CMC) used for generating ranking accuracy, and the mean Average Precision (mAP).

4.3. Ablation Studies

In this subsection, we aim to fully analyze the effectiveness of our ARN via comparing with other baseline settings. As shown in Table 2, we compare our final version model with the ones removing supervised losses $\mathcal{L}_{ctrs}$, $\mathcal{L}_{class}$ in source domain or private components E_S, E_T. For dataset Market-1501 and DukeMTMC-reID, our full model can achieve 70.3% and 60.2% at Rank-1 accuracy, and 39.4% and 33.4% at mAP respectively.

Reconstruction loss $\mathcal{L}_{rec}$. For the target dataset on Market-1501 and DukeMTMC-reID, we observe that the Rank-1 accuracy of baseline model without $\mathcal{L}_{ctrs}$, $\mathcal{L}_{class}$, E_S, and E_T, containing only the reconstruction loss $\mathcal{L}_{rec}$, decrease by 25.8% and by 29% respectively. However, this shows that the reconstruction loss does play a great role in learning basic latent representation, which can still achieve 44.5% and 31.2% at Rank-1 accuracy. We note that without $\mathcal{L}_{ctrs}$, $\mathcal{L}_{class}$, we are not able to fine-tune E_I and let E_I keep its original pre-trained weights on ImageNet.

Source supervised losses $\mathcal{L}_{ctrs}, \mathcal{L}_{class}$. Refer to Table 2 again, we also observe that without supervised loss $\mathcal{L}_{ctrs}, \mathcal{L}_{class}$, the Rank-1 accuracy decrease by 18.1% and by 21.4% on Market-1501 and DukeMTMC-reID respectively. This obvious drop indicates that supervised metrics on source domain has largely improved the performance of our ARN model. We also conclude that the shared latent space does need the losses $\mathcal{L}_{ctrs}, \mathcal{L}_{class}$ to capture the semantics of person information.

Private modules E_T, E_S. In Table 2, without private modules E_T, E_S, the Rank-1 accuracy decrease by 9.8%

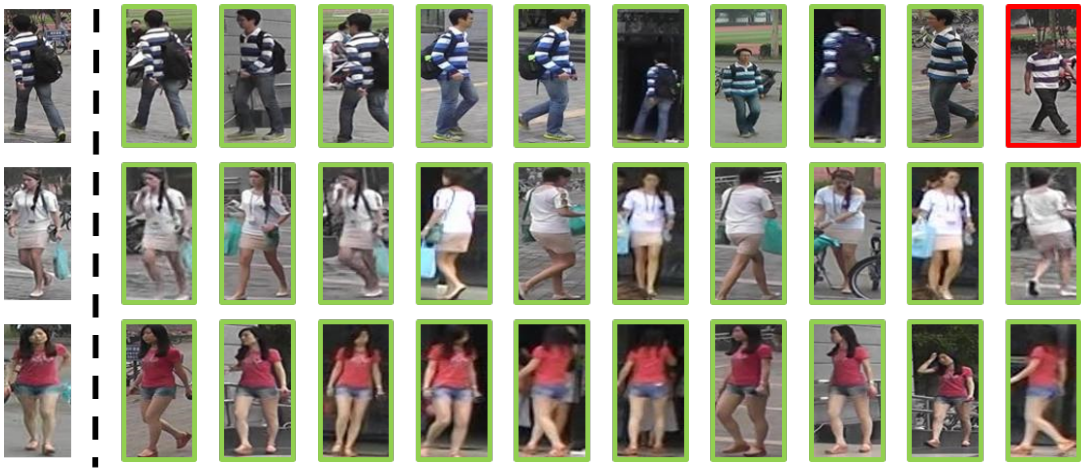

Fig. 5. Pedestrian retrieval samples on Market-1501 in the multi-shot setting using ARN. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right.

