Disentangling Reconstruction Network for Unsupervised Cross-Domain Person Re-Identification

Student Name: Harsh Kumar Jain
Roll Number: IIIT-D-MTech-CS-20-MT18006
June, 2020

Indraprastha Institute of Information Technology
New Delhi

Thesis Advisor
Dr. A. V. Subramanyam

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science & Engineering

Certificate

This is to certify that the thesis titled **"Disentangling Reconstruction Network for Unsupervised Cross-Domain Person Re-Identification"** submitted by **Harsh Kumar Jain** for the partial fulfillment of the requirements for the degree of Master of Technology in Computer Science & Engineering is a record of the bonafide work carried out by him under my guidance and supervision at Indraprastha Institute of Information Technology, New Delhi. This work has not been submitted anywhere else for the reward of any other degree.

.....

Dr. A. V. Subramanyam

Indraprastha Institute of Information Technology, New Delhi

Abstract

Unsupervised cross-domain Person Re-Identification (Re-ID) severely suffers from the domain gap. While different works address this issue, bridging domain gap with high-level representation is hard as it comprises of entangled information including identity, pose, illumination, and other domain-specific variations. In this work, we propose a disentangled reconstruction method to address the domain-shift problem for Re-ID in an unsupervised manner. To this end, we have two major contributions. First, we propose to disentangle identity-related and non-identity related features from person images. We also reconstruct the disentangled features using a decoding layer to increase the generalization capability of identity features. Second, in the target domain, we explicitly consider the camera style transfer images as a data augmentation to address intra-domain discrepancy and to learn the camera invariant features from the target domain. We demonstrate that the auxiliary tasks of disentanglement and reconstruction are helpful to improve the generalization capability of the model and perform cross Re-ID on unlabeled target domain data. Experimental results on the challenging benchmarks of Market-1501 and DukeMTMC-reID demonstrate that our proposed method achieves competitive performance.

Acknowledgment

I would like to extend my deepest gratitude towards my advisor Dr. A. V. Subramanyam for his guidance at all levels, support, patience, and immense knowledge. This work would not have been possible without his experience and motivation he provided to me. I would also like to thank Kajal Kansal for her support and collaboration in this work. And finally, I would like to thank my family who encouraged me and kept me motivated throughout the course of this project.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Person Re-Identification (Re-ID) aims to match the person identities from non-overlapping multi-camera networks [12]. Many existing Re-ID methods adopt a supervised learning approach, which assumes the availability of a large number of manually labeled data [42]. This assumption inherently limits the scalability of Re-ID models in practical deployments. Additionally, the problem of domain shift remains a big challenge. The primary cause of insufficient cross-domain generalization lies in the distribution discrepancy between different domains and the unavailability of label information in the target domain.

Unsupervised Re-ID, on the other hand, has witnessed a surge in the past few years. This is because such a system is easily scalable in a real-world scenario, unlike the supervised system. However, unsupervised Re-ID also poses daunting challenges in terms of domain gap between a labeled source domain and an unlabeled target domain. Unsupervised Domain Adaptation (UDA) methods, such as [25,29], try to reduce the discrepancy between source and target domain directly and completely ignore the camera variation of the target domain. Several Generative Adversarial Networks (GANs) [1,6,15,28,33,37] based works translate the appearance of images from the source domain to the target domain by preserving the annotation information of the source domain. However, the main focus is on increasing the training samples, and it costs more time and computation complexity with difficult convergence rates. Discriminative feature learning for cross-dataset Re-ID has also been exploited in [26, 33].

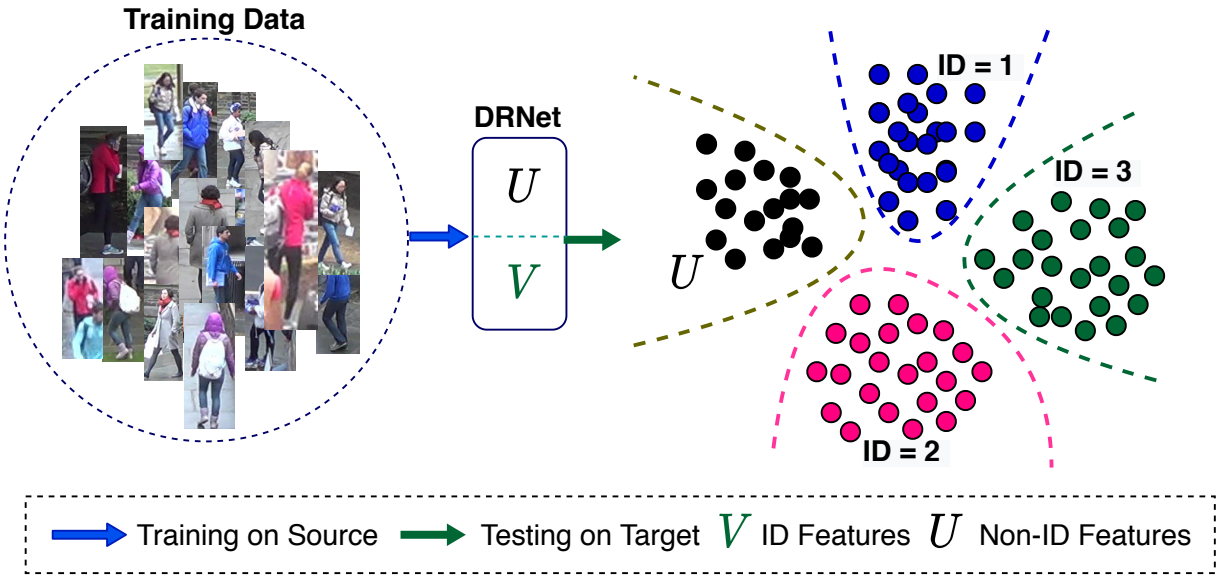


Figure 1.1: Motivation illustration. The domain variations can be quite diverse. Our approach learns to decompose the identity and non-identity features and transfer it so that the target model can only focus on identity related discriminative or generalizable features.

