Investigating Domain Transfer and Viewpoint in the Context of Person Re-Id

Zheng Liu

TU Delft

Investigating Domain Transfer and Viewpoint in the Context of Person Re-Id

by

Zheng Liu

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday February 26, 2021 at 10:00 AM.

Student number:	4798406	
Project duration:	March 1, 2020 – January 12, 2021	
Thesis committee:	Dr. ir. J. van Gemert,	TU Delft, supervisor
	Dr. ir. S. Pintea,	TU Delft
	Dr. ir. K. Hildebrandt,	TU Delft
	ir. Y. Napolean,	TU Delft, co-supervisor

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

This article reports my master thesis work "The Influence of External Dataset and Viewpoint of Pedestrians on Task ReId Dataset Performance". The research was conducted at Pattern Recognition and Bioinformatics group, Computer Vision lab under the supervision of Dr. Jan Van Gemert and Yeshwanth Napolean.

This document contains two main parts, firstly a scientific paper presents our motivation, experiments, and results, the rest are supplemental materials that demonstrate fundamental background knowledge which is relevant to the work.

I sincerely thank my supervisor Dr. Jan van Gemert for his guidance and supports, he gave me lots of advice to do scientific research systematically. Furthermore, I also thank my daily supervisor Yeshwanth Napoleon for the regular meetings, on technique details and experiment suggestions and guidance. I would also like to thank Dr. Silvia P and Dr. Klaus H to be a member of my defence committee. Moreover, I thank Dr. Hongjie Liu for her inspiration and Miaoyin Yan and Weichao Yan for their contribution. I thank all my friends I meet in Delft and the tired but abundant and wonderful two years.

In the end, I thank my girlfriend's accompany in the Netherlands during the COVID-19 pandemic time, which support me to go through that hard time. I appreciate all my family members for their concerns and support. I express my deepest appreciation to my parents for their unconditional support and encouragement.

Zheng Liu
Delft, January 2021

Investigating Domain Transfer and Viewpoint in the Context of Person Re-Id

Zheng Liu
Delft University of Technology
Delft, Netherlands
Z.Liu-14@student.tudelft.nl

Abstract

Deep learning has significantly improved Re-Id performance but it requires a large amount of data, however, obtaining data is expensive from both time and money perspective. Inspired by ImageNet pre-trained models and synthetic data generation techniques, this paper investigates to utilise real-world and synthetic Re-Id datasets to augment task performance. Firstly, we propose two methods to apply external Re-Id data, NDTL (Neighbour-Domain Transfer Learning) and NDDS (Neighbour-Domain Data Stitching). Secondly, we quantitatively illustrate that both real-world and synthetic data could mitigate Re-Id data shortage problems, using Re-Id dataset to pre-train models is better than using ImageNet, we achieve up to 28.2% mAP improvement on DukeMTMC and 5.2% on Market-1501. Moreover, we find out that viewpoint, one of Re-Id relevant factors, has the an influence on the system performance due to viewpoint-wise non-alignment and unbalance of the original dataset, it also assists the performance if train set is augmented balanced. Our research strongly illustrates both real-world and synthetic Re-Id dataset can effectively augment Re-Id task, viewpoint is an essential factor and based on which, train-test distribution dramatically influences Re-Id performance, and balancing train classes are also helpful to improve the performance.

Keywords: *Person Re-Identification; transfer learning; synthetic data; augmentation; viewpoint*

1. Introduction

Person Re-Identification (Re-Id) is the task of recognising and matching people from multiple views of cameras where there is no view duplication, it is a popular research topic in computer vision [46, 47] and widely applied in many fields such as intelligent security, video surveillance and traffic control. The early studies focus on conventional computer vision methods and researchers exploit the characteristics, which are also acknowledged as feature representations, from videos and the frames, others pay attention to metric learning [42] relying on the (dis)similarities between images. Since A. Krizhevsky proposing AlexNet [15] in 2010, deep neural network has been playing more and more essential roles in computer vision and Re-Id research. Deep Convolutional Neural Network (CNN) assistant to deal with deep Re-Id problems by extracting features from large dataset automatically, [8, 25, 46, 47] show the improvement, some other methodologies such as part-based convolutional baseline (PCB) [36] exploiting local features of each person and spatial-temporal attention (STA) method combining spatial attention and temporal attention demonstrate the advancement as well.

Re-Id is confronted with various problems, such as low resolution due to camera situation, viewpoint difference, pedestrians being occluded by objects or other people and illumination variation. Besides, a well-performing Re-Id system requires adequate data to trained, which leads the dataset size and labelling issues, however, plenty of conventional and well-known CNN methodologies require fully labelled dataset to proceed supervised learning, but labelling dataset is a tremendously costing work that usually requires a long time period and relevant knowledge to do so, meaning experts would be involved into these tasks, and these will eventually lead money costing. To address this issue, exploring methods that release dataset issue has great significance.

Recently, some works based on generative adversarial network (GAN) are proposed to deal with dataset problems. Z. Zheng proposed a method to refresh people's appearances based on their original appearances and dresses [48], paper [44] introduced data without label to augment training samples by GAN. Unsupervised learning is a machine learning method that learns without labels. Unlabelled data are added into system by clustering them and assigning and refreshing the pseudo-labels [19]. Semi-supervised methods are also applied to Re-Id [2].

These GAN-based and non-fully-supervised methods have achieved certain successes, but these methods are limited to the exploitation of a specific dataset, the features that can be learned are limited and constrained by the task dataset itself, also considering the complexity of designing specific neural network architectures and multifarious data processing, there is still room for improving the method to deal with dataset size problem. We

take inspiration from ImageNet pre-training [4], which is to use the external Re-Id dataset to pre-train and enhance the model, and introduce additional pedestrian and background features to improve the performance and robustness of the model. Based on this, we attempted two data augmentation methods and multiple augmentation data.

The first method we used is transfer learning, different from conventional ImageNet pre-trained models which require such a large dataset ImageNet or downloading the general models directly from framework websites like PyTorch, we pre-train our model by external Re-Id dataset, in another word, we would like to design our own dedicated pre-training models in a specific research field. Once we have an external dataset to do pre-training, another method also come to our mind, that is to directly stitch it with the target dataset into a new larger dataset, and the model will learn information directly and simultaneously both of these two datasets, rather than sequential order of the transfer method.

The augmentation datasets applied in our research include three commonly used real-world dataset, in which two of them are regarded as the target dataset in the experiment, and one artificially designed dataset named PersonX [34]. Since PersonX is a synthetic dataset devised for viewpoint research, we also explored whether information of various viewpoints caused by different angles of photographed people could influence Re-Id system performance, based on which we designed experiments and find that different distribution of viewpoints does have different influence on synthetic augmentation.

The contributions of our work are:

1) we empirically demonstrate that extra existed real-world Re-Id datasets could improve target Re-Id performance, on DukeMTMC and Market-1501 they beat ImageNet pre-trained model 28.3% and 5.2% respectively, this method could be regarded as external Re-Id dataset compared with the target dataset;

2) we empirically demonstrate that external synthetic dataset could improve real-world dataset performance with 22.4% mAP improvement, even splitting synthetic data into multiple small subsets based on their viewpoint, those subsets are still helpful, which lead to a possibility that augmentation and pre-training process could be done by small dataset;

3) we rearrange Re-Id dataset based on their viewpoint, the distribution of train-test set are approximately aligned, but different classes are unbalanced, from whose train-test distribution we illustrate viewpoint variance leads to different viewpoint-based augmentation performance, and if we augment train set to be balanced, the performance could be improves as well but with the significantly small size of data needed.

2. Related Work

2.1. Deep Learning

With the development of deep learning, computer vision research depend much more than previous stage, since neural network and deep architecture could extract features automatically from input data [15, 16], machine learning especially deep leaning based Re-Id research has become the mainstream methodology in this field. M. Ye and et. al. summarised deep learning based Re-Id into five stages, (1) raw data collection, (2) bounding box generation; (3) data labelling, (4) model training and (5) pedestrian retrieval [39, 40], in which stage (4) is highly depended on deep learning technique. Paper [8] proposed an architecture with two subnets focusing on verification and classification respectively. Neural networks learn not only from people's identities but also their attributes such like genders and dresses [20, 25]. Also, the RNN (recurrent neural network) techniques ensure the possibilities of applying people's partial information [6, 11], where images are clipped into parts like heads, necks, torsos, legs and feet and fed into RNN sequentially from up to down [36, 37]. Euclidean distance in feature space is commonly used to calculate (dis)similarity, multiple loss functions are designed to calculate recognition loss, such as Contrastive Loss [35], Triplet Loss [32], Quadruplet Loss [3].

2.2. Transfer Learning

Transfer learning is the technique to adapt learned knowledge from one domain into another one, it assists to reduce learning time consuming with faster convergence and avoid to learn from zero, plenty of research illustrates that it has great essentiality [12, 28, 41]. Generally, in practice, we call the learned knowledge pre-trained model and the target domain to be adapted is the task requiring fine-tuning. Currently, transfer learning based Re-Id focus on style transfer from one dataset to another [5, 50], some other work applied ImageNet together with Re-Id dataset to train the model [8]. However, K. He and et al. query that ImageNet based transfer learning and pre-trained model might not play the most significant role in tasks if more cost is acceptable [9].

2.3. Person Re-Identification

Deep learning based Re-Id contains many research methodologies, such as representation based Re-ID, metric learning based Re-Id, local feature based Re-Id, and GAN based Re-Id [13, 35, 38, 46, 47].