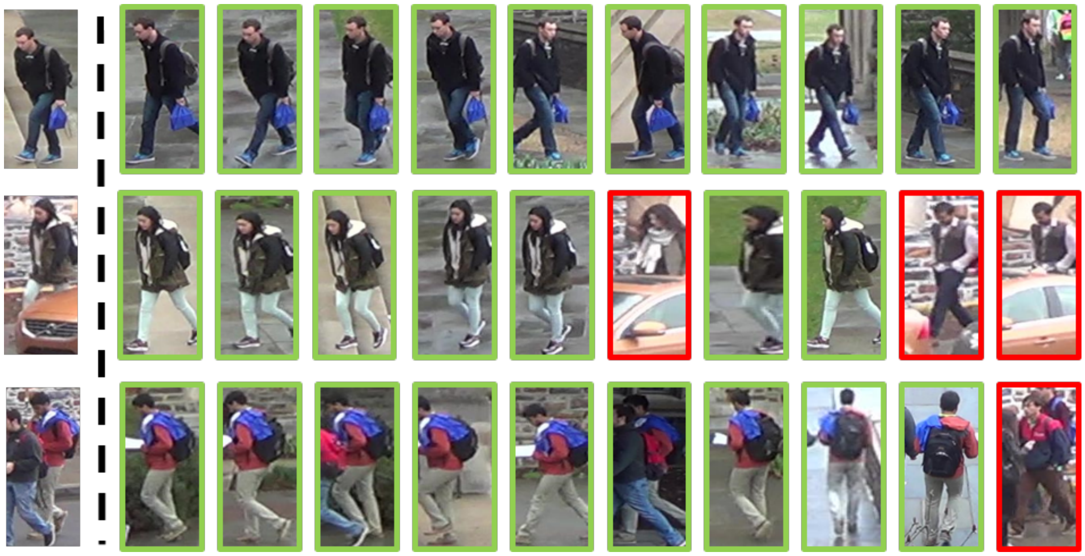

Fig. 6. Pedestrian retrieval samples on DukeMTMC-reID in the multi-shot setting using ARN. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right.

and by 11.8% on Market-1501 and DukeMTMC-reID respectively. We conclude that without partitioning the space to produce a private representation, the feature space may be contaminated with aspects of the noise that are unique for each dataset. Hence, having the private modules E_T, E_S does help perform representation learning in the shared latent space.

4.4. Comparison with State-of-the-art Methods

Market-1501. In Table 3, we first compare our model with the unsupervised methods. For the hand-crafted features based models, we compare our model with Bag-of-Word (BOW) [14]. For the cross-domain Re-ID models, there are Unsupervised Multi-task Dictionary Learning (UMDL) [10], Progressive Unsupervised Learning (PUL) [9], Clustering-based Asymmetric Metric Learning (CAMEL) [11] and Similarity Preserving Generative Adversarial Network (SPGAN) [13]. Our model outperforms these models in Rank-1, Rank-5, Rank-10, and mAP on Market-1501. Note that our model outperforms the second best method by 13.6% in Rank-1 accuracy and by 12.7% in mAP.

In addition, we also compare our model with existing supervised models, observing that our model surpasses BOW [14], LDNS [24] and already boost the performance closely to supervised deep learning based model like SVDNET [2], TriNet [5], CamStyle [6], or DuATM [7].

Method	Target: Market-1501 Source: DukeMTMC-reID					Target: DukeMTMC-reID Source: Market-1501				
	R1	R5	R10	R20	mAP	R1	R5	R10	R20	mAP
Ours w/o $\mathcal{L}_{ctrs}$, $\mathcal{L}_{class}$, E_S, E_T	44.5	63.2	70.4	78.5	20.3	31.2	42.5	50.1	57.4	18.4
Ours w/o $\mathcal{L}_{ctrs}$, $\mathcal{L}_{class}$	52.2	68.4	75.9	82.1	23.7	36.7	48.9	58.2	63.4	19.6
Ours w/o E_S, E_T	60.5	74.2	81.9	88.1	28.7	48.4	62.5	68.8	73.1	26.8
Ours	**70.3**	**80.4**	**86.3**	**93.6**	**39.4**	**60.2**	**73.9**	**79.5**	**82.5**	**33.4**

Table 2. Ablation studies of Adaptation and Re-Identification Network (ARN) under different experimental settings.