1.2 Contribution

In this work, we aim to utilize the discriminative identity information and transfer the learned knowledge on the target domain for more effective cross-domain Re-ID model learning. Our motivation to address the domain bias problem is illustrated in Figure 1.1. Our contributions are as follows:

• We address the problem of cross-domain Re-ID by exploiting multi-task learning. We design the network with an auxiliary task of disentanglement and reconstruction to enhance the generalization capability.

• To achieve the goal of discriminative and camera invariant features, we design two components in our network: disentangled reconstruction on the source domain and camstyle invariance on the target domain. The former component preserves the unique person cues specific to identity and improves the source Re-ID model's generalization capability. On the later component, we apply a camstyle transfer mechanism as a data augmentation on the target domain.

• We demonstrate the effectiveness of our proposed method on challenging cross-domain Re-ID datasets Market-1501 [46] and DukeMTMC-reID [47]. We perform an extensive ablation study and demonstrate the effectiveness of all the components of the network.

Chapter 2

Related Works

2.1 Supervised Person Re-ID

[2, 3, 34–36] has been widely studied in the literature. Classical supervised methods mainly address the challenges of viewpoint variations, background clutter, or occlusion. Zhang et al. [45] explore a feature matching method to align different body parts, and retrieval accuracy has already surpassed the accuracy of human eyes. However, the supervised approaches cannot be applied for cross-domain Re-ID due to the lack of generalization capability. To alleviate the problem of domain bias, we focus on unsupervised cross-domain person Re-ID in this work.

2.2 Unsupervised Person Re-ID

Prior unsupervised Re-ID works address the problem of the domain gap by reducing the discrepancy between the source domain and target domain. They project the source and target data in shared space to learn domain invariant representation, or purely in an unsupervised manner without using any labels. There are three major categories for existing methods: Pseudo-label based methods, domain-translation methods, and unsupervised methods.

2.3 Pseudo-label based method

Pseudo-label based methods [8–10,43,44] propose to learn target-domain features with generated pseudo labels. Fan *et al.* [8] assign hard clustering labels for unlabelled data to learn target-domain features by self-training. However, they ignore the inevitable label noise caused by the

clustering. Ge *et al*. [10] explores such clustering-based methods by introducing local features and mutual mean-teaching scheme to learn better target domain features. Yu *et al*. [43] focuses on learning the target domain features with soft labels. In [17], Kumar *et al*. observe that single domain-based adaptation does not take care of the heterogeneity of the environment and is insufficient for domain adaptation. They propose multi-source based domain adaptation using k-reciprocal tracklet clustering. However, this method assumes that intra-camera labels are available for tracklet generation. Ren *et al*. [28] perform style adaptation using StarGAN [5] and then assign soft labels to the translated images using KNN. Qi *et al*. [27] present a camera aware domain adaptation approach using re-ranking [48], online triplets and cross camera adaptation. PAUL [41] proposes PatchNet to learn the similarity between patches of similar images in the same domain. In order to obtain positive samples in the unlabeled target domain, a random transformation is applied.

2.4 Domain-translation based method

Domain-translation based methods [4, 6, 16, 37] adopt image-to-image translation models, and the translated images are utilized for training Re-ID model in a supervised way. Chen *et al*. [4] introduces CR-GAN to synthesize images by augmentation approach where each source pedestrian is augmented with various contextual images from the target domain. PTGAN [37] employs CycleGAN and maintain color consistency during the domain translation by pixel-level constraints. SPGAN [6] minimizes the feature-level similarities between translated images and the original ones. Huang *et al*. [16] observes that drastic background changes across domains introduce difficulty in extracting robust features. They propose SBSGAN to suppress such a background shift. However, such methods deeply rely on image generation quality.

2.5 Unsupervised methods

Since pseudo-label based methods require a large annotated dataset as the source, unsupervised methods overcome such requirement by completely learning without any labels. Wang and Zhang [32] propose an iterative memory-based positive label prediction and multi-classification loss. TSSL [39] learns similarity between tracklets using triplet loss [14] while applying neighborhood compactness and cluster structure.

2.6 Disentanglement methods

Disentangling approaches have been widely used in many computer vision tasks. Prior face tasks based work propose to disentangle the representations in pose-invariant recognition [31] and identity-preserving image editing [18]. Liu *et al.* [23] propose to learn disentangled face identification in addition to face features. Researchers have also explored disentanglement approaches for the person Re-ID task. Eom *et al.* [7] introduces a GAN based network, IS-GAN, to disentangle identity and non-identity related features for supervised Re-ID. Li *et al.* [19] introduces PDA-Net to learn deep image representation with disentangled pose and domain information for cross-domain Re-ID.

Inspired from the above methods, we choose to disentangle useful person cues and learn discriminative identity features for unsupervised cross-domain Re-ID. We demonstrate that domain bias in Re-ID can be successfully addressed via a disentanglement approach, even with unlabelled target-domain data.

2.7 Camstyle Invariance Learning

Zhong *et al.* [52] introduces a data augmentation approach to smooth camstyle disparities and avoid overfitting for supervised Re-ID models. HHL [49] introduces a Hetero-Homogeneous Learning method to learn camera-invariant features with camera style transferred images to improve the generalization ability of Re-ID models. Zhong *et al.* [51] proposes a camstyle adaptation model to smooth style disparities across the cameras. [50] explores camstyle invariance through an exemplar memory for domain adaptive person Re-ID. Inspired by the promising performance, we apply it to the target domain as a data augmentation approach.