Due to different pedestrians always have unique attributes such as standing posture and gait, how people walk, if dataset comprises a sufficient number of images, these behaviour attributes could be exploited into useful information to support the Re-Id system. Gait is about the way people walk, which is positively correlated with their identities [23, 27]. Pose transferable samples

are used to augment and enhance model training [22], it chose MARS as source data [45], a large size video-based Re-Id dataset containing adequate pose information, and took pose detection algorithm on every sample and obtained skeleton representation. PCB (Part-based Convolutional Baseline) is a novel Re-Id technique exploiting people's partial information, which will be eventually fed into RNN for training sequentially from head to feet [36, 37]. As the demand of surveillance and camera network, large video based Re-Id also become Research hotspots, works paying attention to temporal modelling methods, temporal attention and RNN architectures are proposed in [7, 17, 21, 33].

Confronted with the small dataset size problem and high cost of labelling data, a large number of methods are proposed to solve the problem [5, 19, 29, 49]. As annotating labels is a painful task, utilising data with non-fully-labelled data or even without a label would be available. Unsupervised learning, weakly-supervised learning semi-supervised learning are taking efforts to release the curse of data. The unsupervised method applies the extracted features learned from previous CNN to be clustered and labelled based on probability and repeat refreshing [19]. Paper [24] introduced an innovative video-based Re-Id method which can match different people across views from unaligned image sequences with no labelled pairwise data. A semi-supervised work only focuses on soft label, the label within the camera containing camera information [29]. Transductive Semi-Supervised Metric Learning framework is also a semi-supervised method which is graph-based transductive to obtain triplets in unlabelled data and a degree-based relationship confidence scoring system to reduce incorrect triplets [19]. Weakly supervised learning utilise information like the state of people or identities contained by images [26, 43], not the conventional bounding boxes comparably. Z. Zhou made a brief summary on weakly supervised learning [51]. Generative adversarial network is another way to deal with small dataset problem [49]. Some works show the availability of adapting style or pose from source domain into target domain or generating more training data, in such ways model gains sufficient training samples and robustness [44, 48, 49, 50].

Besides, Re-Id meet other issues like occlusion, illumination, resolution, and viewpoint, requiring model robustness and more data to face these variations. Sun and Zheng quantitatively prove in [34] that people's views have different effects on system.

Widely using synthetic data is a novel and potential solution to release dataset size issue because it is easy to generate data and gain their labels simultaneously because of the operational simplicity of software such as Unity [30], paper [1] and [34] give positive examples of applying synthetic dataset in Re-Id research.

3. Method

The current existing and well-known Re-Id datasets are always in scale $10^4 \sim 10^5$ of images, especially image-based Re-Id. This work is inspired by ImageNet pre-trained models succeeding in various tasks [5, 12]. We hypothesize that Re-Id datasets can be used in a similar transfer learning manner to learn task-specific features and improve Re-Id performance. If it is possible, not only Re-Id topics but also other research domains could exploit their conventional dataset for future tasks, like what ImageNet means significantly to computer vision research these days.

3.1. Neighbour-domain Transfer Learning and Data Stitching

As transfer learning is the machine learning technique that adapts knowledge learned from one field into the others, and currently applied knowledge in computer vision are mostly learned from ImageNet, we query that is too general to adapt certain research domain, although features are widely extracted, there still are plenty of information existing without availably used. We want to pre-train the models from neighbour domains, where data distribution are much closer to our target than the others, and transfer learning method can learn associating features as much as possible with data size required as small as possible comparably. We notate the target domain as $D_T = \{X_T, f_T(X_T)\}$ and the source domain as $D_S = \{X_S, f_S(X_S)\}$, where $X_{T/S}$ is the training data. The learning tasks are T_T and T_S, $T = \{Y, f(\cdot)\}$, respectively, the goal of transfer learning is to improve the performance of the target predictive function $f_T(\cdot)$ by applying the knowledge learned from source domain D_S and its task T_S. The condition is that $D_T \neq D_S$ or $T_T \neq T_S$, referring to the mathematical definition in [28].

Neighbour-domain Transfer Learning (NDTL) is the learning method referring to how conventional transfer learning works have done, by modifying the range of source domain to be limited by $X_S \in \mathcal{R}$, where $\mathcal{R}$ is the set of collection of all, if possible, Re-Id dataset $\mathcal{R} = \{\text{Market-1505}, \text{DukeMTMC}, \cdots, \text{PersonX}\}$, and certainly by this definition $X_T \in \mathcal{R}$. $X_S = \{x_{S_1}, x_{S_2}, \cdots, x_{S_{n_S}}\}$ is the vector of data samples and $Y_S = \{y_{S_1}, y_{S_2}, \cdots, y_{S_{n_S}}\}$ is the vector of corresponding labels, meanwhile $X_T = \{x_{T_1}, x_{T_2}, \cdots, x_{T_{n_T}}\}$ and $Y_T = \{y_{T_1}, y_{T_2}, \cdots, y_{T_{n_T}}\}$ are input and output respectively. To be clear, Y_T and Y_S are not strict mathematical sets in roster notation as multiple instances always share the same label, but duplicate data items are permitted in multiset. Usually, $0 < n_T \ll n_S$, and specifically for Re-Id, the cardinality of samples is always much larger than their labels, which is denoted by $\text{Card}(Y_{T/S}) \ll \text{Card}(X_{T/S})$.

The process is to first pre-train the randomly initialised model on source domain containing an external intro-

duced Re-Id dataset X_S, based on whose knowledge we fine-tune the model on target domain D_T, whose training set is X_T, with supervision, where labels Y_T are available. We keep former learned information partially or fully frozen or train them with comparably lower learning rate such as $\times 0.1$ as latter layers, and the rest layer of neural network could be trained as usual.

Neighbour-domain Data Stitching (NDDS) is the method utilising exactly the same external data X_S to augment the system but in a different way. If above transfer learning method could be described as a sequential process, then NDDS is a simultaneous one. A new dataset is stitched from D_T and D_S, which is notated as D_M, where M represents multiple datasets. Obviously, $X_M = X_T \cup X_S$ in practice, one of X_T and X_S shall be specially marked as some of the training samples share same identities originally, such as ID: 0007 appears in both of Market-1505 and DukeMTMC but corresponding to different pedestrians. The task is then to learn $f_M(\cdot)$ on X_M such that it yields the least error on predicting Y_T. Representation of training sample is $X_M = \{\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \cdots, \mathcal{X}^{(k)}\}$ and $\mathcal{X}^{(i)} \in \mathbb{R}^{H \times W \times N}$, where $\mathcal{X}^{(i)}$ is the i^{th} mini-batch, H and W are height and weight of input images and N is batch size, another way to express is $X = \{x^{(1)}, x^{(2)}, \cdots, x^{(\text{Card}(M))}\}$, where $X \in \mathbb{R}^{H \times W \times N \times \text{Card}(M)}$. We omit network architecture information channel C for simplicity of expression. Thus the task for NDDS is to solve the optimisation problem, shown in equation 1:

$$\arg\min_{f(\cdot)} \frac{1}{\text{Card}(M)} \left(\sum_{i=1}^{\text{Card}(M)} L(y_i, f(x_i)) + \lambda J(f(\cdot)) \right) \tag{1}$$

where, by definition, $x_i \in X_M$ is the input training data, y_i is the label, $L(\cdot)$ is loss function, $J(\cdot)$ is regularisation item to constrain parameters not to overfit, such as L_2-norm, and λ is the hyper-parameter of this item, $f(\cdot)$ is the set of parameters, which is also known as the model, we want to learn based on $\text{Card}(M)$ training samples.

3.2. Synthetic to Real-world Augmentation

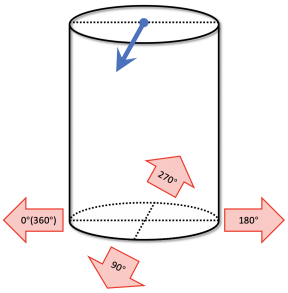
Considering exiting dataset could be helpful to mitigate dataset size issue, but the number of those datasets themselves are limited, especially the commercially used ones are always not open source, so the idea that utilising synthetic data to augment Re-Id system came into our mind. It is clear that differing from real-world datasets which are collected in certain locations and time periods, leading to irrevocable characteristics, synthetic data is much more flexible to be modified once we meet new scene and requirement. The core characteristics contain viewpoint, illumination, scene, dresses and so on. Unity [30] is the platform to generate synthetic engine which is controllable and flexible, PersonX is a 3D engine involved in this research, which is originally applied by X. Sun in [34]. The dataset contains six scenes, but we choose one of them as the source dataset. The six subsets share exactly the same design protocol with only different scene, thus we describe it below omitting the scene protocol. There are $1,266$ identities in this dataset, in order to guarantee the diversity, different identities are endowed different complexion, ages, height, weight and hair styles, etc. For each identity, one has 36 viewpoints divided with interval $10°$, Figure 1(a) illustrates how a person's view is determined abstractly and Figure 1(b) gives an example. We notate this synthetic dataset as source domain D_P and its train set X_P, P stands for the initial of PersonX. From viewpoint perspective, $X_P = \{X_P^{(10°)}, X_P^{(20°)}, \cdots, X_P^{(i)}, \cdots, X_P^{(360°)}\}$, where $X_P^{(i)}$ is the collection of identities in view i, and we have that for any $i, j \in \{10°, 20°, \cdots, 360°\}$, $\text{Card}(X_P^{(i)}) = \text{Card}(X_P^{(j)})$, it means all $1,266$ identities appear in all 36 viewpoints, say all viewpoints in this source domain are balanced.