	Method	Rank-1	Rank-5	Rank-10	mAP
Supervised	BOW [14]	44.4	-	-	20.8
	LDNS [24]	61.0	-	-	35.7
	SVDNET [2]	82.3	-	-	62.1
	TriNet [5]	84.9	-	-	69.1
	CamStyle [6]	89.5	-	-	71.6
	DuATM [7]	91.4	-	-	76.6
Unsupervised	BOW [14]	35.8	52.4	60.3	14.8
	UMDL [10]	34.5	52.6	59.6	12.4
	PUL [9]	45.5	60.7	66.7	20.5
	CAMEL [11]	54.5	-	-	26.3
	SPGAN [13]	57.7	75.8	82.4	26.7
	Ours	**70.3**	**80.4**	**86.3**	**39.4**

Table 3. Performance comparisons on Market-1501 with supervised and unsupervised Re-ID methods.

	Method	Rank-1	Rank-5	Rank-10	mAP
Supervised	BOW [14]	25.1	-	-	12.2
	LOMO [25]	30.8	-	-	17.0
	TriNet [5]	72.4	-	-	53.5
	SVDNET [2]	76.7	-	-	56.8
	CamStyle [6]	78.3	-	-	57.6
	DuATM [7]	81.8	-	-	64.6
Unsupervised	BOW [14]	17.1	28.8	34.9	8.3
	UMDL [10]	18.5	31.4	37.6	7.3
	PUL [9]	30.0	43.4	48.5	16.4
	SPGAN [13]	46.4	62.3	68.0	26.2
	Ours	**60.2**	**73.9**	**79.5**	**33.4**

Table 4. Performance comparisons on DukeMTMC-reID with supervised and unsupervised Re-ID methods.

DukeMTMC-reID. In Table 4, our model outperforms unsupervised methods such as BOW [14], UMDL [10], PUL [9], and SPGAN [13]. Our model achieves **Rank-1 accuracy=60.2% and mAP=33.4%** and outperforms the second best method [13] roughly by 13.8% in Rank-1 accuracy and by 7.2% in mAP. More importantly, the performance of our model is better than some supervised methods such as BOW [14] and LOMO [25].

5. CONCLUSIONS

In this paper, we presented a deep learning model of Adaptation and Re-Identification Network (ARN) for solving cross-domain Re-ID tasks. Our ARN allows us to jointly exploit a pre-collected supervised source-domain dataset and a target-domain dataset of interest by learning domain invariant and discriminative features. As a result, Re-ID in the target-domain can be performed even without any label information observed during training. With this proposed unsupervised domain adaptation network, we conducted experiments on Market-1501 and DukeMTMC-reID datasets, and confirmed the effectiveness of our model in such a challenging unsupervised learning setting. Moreover, our method also performed favorably against a number of baseline supervised Re-ID approaches, which again supports the use of our ARN for practical Re-ID tasks.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Science and Technology of Taiwan under grant MOST 107-2634-F-002-010.

REFERENCES

[1] Liang Zheng, Yi Yang, and Alexander G Hauptmann, “Person re-identification: Past, present and future,” in *arXiv preprint*, 2016.

[2] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, “Svdnet for pedestrian retrieval,” in *arXiv preprint*, 2017.

[3] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang, “Improving person re-identification by attribute and identity learning,” in *arXiv preprint*, 2017.

[5] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," in *arXiv preprint*, 2017.

[6] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong, "Towards unsupervised open-set person re-identification," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016.

[9] Hehe Fan, Liang Zheng, and Yi Yang, "Unsupervised person re-identification: Clustering and fine-tuning," in *arXiv preprint*, 2017.

[10] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1306–1315.

[11] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[13] 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 (CVPR)*, 2018.

[14] 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 (ICCV)*, 2015.

[15] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian, "Deep transfer learning for person re-identification," in *arXiv preprint arXiv:1611.05244*, 2016.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[18] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint*, 2014.

[19] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 343–351.

[20] Ming-Yu Liu and Oncel Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[22] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006, vol. 2, pp. 1735–1742.

[23] 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 (ECCV) Workshop*, 2016.

[24] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2197–2206.