Chapter 3

Proposed Algorithm

In this section, we describe our proposed network. The overall network architecture is illustrated in Figure 3.1. Our model comprises of a baseline network to learn the representations corresponding to inputs. The model has three components – baseline, identity-preserving and disentangled reconstruction module. The identity-preserving and disentanglement components preserve useful Re-ID cues of person. We apply reconstruction along with disentanglement to learn the generalizable identity features. While training on target, we perform fine-tuning and camstyle-transfer. The goal of the camstyle-transfer component is to learn the camera invariant features due to domain shift. We discuss the proposed approach in detail in the following subsections.

3.1 Notations

We first define basic notations followed in this work. The source dataset is defined as $\mathbb{S} = \{I^{\mathbb{S}}, y_{id}^{\mathbb{S}}\}^{N_{\mathbb{S}}}$ with $N_{\mathbb{S}}$ number of images, identity labels $\mathbb{Y}^{\mathbb{S}} = \{y_{id}^{\mathbb{S}}\}^{N_{\mathbb{S}}}$. The target dataset is defined as $\mathbb{T} = \{I^{\mathbb{T}}\}^{N_{\mathbb{T}}}$ with $N_{\mathbb{T}}$ number of total images. The target dataset do not have ground truth identity labels.

3.2 Source Representation Model

3.2.1 Baseline network

The baseline network, denoted as $B(.;\theta)$, consists of a deep convolutional neural network based on ResNet-152 [13] which is pretrained on ImageNet. θ denotes the network parameters. We learn θ through identification loss [30] represented as $\mathcal{L}_B$,

$$\mathcal{L}_B = -\frac{1}{N}\sum_{j=1}^{N} \log p(\hat{y}^{\mathbb{S}}_{id_j} | B(I^{\mathbb{S}}_j;\theta)\,, \tag{3.1}$$

where N is the number of images, $I^{\mathbb{S}}_j$ denotes a sample, $B(I^{\mathbb{S}}_j;\theta)$ denotes encoded representation and $p(.)$ denotes the predicted probability.

3.2.2 Disentangled-Reconstruction module

To learn the discriminative features and generalize the Re-ID model to the unlabeled target domain, we propose a disentangled reconstruction module on source domain $\mathbb{S}$ where learned representations can be disentangled into identity and non-identity related features. Further, the disentangled features can undergo reconstruction using a decoding layer.

Disentanglement: We augment the baseline network B with the disentanglement network, denoted as $D(.;\phi)$. Here $\phi = \{\phi_{ID}, \phi_{NID}\}$ where ϕ_{ID} are the parameters of the network which extracts identity related features, and ϕ_{NID} denotes parameters of the network which extracts non-identity related features. The output features, $X = B(I^{\mathbb{S}};\theta)$, extracted from the backbone are fed to disentanglement model D. Here, $X \in \mathcal{R}^{d\times N}$, where d is the feature dimensionality. It has two branches: identity (ID) branch and non-identity (NID) branch, which performs the disentanglement between identity and non-identity information such as pose and camera, respectively. The non-identity related features are denoted by $U = D(X;\phi_{NID})$ and the identity related features are denoted by $V = D(X;\phi_{ID})$. Both $U, V \in \mathcal{R}^{d\times N}$. The disentanglement loss is defined as,

$$\mathcal{L}_{dis} = \frac{1}{N}\sum_{j=1}^{N} \|X_j - (U_j \odot V_j)\|_2^2 \tag{3.2}$$

where $\|.\|_2^2$ denotes L2 norm, X_j denotes the j^{th} column of X and $\odot$ represents Hadamard product.

Reconstruction: We apply Mean-Squared Error (MSE) reconstruction loss between X and its reconstructed version $\hat{X}$. Let the trainable parameters used to obtain $\hat{X}$ be ψ. The reconstruction loss on the source representation model can be given as,

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{j=1}^{N} ||X_j - \hat{X}_j||_2^2 \tag{3.3}$$

where X is the baseline features and $\hat{X}$ is the reconstructed features.

3.2.3 Identity-preserving module

The goal of this module is to improve the identity discriminative features (V) and enhance the generalization capability on the source domain. Under the disentanglement decomposition, we take additional benefit from the identification loss, $\mathcal{L}_{vid}$ given in eq. 3.4. Since $D(.;\phi_{ID})$ must capture information relevant to person identity, we use the identification loss to supervise the training process on source domain. The identification loss is given by,

$$\mathcal{L}_{vid} = -\frac{1}{N}\sum_{j=1}^{N} \log p(\hat{y}^{\mathbb{S}}_{id_j}|D(X_j;\phi_{ID}))\,, \tag{3.4}$$

We explain the target representation model in the next section.

3.3 Target Representation Model

Once source domain training is completed, the learned knowledge is transferred to the target domain via fine-tuning.