To validate the availability of using synthetic dataset to augment Re-Id system, we apply NDTL and NDDS introduced above. One is to pre-train a model on X_P and fine-tune on our target dataset X_T, the other one is to stitch two datasets X_P and X_T to a new dataset and train from sketch and test on test set of target dataset. From the characteristic that all viewpoints possess same number of images and pedestrians, we assume that, theoretically, every viewpoint shall support performance equally; if not, it means different viewpoints make different improvement with even same identities, scene and size, and then it may come to some principles. Thus there will be 36 parallel experiments for each of NDTL and NDDS, each subset $X_P^{(i)}$ is regarded as source train data to augment X_T to achieve 36 $f^{(i)}(\cdot)$ for testing.

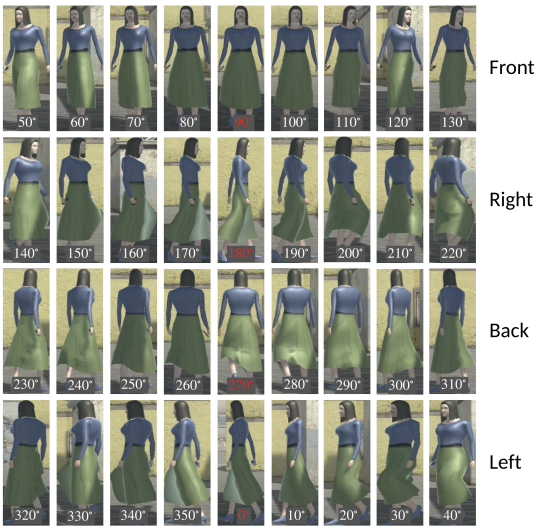
3.3. Viewpoint-Alignment-based Augmentation

As if the existed Re-Id data could be exploited externally on target, say, external dataset could be used directly on target dataset, then we draw an assumption that the dataset itself may be exploited out useful information internally based on their characteristics such as viewpoints distribution, 36 divisions with interval $10°$ is one of the situations, others could be four divisions represented as the front, back, left and right, and to be more general, illumination and other features' distribution inside the target dataset may contain useful information to augment the system.

In this research, since the hypothesis above may not be true, which means different viewpoints may support Re-Id system differently even with the same number of (pre-)train set, we aim to explore influence of viewpoint distribution of the original dataset on model performance, whose 36 viewpoints may be distributed unbalanced on both train-set-classes and train-test-alignment perspec-



(a) 3D demo



(b) A PersonX pedestrian example containing 36 images

Figure 1. (a) Abstract schematic diagram of pedestrian in 3D space. The cylinder represents the human torso, the blue arrow represents the front direction of person, and the four arrows at the bottom represent centers of four directions; (b) dataset demonstration, red tags is equivalent to the bottom arrows in (a) corresponding to due right, front, back and left.

tive, the distribution of dataset might be the explanation. If augmentation shall be accomplished based on viewpoint distribution, then the viewpoint distribution of target dataset is required, which ask us to modify the expression of dataset storage in devices from ID-based shown in equation 2, which is the original dataset, to viewpoint-based dataset and the structure is shown in equation 3, for both train and test set.

$$X_T = \{X_T^{(0002)}, X_T^{(0003)}, \cdots, X_T^{(i)}, \cdots\} \\ i \in \{\text{labels of identities in } X_T\} \tag{2}$$

$$X_T = \{X_T^{(10^\circ)}, X_T^{(10^\circ)}, \cdots, X_T^{(j)}, \cdots\} \\ j \in \{X_T^{(10^\circ)}, \cdots, X_T^{(i)}, \cdots, X_T^{(360^\circ)}\} \tag{3}$$

This modification requires manually labelling referring to viewpoint standard and examples from Figure 1; in another word, new datasets are regenerated. Here we proposed two ways to align the new dataset, one is to augment train set such that it aligns best with test set, the other one is to augment train set itself such that it is balanced, the former one is to counteract the possible trick that some views in train and test fit well but the others not, which is biased distribution and may be led by sampling, the latter one is to counterbalance the possible disproportion of train set. So to augment train set to align with test set, the task is to calculate the size difference $\Delta = \{\Delta^{(10^\circ)}, \cdots, \Delta^{(i)}, \cdots, \Delta^{(360^\circ)}\}$ for each viewpoint i and interpolate where train is smaller than test, the process is mathematically expressed in equation 4. To balance train set, the viewpoint containing the maximum number of images is the reference indicator that augmentation should make all the others meet this number of images, equation 5 frankly explains.

$$\Delta^{(i)} = (\text{Card}(X_{T,\text{test}}^{(i)}) - \text{Card}(X_{T,\text{train}}^{(i)}))_+ \\ (z)_+ = max(z, 0) \tag{4}$$

$$\Delta^{(i)} = \text{Card}(X_{T,\text{train}}^{(i^*)}) - \text{Card}(X_{T,\text{train}}^{(i)}) \\ i^* = \underset{i \in 10^\circ \sim 360^\circ}{\arg\max} \ \text{Card}(X_{T,\text{train}}^{(i)}) \tag{5}$$

Based on the 1×36 vector Δ, augmentation data should be randomly selected from source data for every viewpoint, respectively, and here the source dataset is PersonX. For those data picked out, the augmentation is to apply them into original target Re-Id dataset, and there are two ways to do so. One is to use them based on IDs, and the other one is to keep them in viewpoint form, both of them, due to following Δ, maintaining exactly the same data and angularly meet the requirement of two protocols to align train set.

4. Experiments

Our target of experiment design contains the following several parts:

- Validate whether real-world external Re-Id dataset improve system performance;
- Validate whether synthetic external Re-Id dataset improve system performance;
- Explore whether different viewpoints have different augmentation effects;
- If viewpoints do augment Re-Id system differing from each other, is it the reason that the original dataset distribution of train and test set does not align?
- or different views are unbalanced?

This section simply demonstrates the dataset involved in the experiment and setting protocols. Some settings are amply illustrated in section 5, as results of the pre-experiment are achieved.

4.1. Datasets

In this research, to explore that our methods do assistant Re-Id system performance on dealing with small size dataset, we introduced two classical real-world datasets collected years ago as our target, an extra real-world one to provide external sources, a synthetic dataset as external dataset and in the viewpoint research those two task dataset are manually labelled as two novel dataset. Besides, model pre-trained on ImageNet [4, 10, 12] is regularly used, yet this dataset itself is not directly operated, so its details will not be discussed in the followings.

*(1) Real-world data.***(a)DukeMTMC-reID** [31] was collected in Duke University, it contains 16,522 train images and 2,228 query images from different 702 people, 17,661 images are in gallery; **(b)Market-1501** [46] includes 32,668 images under 1501 identities, in which 19,372 and 12,936 are used for testing and training separately, query folder contains 3,368 images; **(c)CUHK03** is the first dataset which is sufficient enough for deep learning, it was collected in campus of The Chinese University Hong Kong [18] containing 1467 pedestrians collected from 5 cameras.

(2) synthetic data. In paper [34], X. Sun and L. Zheng introduced a artificially designed dataset **PersonX** in viewpoint research. PersonX is the dataset generated via Unity [30].

(3) Manually labelled data. For searching whether each viewpoint has special influence on Re-Id system and is the performance determined by train-test alignment on viewpoint perspective, we developed two novel datasets by manually labelling DukeMTMC-reID and Market-1501, involving train, validation, gallery and query, based on their appearance, referring to the standard illustrated in Figure 1. For each labelled dataset respectively, the parents' folder contains 36 folders representing the 36 classes with various views divided by interval 10°. In order to distinguish the original datasets from these labelled ones, we call them **DukeMTMC-view** and **Market-view** respectively.

There are mainly four datasets involved in this project[1]. Market-1501 [46] was collected by 6 cameras in open environment in Tsinghua University. This dataset contains 1501 pedestrians, 751 of them are used for training, and the rest 750 are for testing, 32668 images are in total, within which 12936 are used for training, the rest are for testing. Duke [31] CUHK03 [18] ImageNet [4].

4.2. Evaluation Metrics

In order to evaluate the performance of the Re-Id system, the commonly applied mAP (mean Average Precision, shown in equation 6), Rank-n (n = $1, 5, 10$) and CMC (cumulative matching characteristics) curve [46] are used to quantify the result.

$$\text{mAP} = \frac{\sum_{i=1}^{N_q} \text{AP}_i}{N_q} \tag{6}$$

where N_q is number of query inages and AP is defined based on precision and recall value as AP $= \sum_k precision_k \frac{1}{N_r}$ ($k \in \Omega$), N_r is the counting of recall being refreshed; Rank-n is the value on CMC curve when rank=n.

4.3. Protocols

The implementation of network in this work is based on ResNet50 [10], all image boxes are reshaped into $3 \times 28 \times 28$ resolution before fed into networks. Based on several pre-experiments, Adam [14] is picked as the optimiser, learning rate is set 0.01 for training and 0.001 for starting of fine-tune, due to the usual size of Re-Id image, batch size is 64.

We trained the models based on relevant augmentation protocols, such as (a) augmenting DukeMTMC-reID by Market-1501, CUHK03 and Market-1501+CUHK, similarly for Market-1501 when it is regarded as task set; (b) augmenting DukeMTMC-reID by each viewpoint of PersonX. For same group of experiments, we keep the same initialisation, like DukeMTMC-reID Re-Id systems augmented by CUHK03 and Market-1501 are started from the same sketch.