3.3.1 Fine-tuning

In case of target representation model, we first initialize the model with trained weights of source model. Then, we apply the following identification loss to fine-tune the model,

$$\mathcal{L}_{id} = -\frac{1}{C}\sum_{j=1}^{C} \log p(\hat{y}^{\mathbb{T}}_{id_j}|D(B(I^{\mathbb{T}}_j;\theta);\phi_{ID}))\,, \tag{3.5}$$

where C is the number of classes and p is the predicted probability. Here we assume that each image has a unique label and there are as many classes as images. Though such assumption

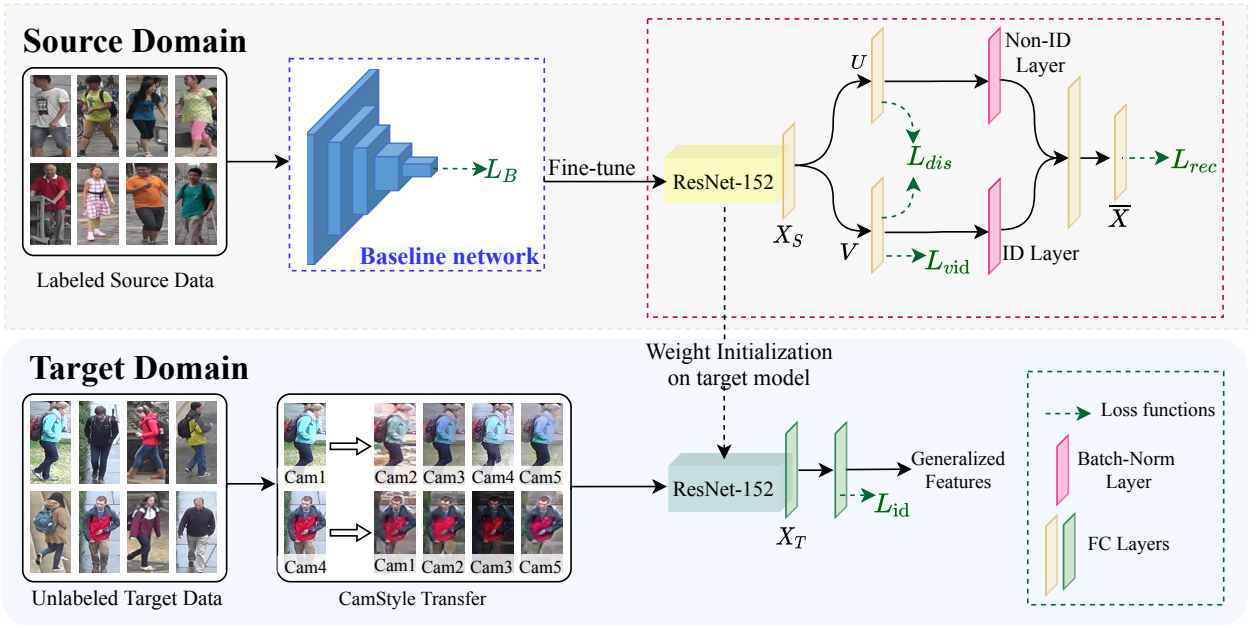


Figure 3.1: Proposed Architecture. On the source domain, labeled source data is given as an input to the baseline network and trained using $\mathcal{L}_B$. The baseline features are decomposed into U and V using disentanglement loss $\mathcal{L}_{dis}$ and classification loss L_{vid}. Batch-normalized U and V are concatenated for reconstruction of X using reconstruction loss $\mathcal{L}_{rec}$. On target domain, unlabelled target data and cam-style transferred images are fed as an input to ResNet-152, whose weight initialization is done through the trained source model. The model is further fine-tuned through $\mathcal{L}_{id}$.

provides weak supervision, [50] shows that it helps the network to learn the representation of a person.

3.3.2 Camstyle Invariance Learning

One of the most challenging problems during domain transfer in Re-ID is the variation between source cameras and target cameras. The model trained on the source domain suffers because of the variance in the target cameras. To overcome this problem, we use [50] to learn the camera invariance property. For each target image, the camera style transferred image for each camera in the target dataset is generated while preserving the identity information. The generated images capture the camera invariance by transferring the appearance and illumination related information. Further, original and camera-style images form the augmented training set can be fed to the target representation model. Once the domain-invariant feature is learned via a fine-tuned model, we can perform cross-domain Re-ID by matching the query image against gallery images.

3.4 Training strategy

We train the model in three steps: fine-tune the backbone network followed by source representation model training and fine-tuning on target.

Fine-tuning the backbone: We first fine-tune the backbone network of ResNet-152 using the baseline loss ($\mathcal{L}_B$) as given in Section 3.2.1.

Source representation model training: Here, we apply identification ($\mathcal{L}_{vid}$), disentanglement ($\mathcal{L}_{dis}$) and reconstruction ($\mathcal{L}_{rec}$) loss to optimize the source network. The losses are given in eq. 3.2, 3.3 and 3.4, respectively.

Target fine-tuning: On target representation model, we apply identification loss ($\mathcal{L}_{id}$) which is given in eq. 3.5.

ALGORITHM 1: Optimization of the Model

Input: $\mathbb{S}$, $\mathbb{T}$, $B(.;\theta)$
Output: Network parameters $\{\theta\}$
repeat
  for *each epoch* **do**
    Randomly sample PK images from $\mathbb{S}$;
    $\theta \leftarrow \theta - \frac{\partial \mathcal{L}_B}{\partial \theta}$;
    $\{\theta, \phi_{ID}\} \leftarrow \{\theta, \phi_{ID}\} - \frac{\partial \mathcal{L}_{vid}}{\partial \{\theta, \phi_{ID}\}}$;
    $\{\theta, \phi\} \leftarrow \{\theta, \phi\} - \frac{\partial \mathcal{L}_{dis}}{\partial \{\theta, \phi\}}$;
    $\{\theta, \phi, \psi\} \leftarrow \{\theta, \phi, \psi\} - \frac{\partial \mathcal{L}_{rec}}{\partial \{\theta, \phi, \psi\}}$;
  end
until *Convergence*;
Return $\{\theta, \phi\}$
Initialize target network with $\{\theta, \phi_{ID}\}$;
repeat
  for *each epoch* **do**
    Randomly sample N images from $\mathbb{T}$ and cam-style transferred images;
    $\{\theta, \phi_{ID}\} \leftarrow \{\theta, \phi_{ID}\} - \frac{\partial \mathcal{L}_{id}}{\partial \{\theta, \phi_{ID}\}}$;
  end
until *Convergence*;
Return $\{\theta\}$

3.5 Testing

Once the source model is trained, we test on the target dataset to perform cross-domain Re-ID. We first extract the features of the probe and gallery image using $B(.;\theta)$. We then compute the Euclidean distance between the query and gallery features to measure the similarity.

Chapter 4

Experiments

In this section, we describe the datasets and the evaluation protocol. Further, we discuss the ablation study to evaluate the contribution of each component of the proposed method. Furthermore, we compare our proposed method with state-of-the-art methods.

4.1 Datasets

4.1.1 Market-1501

Market-1501 [46] is composed of 32,668 labeled images of 1,501 identities collected from 6 camera views. The dataset is split into two non-overlapping 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.

4.1.2 DukeMTMC-reID

The DukeMTMC-reID [47] 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.

4.2 Evaluation Protocol

We employ the standard metrics, namely the Cumulative Matching Curve (CMC) used for generating rank accuracy, and the mean Average Precision (mAP). We report CMC-1, CMC-5, CMC-10 accuracy, and mAP for evaluation on both datasets.

4.3 Implementation details

The model is implemented on PyTorch with NVIDIA GeForce RTX 2080Ti. Data augmentation includes random cropping, horizontal flipping, pixel value normalization, and random erasing. We randomly sample P classes (person identities) and sample K images of each class (person), resulting in a batch of PK input images.

4.3.1 Backbone

: We use the ResNet-152 [13] architecture as backbone network. We fine-tune the baseline network using identification loss on the source Re-ID dataset. We discard the last fully connected layer and add two fully connected layers. The output of the last fully connected layer has 512 units. The number of epochs is set to 60, and the mini-batch size is 64. Dropout is set to 0.5. We use Adam optimizer for optimizing the overall objective function. The initial learning rate is set to 5×10^{-3} for the ResNet-152 and 5×10^{-2} for the other layers in the baseline. It is decreased to 0.1 of its previous value every 40 epochs. The best model is saved and used as the backbone for all further experiments.

4.3.2 Source representation model training:

It is trained with three different loss functions. The number of epochs is set to 1000 and mini-batch size is 64 (PK style sampling is used with $P = 16$ and $K = 4$). The initial learning rate for updating the backbone weights is set to 5×10^{-4}, whereas other layers of the DRNet are updated at a learning rate of 5×10^{-2}. The learning rate is decreased to 10% of its previous value after every 200 epoch. We use Stochastic Gradient Descent (SGD) optimizer.

4.3.3 Fine-tuning on target

: The model is then trained for 10 epochs with a mini-batch size of 64. We use the SGD optimizer for model optimization of the network. The learning rate of the backbone is reduced to 1×10^{-4}, and other components learn at a rate of 1×10^{-2}. The learning rate is reduced to 0.1 of its previous value after every epoch. After the fine-tuning, the best model is used for Re-ID evaluation.

4.4 Ablation Study

Table 4.1: Ablation Study on Market-1501-to-DukeMTMC-reID dataset. CMC-1, CMC-5, CMC-10 (%) and mAP (%) are reported. × refers to the loss function is not used. ✓represents the applied loss function.

	Source				**Target**					
Variant	$\mathcal{L}_B$	$\mathcal{L}_{vid}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{id}$	Cam-style	CMC-1	CMC-5	CMC-10	mAP
①	✓	×	×	×	×	×	38.50	54.80	61.08	21.01
②	✓	×	×	×	✓	×	42.41	58.79	64.76	24.12
③	✓	×	×	×	✓	✓	60.45	73.24	77.87	38.54
④	✓	✓	×	×	✓	✓	60.59	73.47	78.05	38.63
⑤	✓	✓	✓	×	✓	✓	62.00	75.25	79.82	39.52
⑥	✓	✓	✓	✓	✓	✓	**63.01**	**76.21**	**81.04**	**40.84**

In this subsection, we evaluate the proposed network under different variants. We use Market-1501 [46] as source dataset and DukeMTMC-reID [47] as target dataset to do the ablation study. We compare all the different variants in Table 4.1.

Variant ①: We first experiment with the pre-trained network of ResNet-152 [13] as the backbone to train on source dataset of Market-1501 and test it on target dataset of DukeMTMC-reID. The network is trained via loss $\mathcal{L}_{\mathcal{B}}$. We observe a CMC-1 accuracy of 38.50%. The poor generalization capability of the baseline model leads to poor performance.

Variant ②: Here we keep the Variant ① setting at source dataset and fine-tune the target model using identification loss $\mathcal{L}_{id}$. We see a boost of 3.91% in the CMC-1 accuracy.

Variant ③: In this variant, we keep the the Variant ② setting. Additionally, we generate camera-style images from each camera and use them for data augmentation. We fine-tune the target model with original and camera-generated images on the target dataset. Here, we achieve 60.45% in the CMC-1 accuracy.

Variant ④: We add another identification loss $\mathcal{L}_{vid}$. By keeping the Variant ③ setting of target

representation model, we observe a boost of 0.14% in the CMC-1 accuracy.

Variant ⑤: We add reconstruction loss $\mathcal{L}_{rec}$ in this variant. We observe a CMC-1 accuracy of 62.00%. This shows that the generalization capability of the model leads to better identity discriminative representation learning.

Variant ⑥: We add disentanglement loss $\mathcal{L}_{dis}$, which attempts to decompose shared features X into U and V. This variant is known as DRNet. We observe a CMC-1 accuracy of 63.01%. This helps to align source domain to target domain and thus further helps to extract generalized features.