5. Performance Analysis

5.1. NDTL/NDDS methods availability and real-world dataset augmentation

To validate the availability of real-world data whether it can mitigate dataset size issue and beat ImageNet with a small body, we take two dataset DukeMTMC and Market-1501 as target sets, parallelly, once on of them is regarded as the target, the other one could be source data. Moreover, CUHK03 is purely a source dataset in this research. So based on NDTL and NDDS, we have multiple possible combinations to train models. We use the model pre-trained on ImageNet as the baseline, and the result is shown in Table 1.

It is obvious that for both DukeMTMC and Market-1501, the performances of the models trained based on NDTL and NDDS are significantly better than the one of ImageNet pre-trained. These results meet our hypothesis that Re-Id datasets are capable of leading more crucial information for the system even with a much smaller size. The sizes of real-world datasets involved in this research are DukeMTMC: 67.2MB, Market-1501: 31.6MB and CUHK03 38.9MB, yet ImageNet contains 147GB. As the core interest of this hypothesis is to find out either real-world dataset or ImageNet is better, so Table 1 only gives the comparison for that, the question of whether

[1]The details dataset

Table 1. Results of real-world Re-Id dataset augmentation on target dataset. DukeMTMC and Market-1501 are also used as source data while not augmenting themselves, respectively. The initials C., M., and D. are those three real-world datasets applied in our experiment, I.Net Pre. ✓/✗ means whether the model is pre-trained on ImageNet. For Duke, models trained by NDTL and NDDS all perform better than the baseline, and similar for Market-1501 except one case, external Re-Id datasets do assist target Re-Id better than ImageNet.

Dataset	Method	Ext. Data	mAP	R-1	R-10
Market	NDTL	Duke	**59.2**	82.8	95.3
		CUHK	55.1	79.5	94.8
	NDDS	Duke	49.7	68.9	87.3
		CUHK	**58.6**	80.2	94.8
	-	I.Net Pre. ✗	39.8	64.9	89.8
		I.Net Pre. ✓	54.0	77.6	94.2
Duke	NDTL	Market	**51.4**	72.3	87.9
		CUHK	45.8	67.7	85.3
	NDSS	Market	**59.6**	76.0	91.1
		CUHK	49.3	68.8	87.2
	-	I.Net Pre. ✗	28.1	44.5	73.2
		I.Net Pre. ✓	31.3	48.2	74.4

ImageNet is still useful is answered in Table 2, the result drawn from model pre-trained on Imagenet and then pre-trained one more time on Market-1501, is better than both separately pre-trained on those two datasets, so ImageNet still can offer useful information but not as vital as Re-Id dataset.

Table 2. The performance in percentage on DukeMTMC dataset using Market-1501 as source dataset. Four models represent the combination of whether the skeleton is pre-trained on ImageNet and which of NDTL/NDDS is applied. Even we have the conclusion that Re-Id dataset is more useful than ImageNet but utilising ImageNet could still help; this is not the core interest here but a clarification for ImageNet.

Model	mAP	R-1	R-5	R-10
NDTL (I.Net Pre. ✓)	61.8	77.7	88.8	92.3
NDTL (I.Net Pre. ✗)	51.4	72.3	84.1	87.9
NDDS (I.Net Pre. ✓)	66.4	85.8	93.7	96.5
NDDS (I.Net Pre. ✗)	59.6	76.0	87.9	91.1

5.2. Synthetic dataset augmentation and viewpoint influence

We validated the feasibility of synthetic dataset via PersonX. The result illustrated in Table 3 of both NDTL and NDDS methods convinces that synthetic data can be beneficial to Re-Id performance.

Table 3. Synthetic data performance on augmenting DukeMTMC. For the right and left viewpoint, considering the symmetry, we randomly pick "right" and "left" images to meet the same number as "front" or "back"; "mix" is all four viewpoints together but only owe the same number as a full single viewpoint; "all" represents the full PersonX dataset. For each side, there is no significant difference between them; full dataset behaves better than each viewpoints.

Dataset	Viewpoint	mAP
NDTL	front	53.4
	back	53.3
	side	53.2
	mix	53.5
	all	**53.7**
NDDS	front	48.5
	back	48.5
	side	48.9
	mix	48.9
	all	**49.6**

As PersonX is designed originally for viewpoint research, we would like to see whether viewpoints influence the performance. Theoretically, with the same number of augmentation data on each viewpoint, the performance improvement should be approximately similar; otherwise, either different viewpoints assistant the system discrepantly because of their own unexplored and unknown features; for instance, some front viewpoints conduct higher accuracy than the ones of back angles on account of fronts consist of more facial information than backs, or the target dataset itself containing certain distributions that leads the bias augmentation result, such as training sample bias which induces certain viewpoints owning adequate data are well trained while the others lack of training.

Table 4 is the test results of DukeMTMC augmented by each view of PersonX, the table is made referring to Figure 1(b), it is clear that even with such small dataset augmentation, 410 images in 4.9MB, the performance is still better than ImageNet pre-trained one, however, we see that results from different viewpoints differ from each other. To demonstrate the internal variance schematic, we drew the heat-map 2 based on the mean value of all results. The models are trained by NDTL on each viewpoint, the value in every block is the difference between each mAP and mean value of all of them, $\Delta\text{mAP}^{(i)} = \text{mAP}^{(i)} - mean(\text{mAP})$, white colour is where mAP is equal to the average, red represents improvement and blue represents decline. From the figure, we can intu-

Table 4. The performance in mAP of models applying NDTL on PersonX data on each viewpoint. This table refers structure of Figure 1(b), to be clear for comparison, our baseline value is 31.3. It is obvious that all 36 viewpoints in PersonX could improve DukeMTMC Re-Id result, but their performance differ from each other.

$54.0^{(50°)}$	$53.8^{(60°)}$	$53.2^{(70°)}$	$52.9^{(80°)}$	$52.3^{(90°)}$	$53.3^{(100°)}$	$52.1^{(110°)}$	$53.6^{(120°)}$	$52.8^{(130°)}$
$53.3^{(140°)}$	$53.5^{(150°)}$	$52.7^{(160°)}$	$53.2^{(170°)}$	$52.4^{(180°)}$	$52.4^{(190°)}$	$53.8^{(200°)}$	$53.5^{(210°)}$	$54.1^{(220°)}$
$54.2^{(230°)}$	$53.0^{(240°)}$	$51.2^{(250°)}$	$53.8^{(260°)}$	$53.8^{(270°)}$	$53.1^{(280°)}$	$53.8^{(290°)}$	$52.9^{(300°)}$	$54.2^{(310°)}$
$54.5^{(320°)}$	$53.9^{(330°)}$	$53.7^{(340°)}$	$52.0^{(350°)}$	$53.4^{(360°)}$	$53.5^{(10°)}$	$53.2^{(20°)}$	$51.4^{(30°)}$	$52.4^{(40°)}$

itively distinguish that the margin of each side, such as $40°$, and $130°$, performs better than the four due viewpoints, graphically, in the heat-map side part obtains more red blocks than central part, thus is leads our interests to figure out the reason of these biases.

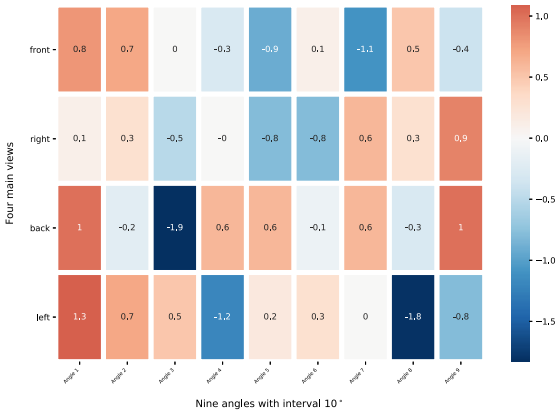


Figure 2. Augmentation performance of 36 viewpoints generated based on Table 4. The warmer te colour is, the better a viewpoint performs augmentation. Red blocks concentrate on the side of the map yet blue ones are in central location, and it presumes that side viewpoints of a direction can augment Re-Id system better than the central ones.

5.3. Dataset alignment

As target dataset alignment is one of the possible reason for bias viewpoint data augmentation discussed above, the distribution of viewpoints on Re-Id dataset is required; thus we manually labelled the original dataset into viewpoint-based version, Duke-view and Market-view. The result of the distribution of Duke-view is shown in Figure 3, it demonstrates distinctly that number of people appear in different viewpoints vary significantly from each other, this directly shatters our intuition that pedestrians including their viewpoints in Re-Id dataset are randomly shot by camera, selected and clipped by researcher, which would make viewpoint-based dataset approximately balanced, so we coincidentally find out that either cameras are specifically located in certain places and angle, or people under cameras where they are located always walk to some certain directions, for example, people can only walk in the direction of the road on a narrow road, but in a place like square, people have many choices in the direction of walking.

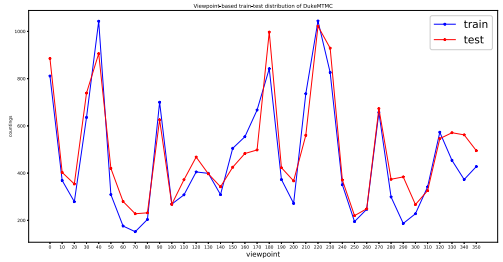


Figure 3. Class distribution of Duke-view. 36 classes are 36 viewpoints with interval $10°$, different viewpoints possess significantly different number of images, meanwhile, train set approximately align with test set, they have similar distribution trend.