4.5 Comparison with State-of-the-art

We compare our proposed method with the state-of-the-art unsupervised person Re-ID methods on Market-1501 [46] and DukeMTMC-reID [47]. We report the experimental results in Table 4.2. We obtain competitive results compared with the state-of-the-art.

Table 4.2: Comparison with the state-of-the-art methods on DukeMTMC-reID-to-Market-1501 and Market-1501-to-DukeMTMC-reID. CMC-1, CMC-5, CMC-10 (%) and mAP (%) are reported. '–' means that the results are unavailable, '*' means intra-camera labels are used and '†' indicates fully unsupervised methods.

Approach	**Reference**	**DukeMTMC-reID-to-Market-1501**				**Market-1501-to-DukeMTMC-reID**			
		CMC-1	CMC-5	CMC-10	mAP	CMC-1	CMC-5	CMC-10	mAP
MMFA [20]	BMVC 2018	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
PTGAN [37]	CVPR 2018	38.6	–	66.1	–	27.4	–	50.7	–
SPGAN [6]	CVPR 2018	57.7	–	26.2	–	57.7	–	–	26.7
TJ-AIDL [33]	CVPR 2018	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
HHL [49]	ECCV 2018	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
UDAA [53]	ICIP 2019	43.0	63.3	70.6	17.1	28.6	44.2	51.7	13.1
CCE [27]	ICCV 2019	64.3	–	–	34.5	55.4	–	–	36.7
CASCL [38]	ICCV 2019	64.7	–	–	35.6	51.5	–	–	30.5
EUG [40]	TIP 2019	49.8	66.4	72.7	22.5	45.2	59.2	63.4	24.5
ATNet [22]	CVPR 2019	55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
PAUL	CVPR 2019	66.7	–	–	36.8	56.1	–	–	35.7
TSSL†1001[39]	AAAI 2020	71.2	–	–	43.3	62.2	–	–	38.5
SDA [11]	Arxiv 2020	49.5	62.2	67.7	23.2	54.4	66.4	71.3	25.6
CSGLP [28]	TIFS 2020	61.2	77.5	83.2	31.5	47.8	62.3	68.3	27.1
ktCUDA* [17]	WACV 2020	68.6	–	–	49.4	58.7	–	–	40.9
SSL [21]	CVPR 2020	71.7	83.8	87.4	37.8	52.5	63.5	68.9	28.6
DRNet	**Ours**	**72.0**	**83.9**	**88.6**	**39.8**	**63.0**	**76.2**	**81.0**	**40.8**

4.5.1 Results on DukeMTMC-reID-to-Market-1501

To verify that our method is able to extract generalized features across domains, we compare it with recent methods like CSGLP [28], SDA [11], ATNet [22] and EUG [40]. We also compare to the state-of-the-art pseudo-label based and domain translation based algorithms. When tested on DukeMTMC-reID, the proposed method achieves 72.0% CMC-1 and 39.8% mAP score. The comparison indicates the effectiveness of the proposed method to learn more generalizable features for bridging domain gaps.

4.5.2 Results on Market-1501-to-DukeMTMC-reID

We achieve CMC-1 accuracy of 63.0% and 40.8% mAP score for this case. Our results state that our proposed method achieves significant performance compared to the existing literature.

4.5.3 Other Results

In Table 4.3, we show the improvement of CMC-1 (%) on the source datasets for different variants. It shows that the objective functions help to improve the generalization capability of the source dataset.

Table 4.3: Results of proposed model on the source dataset. Variants are as described in Table 4.1

Variant	**Objective Function**				**Market-1501**
	$\mathcal{L}_B$	$\mathcal{L}_{vid}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{dis}$	CMC-1
①	✓	×	×	×	89.99
④	✓	✓	×	×	90.26
⑤	✓	✓	✓	×	90.43
⑥	✓	✓	✓	✓	90.79

In Table 4.4, we show the results obtained by extracting features from different layers of DRNet model on target domain. Since, we obtain better results from X layer instead of V, we have used the features from X for all the experiments stated earlier.

In Table 4.5, we compare the results using two strategies. The DRNet model is as explained in the previous sections. The DRNet* model uses the intra-camera labels in the target domain. We use the camera labels to generate labels along with tracking details. Hence, each person in a single camera can be classified into a single identity. However, the tracking details are not available across different cameras.

Table 4.4: Comparison of results by extracting features from different layers on Market-1501-to-DukeMTMC-reID dataset. CMC-1 (%) is reported. × refers to the loss function is not used. ✓represents the applied loss function.

Variant	Source				Target		CMC-1	
	$\mathcal{L}_B$	$\mathcal{L}_{vid}$	$\mathcal{L}_{rec}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{id}$	Cam-style	*LayerX*	*LayerV*
①	✓	×	×	×	×	×	38.50	–
②	✓	×	×	×	✓	×	42.41	40.20
③	✓	×	×	×	✓	✓	60.45	58.57
④	✓	✓	×	×	✓	✓	60.59	58.91
⑤	✓	✓	✓	×	✓	✓	62.00	60.19
⑥	✓	✓	✓	✓	✓	✓	63.01	61.54

Table 4.5: Proposed results using different methodologies ('*' refers the use of intra-camera labels in target domain)

Approach	**DukeMTMC-reID-to-Market-1501**				**Market-1501-to-DukeMTMC-reID**			
	CMC-1	CMC-5	CMC-10	mAP	CMC-1	CMC-5	CMC-10	mAP
DRNet	72.0	83.9	88.6	39.8	63.0	76.2	81.0	40.8
DRNet*	77.1	87.5	92.4	44.6	67.4	80.1	86.9	47.2

4.6 Visualization Results

We show visualization results from our proposed network in Figure 4.1 and 4.2 and compare it against the baseline network to show the superiority of our model.

4.6.1 Retrieval Results

We show retrieval results in Figures 4.1 using the baseline and the proposed network on the same query images. Here, we randomly pick images from the probe set as a query and retrieve the corresponding images from the gallery set. The images in the first column are the query images. The retrieved images are sorted according to the similarity scores from left to right (from the second column till last). The green boundary indicates a positive match, and red shows a negative match. We can see that the proposed network gets accurate shots under most of the challenging situations, as shown in Figure 4.1.

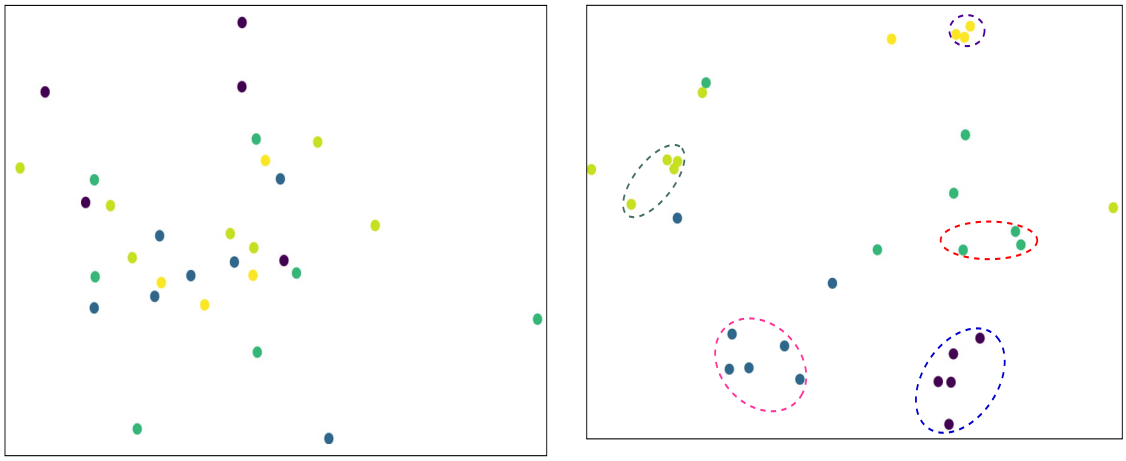
4.6.2 Visualization of Person Features through t-SNE

We randomly select 5 identities from the testing set and visualize the features using t-SNE [24] in Figure 4.2. We use the same color for the features corresponding to the same identity. We show t-SNE for baseline features and features from the proposed approach in Figure 4.2.

(a) Top-10 retrieval results from baseline

(b) Top-10 retrieval results from the proposed network

Figure 4.1: Top-10 retrieval results. The images in the first column are the query images. The retrieved images are sorted according to the similarity scores from left to right (from second column till last). Red boundary indicates a negative match and green shows a positive match.



(a) Baseline (b) Proposed

Figure 4.2: **2D visualization of the features from baseline and proposed network.** Same color refers to same identity and different color represent different identity.

Baseline: In Figure 4.2 (a), we observe that the features obtained from the trained baseline are merged into samples of another identity. Hence, the baseline features are not robust enough to perform Re-ID.

Proposed Network: On the other hand, features from our proposed network show that the individual identity clusters form very well, as in Figure 4.2 (b). For example, the intra-distance between purple, green, yellow, and blue samples has reduced drastically. This shows the superiority of our method over the baseline. Thus, we conclude that the network has a strong capability to extract generalizable features and perform cross-domain Re-ID to bridge the domain gap accurately.

Chapter 5

Conclusion

In this work, we propose a novel method for unsupervised cross-domain person Re-ID tasks. We first disentangle the features into ID and non-ID related features using disentangling and reconstruction loss. Further, to adapt to the target domain, we use camstyle transferred images. We perform extensive ablation study and demonstrate the significant improvement achieved due to the proposed components. Experimental results show that our proposed method can learn better features to address domain shifts in the unsupervised person Re-ID problem.

Bibliography

[1] Bak, S., Carr, P., and Lalonde, J.-F. Domain adaptation through synthesis for unsupervised person re-identification. In *ECCV* (2018), pp. 189–205.

[2] Bhuiyan, A., Liu, Y., Siva, P., Javan, M., Ayed, I. B., and Granger, E. Pose guided gated fusion for person re-identification. In *WACV* (2020), pp. 2675–2684.

[3] Chen, H., Lagadec, B., and Bremond, F. Learning discriminative and generalizable representations by spatial-channel partition for person re-identification. In *WACV* (2020), pp. 2483–2492.

[4] Chen, Y., Zhu, X., and Gong, S. Instance-guided context rendering for cross-domain person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 232–242.

[5] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8789–8797.

[6] Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., and Jiao, J. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR* (2018), pp. 994–1003.

[7] Eom, C., and Ham, B. Learning disentangled representation for robust person re-identification. In *Advances in Neural Information Processing Systems* (2019), pp. 5298–5309.

[8] Fan, H., Zheng, L., Yan, C., and Yang, Y. Unsupervised person re-identification: Clustering and fine-tuning. *ACM-TOMM 14*, 4 (2018), 1–18.

[9] Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., and Huang, T. S. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV* (2019), pp. 6112–6121.

[10] Ge, Y., Chen, D., and Li, H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526* (2020).

[11] Ge, Y., Zhu, F., Zhao, R., and Li, H. Structured domain adaptation for unsupervised person re-identification. *arXiv preprint arXiv:2003.06650* (2020).

[12] Gong, S., Cristani, M., Yan, S., and Loy, C. C. Person re-identification. *Incorporated 1447162951* (2014), 9781447162957.

[13] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR* (2016), pp. 770–778.