Comparing Figure 2 with 3, we can see that roughly where there is peak value in distribution figure, there is red block in heat-map, such as viewpoints $40°$ and $130°$. Thus, we believe the dominant factor of the augmentation bias is not the variance of viewpoint-based PersonX data but data distribution of original dataset on viewpoint perspective. Based on this conclusion, we realise that the dataset itself could offer useful information, and in this research, the information is data distribution of 36 viewpoints, and full PersonX is not indispensable as there are some viewpoints that possess sufficient data for training. Thus we aim to find out whether augmenting train set such that it aligns best with test set helps or augmenting train set to be balanced helps. To be detailed, we take due back as an example, the reason that the augmentation is bad is due to there are more test samples of back side than ones of train, which is train-test non-alignment, or there are too few back images compared with front, right and left on train unbalance perspective.

Train-Test Alignment. We pick data from PersonX based on the difference of train-test distribution on viewpoint perspective such that train set aligns best with test set. Once we have these data, we have two ways to use them as described in section 3.3. Table 5 illustrates result of augmenting train-test alignment by PersonX data.

From Table 5, we have a clear result that augmenting Re-Id train-test-alignment-wise and viewpoint-wise does

Table 5. Model performance augmented based on train-test alignment. "ID." means we split the picked augmentation data based on their identities, and "View." represents viewpoint leading to the separation of augmentation data based on viewpoints. NDTL wins the best with augmentation data being arranged into different classes based on their viewpoints.

Aug. Method	Aug. Arrangement	mAP	R-1	R-10
NDTL	ID.	48.5	69.1	86.8
	View.	**59.3**	**90.7**	**96.6**
NDDS	ID.	51.6	69.8	88.2
	View.	52.1	70.6	88.2

improve system performance compared with the baseline result, and in these four results, the model augmented by viewpoint-based assignment synthetic data and NDTL method performs best, the augmentation data only contains 1916 images.

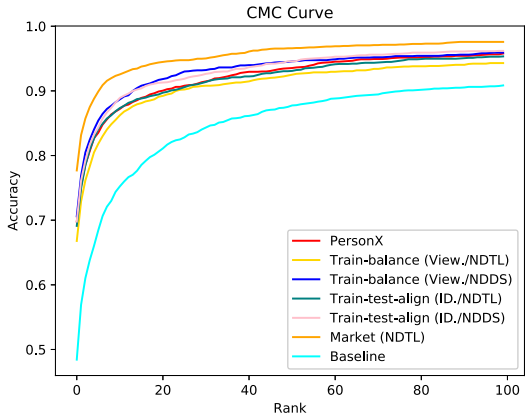


Figure 4. Comparison of multiple CMCs drawn from multiple models. Our methods are all better than baseline, the Rank-1 values concentrate upon 70% approximately.

Train Set Balance. Similarly we do augmentation based on train set itself to make all viewpoints balanced, referring to the viewpoint with maximum training images. The result is shown in Table 6. ID-based assignment conducts the best performance within the four result, and all of them are better than the baseline, the augmentation data size is 13112 here.

From the results of these two ways to pick up augmentation data, we can see that both of them could beat the performance pre-trained on ImageNet with much less data usage than full PersonX, NDTL method even achieved a result as good as the one we apply full Market-1501 dataset. Compare these two augmentation ways, train-test-alignment-wise is better overall than train-balance-wise 6.3% averagely, and it means to augment Re-Id dataset on viewpoint perspective, train-test alignment is more crucial than train set balance even with

Table 6. Model performance augmented based on train set classes balance. The overall performance is worse than train-test alignment but better than baseline. NDDS with augmentation data arranged based on their identities performs best.

Aug. Method	Aug. Arrangement	mAP	R-1	R-10
NDTL	ID.	45.4	65.4	85.5
	View.	47.0	66.8	85.5
NDDS	ID.	**49.5**	**68.9**	**87.1**
	View.	44.5	63.7	84.0

applying fewer augmentation data. However, the former better protocol is not available in practice because we need to regard test set unknown while train set is completely operable. Nevertheless it does not matter because the aim to do the train-test-alignment experiment is not to design the augmentation method to improve performance but find out the possible reason for bias appears in different viewpoints augmentation.

Besides the accuracy improvement, we also monitored the model training process. Figure 4 demonstrates CMC curve where we pick Rank-n value, and all of our methods are better than the baseline. Figure 5 demonstrates the comparison of training curves of model trained on ImageNet pre-trained model, which is the baseline, Market-1501 NDTL/NDDS model, PersonX NDTL/NDDS model and train-balanced-wise PersonX NDTL/NDDS model, and it is clear that Re-Id neighbour domain assists its "own" training process better than "the others" out of Re-Id field. Furthermore, of course, the main task of our work is to consider the cost to save time and money and find a way better than conventional ImageNet, K. He in [9] argued that ImageNet might not be significant in some situation like epochs are large, our neighbour domain augmentation also might not be crucial on mathematical and accuracy perspective if the cost is ignored. Yet, this work is to find the way to achieve better performance than the conventional method with cost as low as possible. Moreover, Figure 6 illustrates feature response examples of three pedestrians that most of them can learn and contain more useful information than the baseline model. Figure 7 gives an example of Re-Id performance on 4 models, which also shows our methods advancement.

6. Conclusion

In this research, we empirically demonstrate our hypothesises that external Re-Id datasets are indeed accommodating to be adapted on target Re-Id dataset and assistive to mitigate dataset size and labelling issues, and this idea may be generalised to wider range to be validated. Furthermore, not only the real-world data could support Re-Id system but synthetic data as well, it means in future research, business and industrial application, people

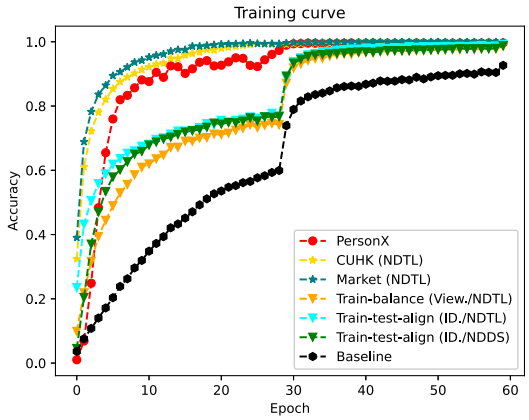


Figure 5. Different models have different convergence trends. Our methods converge faster than baseline, models pre-trained on real-world dataset CUHK03 and Market-1501 converge fastest.

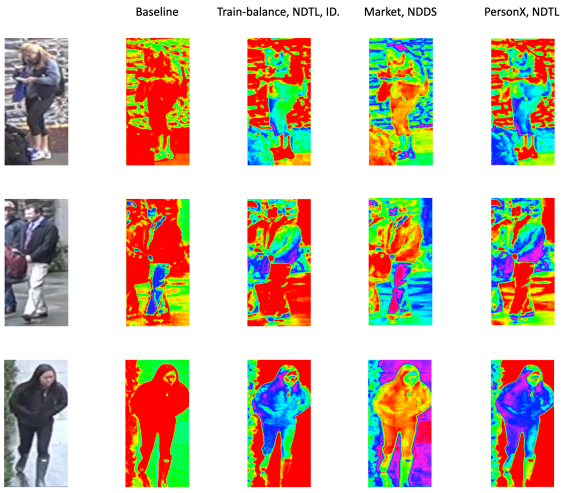


Figure 6. Feature response demonstration of different augmentation methods. The response of our methods is more distinguishable than the baseline. From red to blue, the colder the colour is, the higher response a image has.

may generate synthetic data rather than collecting requisite ones all from real-world for system augmentation, which is able to save time and money simultaneously and significantly. To be specific, some viewpoints do assistant Re-Id system more than other others, which means we can take one more step to save the cost to collect and generate data that are more helpful, but still, the experiments are still needed on those datasets having already aligned train-test set or balanced train set, based on which viewpoint influence could be discussed more detailed, and it may lead commercial applications to collect data and train the model within the best condition. On the other perspective, augmenting data domain-internally and augmenting synthetic data are the ideas that could be extended, as ImageNet is too general with containing too many images variously to be domain-specifically augmented, for instance, vehicle Re-Id task could use either existed real-world vehicle recognition datasets or synthetic vehicle images to augment and pre-train. Just like what ImageNet has been done in computer vision field, neighbour-domain augmentation and synthetic augmentation may become supportive methods for the other fields, there could be numbers of novel research topics, and the conventional dataset may reignite anew.

References

[1] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *CoRR*, abs/1701.03153, 2017.

[2] X. Chang, Z. Ma, X. Wei, X. Hong, and Y. Gong. Transductive semi-supervised metric learning for person re-identification. *Pattern Recognition*, 108:107569, 2020.

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. 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)*, June 2018.

[6] M. Egmont-Petersen, D. de Ridder, and H. Handels. Image processing with neural networks—a review. *Pattern recognition*, 35(10):2279–2301, 2002.

[7] J. Gao and R. Nevatia. Revisiting temporal modeling for video-based person reid. *CoRR*, abs/1805.02104, 2018.

[8] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *CoRR*, abs/1611.05244, 2016.

[9] K. He, R. Girshick, and P. Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[13] R. Iguernaissi, D. Merad, K. Aziz, and P. Drap. People tracking in multi-camera systems: a review. *Multimedia Tools and Applications*, 78(8):10773–10793, 2019.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

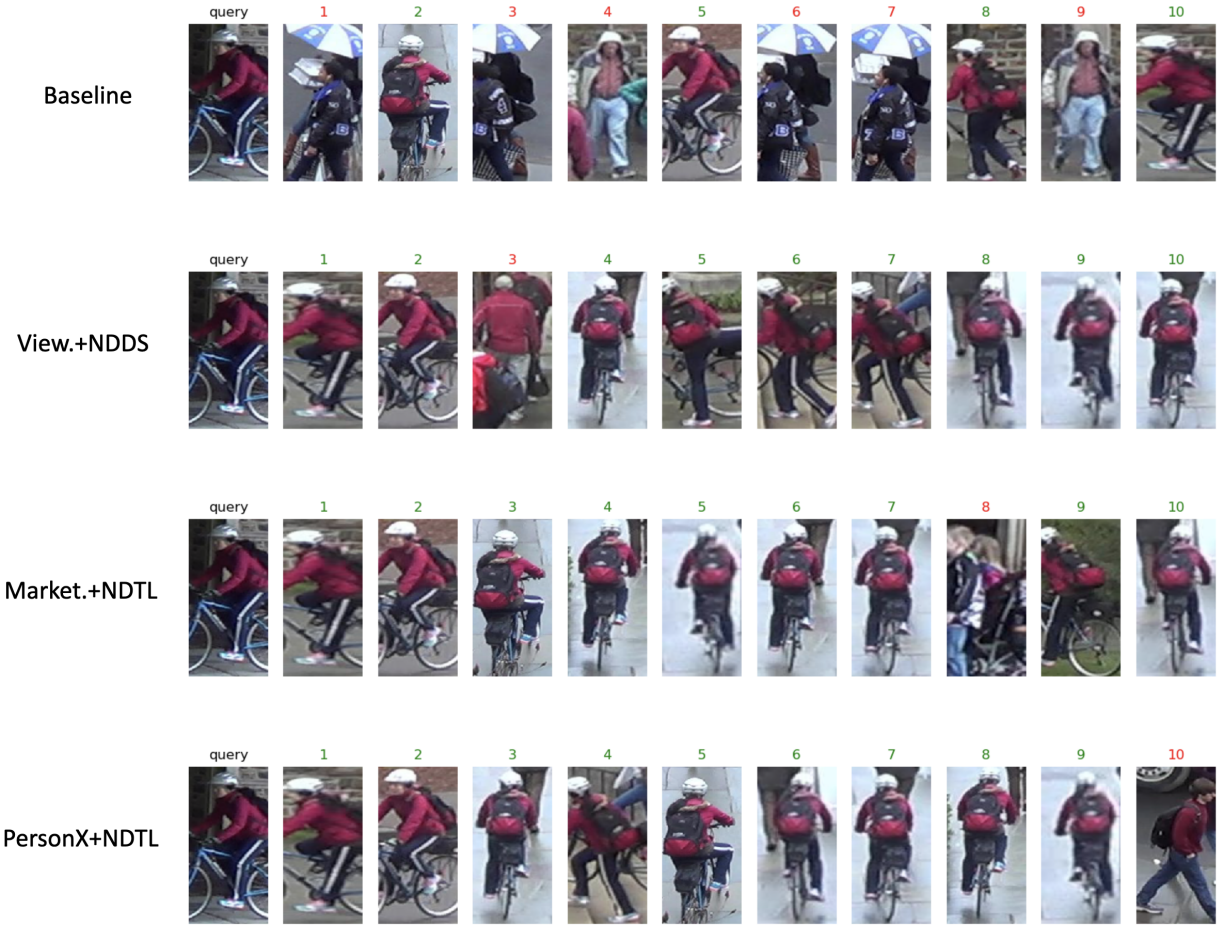


Figure 7. A demonstration of DukeMTMC Re-Id given a query that top 10 recognised images are listed. Green is correct and red is wrong. Compared with the baseline, our approaches perform better.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[19] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8738–8745, 2019.

[20] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019.

[21] C.-T. Liu, C.-W. Wu, Y.-C. F. Wang, and S.-Y. Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019.

[22] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[23] Z. Liu, Z. Zhang, and Q. Wu. Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168, 05 2015.

[24] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[25] T. Matsukawa and E. Suzuki. Person re-identification using cnn features learned from combination of attributes. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2428–2433. IEEE, 2016.

[26] J. Meng, S. Wu, and W.-S. Zheng. Weakly supervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2019.

[27] A. Nambiar, A. Bernardino, and J. C. Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys (CSUR)*, 52(2):1–34, 2019.

[28] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[29] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao. Progressive cross-camera soft-label learning for semi-supervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2815–2829, 2020.

[30] J. Riccitiello. John riccitiello sets out to identify the engine of growth for unity technologies (interview). January, 18, 2015.

[31] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[33] W. Song, Y. Wu, J. Zheng, C. Chen, and F. Liu. Extended global–local representation learning for video person re-identification. *IEEE Access*, 7:122684–122696, 2019.

[34] X. Sun and L. Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, 2019.

[35] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016.

[36] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. *CoRR*, abs/1607.08381, 2016.

[37] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 420–428, 2017.

[38] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371, 2019.

[39] F. Xiong, Y. Xiao, Z. Cao, K. Gong, Z. Fang, and J. T. Zhou. Good practices on building effective cnn baseline model for person re-identification. In *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*, volume 11069, page 110690I. International Society for Optics and Photonics, 2019.

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

[41] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.

[42] J. You, A. Wu, X. Li, and W. Zheng. Top-push video-based person re-identification. *CoRR*, abs/1604.08683, 2016.

[43] H.-X. Yu and W.-S. Zheng. Weakly supervised discriminative feature learning with state information for person identification. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recogni- tion (CVPR)*, 2020.

[44] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9021–9030, 2020.

[45] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 868–884, Cham, 2016. Springer International Publishing.

[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[47] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[48] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.

[49] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[50] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018.

[51] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

Contents

1 Introduction

1.1. Motivation

As deep learning technology are widely developed and applied in computer vision research and business practice, the more complex models are designed and more data are collected and generated. However, gaining data is a significantly time-consuming and money-consuming process which includes and is more than video recording, image shooting, data pre-processing and data labelling. Generally, the more data a programme needs, the more it cost to gain data. Theoretically, take manual data labelling as example, the time spent on data labelling is linearly related with the size of dataset, considering the large workload may reduce work efficiency and productivity, the relationship of time-money-cost and size of dataset maybe worse than linearity. As table 1.1 shown below, the most well-known dataset are size-limited.

On the other hand, in computer vision research even ImageNet is already broadly applied to pre-trained model, which is easy to be achieved in deep learning framework such as PyTorch and TensorFlow, ImageNet is still too general to be specific for a certain research domain, although it does improve the performance in a comparable short time for training (Kaiming He and et. al. claim in [6] that if time costs are not considered ImageNet is not indispensable), we still expect that the domain-internal dataset would provide better supports than ImageNet, meanwhile the dataset size is certainly smaller than ImageNet.

1.2. Research Question

As we see the disadvantage of using ImageNet to pre-train model due to the low domain relevance, the idea that applying re-Id dataset to pre-train the model would benefit the system performance. Furthermore, directly training the external re-Id dataset together with the task set is considered as a reasonable method.

- Do external dataset influence Re-ID performance?
  - What is the efficient method to use external dataset?
  - How much does external dataset improve the performance if it is helpful?
  - How it influence the training phase like convergence?

Even we have real world re-Id data to augment the system, they are still limited in a way, some of them are specifically designed for certain tasks, some of them take settled scene and video sequence losing randomness. To deal these issues, synthetic data, which could be designed based on tasks' requirements, such as viewpoint, illumination and occlusion, are included in our consideration. So we would like to find out how synthetic data would benefit real world re-Id performance and draw the research question and sub-questions as follow:

- Do synthetic data support real world data?
  - Can synthetic dataset make improvement on real world dataset?
  - Which parts of synthetic data make the most contributions?

It is wildly known that data distribution may influence prediction accuracy, based on the results we achieved in previous steps, we wonder that the prediction performance may be determined by the distribution alignment of train and test data. So we take effort to explore whether train-test-oriented data distribution influence re-Id performance, we have third research question and related sub-questions:

- How does data distribution influence augmentation performance?
  - What is the train-test distribution of task datasets? Do they align?
  - Can performance be improved if train set is augmented that it align best with test set?

Table 1.1: Example of some some well-known Re-Id dataset and their basic information. Due to the price of collecting and annotating data as well as the privacy issues, gaining larger scale dataset is not easy.

Dataset	CUHK01[12]	VIPeR[4]	PRID[8]	Market-1501[26]	DukeMTMC[17]	MSMT-17[24]
BBoxes	3,884	1,264	1,134	32,668	36,411	126,411
Identities	971	632	934	1501	1812	4104
Cameras	10	2	2	6	8	15
Detector	hand	hand	hand	DPM	hand	FasterRCNN
Scene	indoor	outdoor	outdoor	outdoor	outdoor	indoor/outdoor

2

Deep Learning

Deep learning is a branch of machine learning, meanwhile, machine learning is also a branch of artificial intelligence. Unlike machine learning to extract features manually, deep learning completes the task of automatically extracting features by machine. LeNet [11] opened the era of using gradient descent to train convolutional neural networks. However in the later research neural network did not achieve excellent performance as the other conventional machine learning algorithm due to the hardware and algorithm limitation. General-purpose GPUs and programming frameworks like CUDA changed this situation, depending on this, AlexNet [10], the deep neural network was designed to break through manual feature extraction. After the inspiration of AlexNet, neural network development has shown the blowout, such as VGG [20] containing repeated using of blocks, GoogLeNet containing parallel blocks, ResNet and DenseNet considering information from former layers and Recurrent Neural Network (RNN) containing sequence information which dominate the performance in Natural Language Processing (NLP), speech recognition, machine translation and etc..