[14] Hermans, A., Beyer, L., and Leibe, B. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[15] Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multimodal unsupervised image-to-image translation. In *ECCV* (2018), pp. 172–189.

[16] Huang, Y., Wu, Q., Xu, J., and Zhong, Y. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *ICCV* (2019), pp. 9527–9536.

[17] Kumar, D., Siva, P., Marchwica, P., and Wong, A. Unsupervised domain adaptation in person re-id via k-reciprocal clustering and large-scale heterogeneous environment synthesis. *arXiv preprint arXiv:2001.04928* (2020).

[18] Li, M., Zuo, W., and Zhang, D. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586* (2016).

[19] Li, Y.-J., Lin, C.-S., Lin, Y.-B., and Wang, Y.-C. F. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV* (2019), pp. 7919–7929.

[20] Lin, S., Li, H., Li, C.-T., and Kot, A. C. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. *arXiv preprint arXiv:1807.01440* (2018).

[21] Lin, Y., Xie, L., Wu, Y., Yan, C., and Tian, Q. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3390–3399.

[22] Liu, J., Zha, Z.-J., Chen, D., Hong, R., and Wang, M. Adaptive transfer network for cross-domain person re-identification. In *CVPR* (2019), pp. 7202–7211.

[23] Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., and Wang, X. Exploring disentangled feature representation beyond face identification. In *CVPR* (2018), pp. 2080–2089.

[24] Maaten, L. v. d., and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research 9*, Nov (2008), 2579–2605.

[25] Panareda Busto, P., and Gall, J. Open set domain adaptation. In *ICCV* (2017), pp. 754–763.

[26] Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., and Tian, Y. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR* (2016), pp. 1306–1315.

[27] Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y., and Gao, Y. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *ICCV* (2019), pp. 8080–8089.

[28] Ren, C.-X., Liang, B., Ge, P., Zhai, Y., and Lei, Z. Domain adaptive person re-identification via camera style generation and label propagation. *IEEE Transactions on Information Forensics and Security 15* (2019), 1290–1302.

[29] Saito, K., Yamamoto, S., Ushiku, Y., and Harada, T. Open set domain adaptation by backpropagation. In *ECCV* (2018), pp. 153–168.

[30] Sun, Y., Zheng, L., Deng, W., and Wang, S. Svdnet for pedestrian retrieval.

[31] Tran, L., Yin, X., and Liu, X. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR* (2017), pp. 1415–1424.

[32] Wang, D., and Zhang, S. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2020), pp. 10981–10990.

[33] Wang, J., Zhu, X., Gong, S., and Li, W. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR* (2018), pp. 2275–2284.

[34] Wang, Z., Hu, R., Chen, C., Yu, Y., Jiang, J., Liang, C., and Satoh, S. Person re-identification via discrepancy matrix and matrix metric. *IEEE Transactions on Cybernetics 48*, 10 (2018), 3006–3020.

[35] Wang, Z., Jiang, J., Wu, Y., Ye, M., Bai, X., and Satoh, S. Learning sparse and identity-preserved hidden attributes for person re-identification. *IEEE TIP* (2019).

[36] Wang, Z., Jiang, J., Yu, Y., and Satoh, S. Incremental re-identification by cross-direction and cross-ranking adaption. *IEEE Transactions on Multimedia 21*, 9 (2019), 2376–2386.

[37] Wei, L., Zhang, S., Gao, W., and Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In *CVPR* (2018), pp. 79–88.

[38] Wu, A., Zheng, W.-S., and Lai, J.-H. Unsupervised person re-identification by camera-aware similarity consistency learning. In *ICCV* (2019), pp. 6922–6931.

[39] Wu, G., Zhu, X., and Gong, S. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI* (2020), vol. 2.

[40] Wu, Y., Lin, Y., Dong, X., Yan, Y., Bian, W., and Yang, Y. Progressive learning for person re-identification with one example. *TIP 28*, 6 (2019), 2872–2881.

[41] Yang, Q., Yu, H.-X., Wu, A., and Zheng, W.-S. Patch-based discriminative feature learning for unsupervised person re-identification. In *CVPR* (2019), pp. 3633–3642.

[42] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193* (2020).

[43] Yu, H.-X., Zheng, W.-S., Wu, A., Guo, X., Gong, S., and Lai, J.-H. Unsupervised person re-identification by soft multilabel learning. In *CVPR* (2019), pp. 2148–2157.

[44] Zhang, X., Cao, J., Shen, C., and You, M. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *ICCV* (2019), pp. 8222–8231.

[45] Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., and Sun, J. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184* (2017).

[46] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. Scalable person re-identification: A benchmark. In *ICCV* (2015), pp. 1116–1124.

[47] Zheng, Z., Zheng, L., and Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV* (2017), pp. 3754–3762.

[48] Zhong, Z., Zheng, L., Cao, D., and Li, S. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1318–1327.

[49] Zhong, Z., Zheng, L., Li, S., and Yang, Y. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV* (2018), pp. 172–188.

[50] Zhong, Z., Zheng, L., Luo, Z., Li, S., and Yang, Y. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR* (2019), pp. 598–607.

[51] Zhong, Z., Zheng, L., Zheng, Z., Li, S., and Yang, Y. Camera style adaptation for person re-identification. In *CVPR* (2018), pp. 5157–5166.

[52] Zhong, Z., Zheng, L., Zheng, Z., Li, S., and Yang, Y. Camstyle: A novel data augmentation method for person re-identification. *IEEE Transactions on Image Processing 28*, 3 (2018), 1176–1190.

[53] Zhu, X., Morerio, P., and Murino, V. Unsupervised domain-adaptive person re-identification based on attributes. In *ICIP* (2019), IEEE, pp. 4110–4114.