2.1. Convolutional Neural Network

Convolutional Neural Network (CNN) is the representing algorithm of deep learning which conclude convolution al calculation and feed-forward neural network.

2.1.1. Architecture

- convolutional layer

The expression of convolution in calculus (equation 2.1) is:

$$\boldsymbol{S}(t) = \int \boldsymbol{x}(t-a)\boldsymbol{w}(a)da \tag{2.1}$$

The discrete form (equation 2.2) is:

$$\boldsymbol{s}(t) = \sum_{a} \boldsymbol{x}(t-a)\boldsymbol{w}(a) \tag{2.2}$$

This formula can be expressed as a matrix (equation 2.3):

$$\boldsymbol{s}(t) = (\boldsymbol{X} * \boldsymbol{W})(t) \tag{2.3}$$

we express 2-D convolution (equation 2.4) as:

$$\boldsymbol{s}(i,j) = (\boldsymbol{X} * \boldsymbol{W})(i,j) = \sum_{m}\sum_{n} \boldsymbol{x}(i+m, j+n)\boldsymbol{w}(m,n) \tag{2.4}$$

The asterisk $(*)$ indicates convolution.

The function of convolutional layer is to extract features from input data, it contains multiple convolutional kernels which detect local features and map them to feature maps. The size of kernel determines receptive field. When kernels work, they will regularly do matrix multiplication and add the bias. If it is a two-dimensional convolution, the expression (equation 2.5) is:

$$\boldsymbol{Z}^{l+1}(i,j) = (\boldsymbol{Z}^{l} * \boldsymbol{w}^{l+1}) + \boldsymbol{b} = \sum_{k=1}^{K_l} \sum_{x=1}^{f} \sum_{y=1}^{f} [\boldsymbol{Z}_k^l(s_o i + x, s_0 j + y) \boldsymbol{w}_k^{l+1}(x,y)] + \boldsymbol{b} \tag{2.5}$$

where $(i,j) \in 0,1,...,L_{l+1}$, $L_{l+1} = \frac{L_l + 2p - f}{s_0} + 1$, where $\boldsymbol{b}$ is bias, $\boldsymbol{Z}_l$ and $\boldsymbol{Z}_{l+1}$ represents the input and output of layer $l+1$, which are regarded as feature map. L_l is the size of $\boldsymbol{Z}_l$ (assume width equals to height here), $\boldsymbol{Z}(i,j)$ represents the pixels of feature map, K is number of channels of feature map, f, s_0, p are kernel size, stride and padding size respectively.

- Activation Function

As the neuron is expressed as $\boldsymbol{Z}^{l+1} = \boldsymbol{Z}^{l} * \boldsymbol{w}^{l+1}$ which is linear module, overlaying of neural network layers cannot solve non-linear classification problems, the idea of activation function is to make neural network be able to solve non-linear problems. The general form of activation function can be expressed as below (equation 2.6):

$$\boldsymbol{A}_{i,j,k}^{l} = f(\boldsymbol{Z}_{i,j,k}^{l}) \tag{2.6}$$

where $f(\cdot)$ is the activation function. The wildly used activation function in CNN are ReLU (Rectified Linear Unit), Sigmoid, hyperbolic tangent and Softmax. Other variants similar to ReLU include: ReLU with slope (Leaky ReLU, LReLU), Parametric ReLU (PReLU), Randomized ReLU (RReLU), Exponential Linear Unit (ELU). Figure 2.1 shows six examples of activation function.

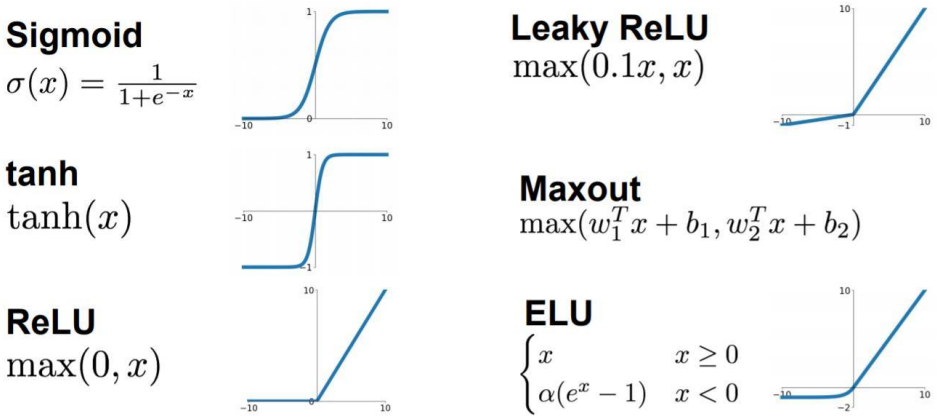


Figure 2.1: Several activation functions commonly used in deep learning.

- Pooling layer

After feature extraction process in convolutional layer, the output feature map will be sent to the pooling layer for feature selection to filter essential information, meanwhile the size of parameters could be reduced. The pooling layer contains a pre-set pooling function, it is to replace the result of a single pixel in the feature map with its neighbouring regions information. The pooling layer selects target pixels in the same ways as convolution kernels scanning feature map, which is controlled by the pooling size, stride and padding. The general expression of $L_p p$ pooling is (equation 2.7):

$$\boldsymbol{A}_k^l(i,j) = \left[\sum_{x=1}^{f} \sum_{y=1}^{f} \boldsymbol{A}_k^l(s_0 i + x, s_0 j + y)^p \right]^{\frac{1}{p}} \tag{2.7}$$

the definition of stride s_0, pixel (x,j) are same as convolutional layer. To be specific, when $p = 1$, the function takes mean value in pooling area for which it is called average pooling; when $p \to \infty$, the function takes maximum value in pooling area, for which it is called max pooling. Figure 2.2 illustrates the difference between average pooling and max pooling and how they work.

- Fully connected layer

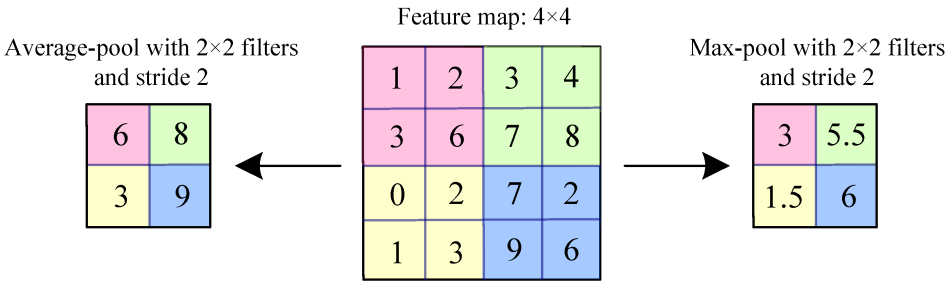


Figure 2.2: Average pooling and max pooling, the input from convolution layer is 4 × 4, pooling filter size is 2 × 2, stride is 2.

From previous steps, original data are extracted by convolutional layers, pooling layer and activation function, the information are mapped into hidden feature space, fully connected layer is to map those learned features into labelling space of samples. In fully connected layer, feature maps from last layer are flattened and turned into single vector, it takes the responsibility to to make classification decision.

Figure 2.3 is a demonstration of a convolutional neural network, it explains the previous structures introduced above.

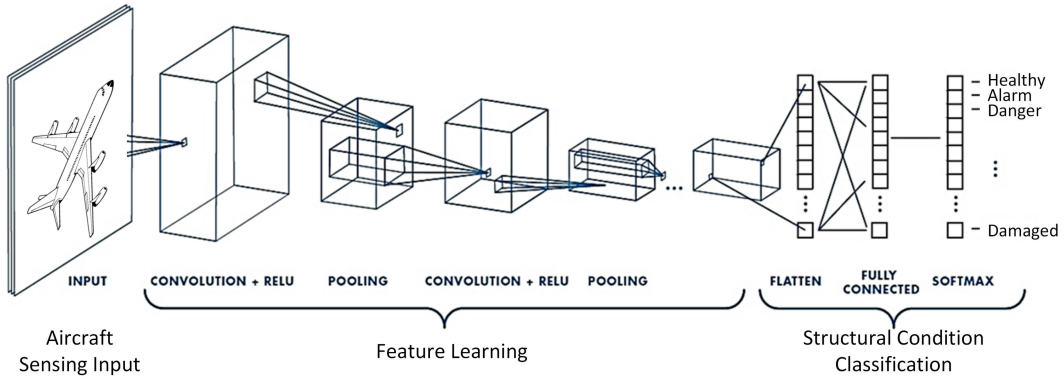


Figure 2.3: Demonstration of CNN architecture on aircraft condition classification, including convolutional layers, pooling, activation and fully connected layer.

- Backpropagation (BP)

Backpropagation algorithm is an essential constitution of deep learning algorithm to update parameters by calculating derivatives. The general expression of BP is (equation 2.8):

$$\begin{aligned}\left(\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}}\right)_{i,j}^{l} &= \sum_{k=1}^{K_l}\sum_{x=1}^{f}\sum_{y=1}^{f}\left[\boldsymbol{w}_k^{l+1}(x,y)\left(\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}}\right)_{s_0+x,s_0+y,k}^{l+1}\right] f'\left(\boldsymbol{A}_{i,j}^{l}\right) \\ \boldsymbol{w}^{l} &= \boldsymbol{w}^{l+1} - \alpha\left(\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}}\right)_k = \boldsymbol{w}^{l+1} - \alpha\left[\boldsymbol{A}^{l+1}\left(\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}}\right)_k^{l+1}\right]\end{aligned} \tag{2.8}$$

where $\boldsymbol{L}$ is loss function, f' is the derivative of activation function, α is learning rate.

2.1.2. Loss functions

Loss function is also known as cost function, it is used for evaluating the discrepancy between prediction made by deep model and true value, the essence of deep learning is basically to optimise the loss function, the training process is to minimise the loss function.The lower the loss is, the better the model performs. Some well-known loss functions are listed below:

Zero-One Loss (0-1 loss, equation 2.9):

$$L(\boldsymbol{Y}, f(\boldsymbol{X})) = \begin{cases} 1, & \boldsymbol{Y} \neq f(\boldsymbol{X}) \\ 0, & \boldsymbol{Y} = f(\boldsymbol{X}) \end{cases} \tag{2.9}$$

Perceptron Loss (equation 2.10):

$$L(\boldsymbol{Y}, f(\boldsymbol{X})) = \begin{cases} 1, & \left|\boldsymbol{Y} - f(\boldsymbol{X})\right| \geqslant t \\ 0, & \left|\boldsymbol{Y} - f(\boldsymbol{X})\right| < t \end{cases} \tag{2.10}$$

Absolute Loss (equation 2.11):

$$L(\boldsymbol{Y}, f(\boldsymbol{X})) = \left|\boldsymbol{Y} - f(\boldsymbol{X})\right| \tag{2.11}$$

Quadratic Loss (equation 2.12):

$$L(\boldsymbol{Y}, f(\boldsymbol{X})) = (\boldsymbol{Y} - f(\boldsymbol{X}))^2 \tag{2.12}$$

Log Loss (equation 2.13):

$$L(\boldsymbol{Y}, P(\boldsymbol{Y}|\boldsymbol{X})) = -logP(\boldsymbol{Y}|\boldsymbol{X}) \tag{2.13}$$

Cross-Entropy Loss (2.14):

$$C = -\frac{1}{n}\sum_{x}[y\ln a + (1-y)\ln(1-a)] \tag{2.14}$$

where n is total number of samples, x is sample, y is ground truth and a is prediction.

From Figure 2.4 we can see different tasks always apply different loss surface, corresponding model training, which is to search the minimum of loss function, to the surface, the task is to find "valley" bottom (deepest point) of these figures.

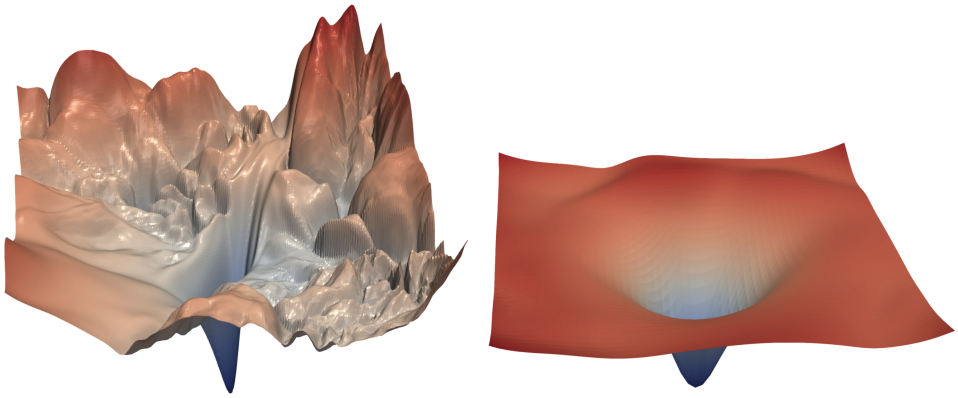


Figure 2.4: Demonstration of two loss surface with different sharpness/flatness. The left loss function may not achieve the global minimum due to the cragginess, on the contrary, the right one is more probably converged faster.

2.1.3. Regularisation

As loss function only consider the performance on train set which may lead over-fitting, to overcome this regularisation should be introduced into system. Various regularisation methods could be used in CNN algorithms to prevent over-fitting, common regularisation methods include $L-p$-norm regularisation, random inactivation (spatial dropout) and random connection Inactivation (drop connect).

L_p-norm regularisation define a loss function with a regularisation item shown in equation 2.15:

$$\tilde{L}(w; X, y) = L(w; X, y) + \lambda \sum \|w\|_p \tag{2.15}$$

where $L(w; X, y)$ is loss function, $\sum \|w\|_p$ is regularisation item, λ is regularisation parameter, when $p \geqslant 1$, regularisation item is convex function [7], and specifically, when $p = 2$, L_2-norm regularisation is called Tikhonov regularisation [2].

2.1.4. Normalisation

In deep neural network, as the input data is transmitted step by step in hidden layers, the mean and standard deviation will change, resulting in covariate shift phenomenon [9], it is considered to be one of the reasons for vanishing gradient in deep neural network.

Batch normalisation (BN) partially solves such problems at the cost of introducing additional learning parameters. Its strategy is to firstly standardise the features in hidden layers, and then use two linear parameters to amplify the standardised features as new inputs. The neural network will update BN parameters in learning process [9]. BN parameters in CNN have the same properties as the convolution kernel parameters, that is, the pixels of same channel in the feature map share a set of BN parameters. Y. Wu and K. He introduced a new method Group Normalisation (GN) [25] which normalises based on division of channels. Other normalisation methods include Layer Normalisation, (LN) [1] and Instance Normalisation (IN) [22]. Equation 2.16 shows how BN works.

$$\begin{aligned} \mu_{\mathscr{B}} &\leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma^2_{\mathscr{B}} &\leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_{\mathscr{B}})^2 \\ \hat{x}_i &\leftarrow \frac{x_i-\mu_{\mathscr{B}}}{\sqrt{\sigma^2+\epsilon}} \\ y_i &\leftarrow \gamma\hat{x}_i+\beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \end{aligned} \tag{2.16}$$

where μ is mean and σ is standard deviation, m is batch size, ϵ is to prevent invalid calculation due to $\sigma^2 = 0$, γ and β are the scale and shift parameter to be learned.

2.2. ResNet

The design of ResNet (Deep residual network) is a milestone of deep learning and CNN development, it was proposed in 2016 by Kaiming He et al [5], which won the first at ILSVRC and COCO 2015. Conventional CNNs suffer degradation problem when layers of network are increased due to the impossibility of identity mapping causing by activation functions like ReLU [18], which leads irreversible information loss. ResNet introduced residual blocks into networks which aims to learn residual information $\mathscr{F}(\mathbf{x}) = \mathscr{H}(\mathbf{x}) - \mathbf{x}$, where $\mathscr{H}(\mathbf{x}) = \mathscr{F}(\mathbf{x}) + \mathbf{x}$ is the output from a residual block, $\mathbf{x}$ is identity mapping, fitting $\mathscr{F}(\mathbf{x}) = 0$ is easier than fitting $\mathscr{H}(\mathbf{x}) = 0$ (conventionally) directly. Currently, wildly applied ResNet architecture includes ResNet-18, ResNet-34, ResNet-50, ResNet-101 and ResNet-152.

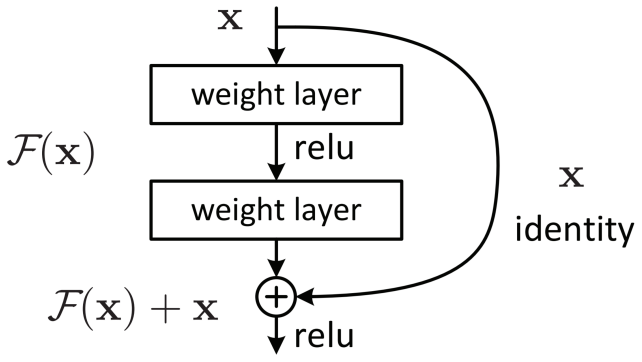
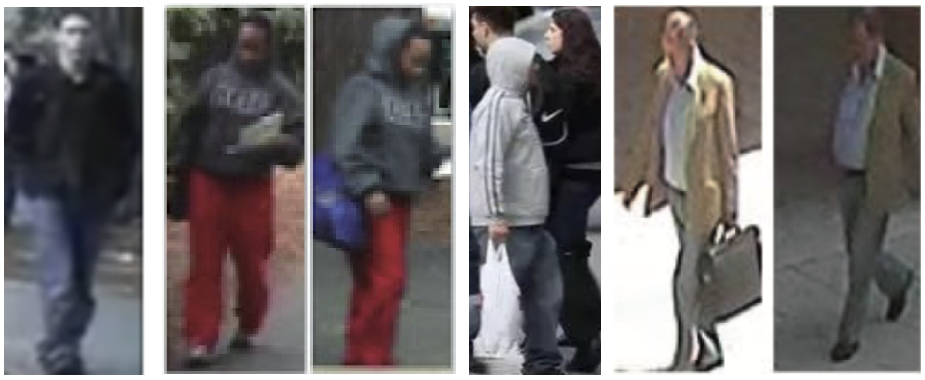


Figure 2.5: Residual learning block.

3

Person Re-Identification

Person Re-Identification is widely known as Re-Id, one important and difficult task under computer vision. It is the task to recognise person from one camera given information from the other cameras, the given images or videos are defined as train set and the people who are aimed to be recognised are regarded as test set, the probes are called query. Thus this research could also be treated as an image retrieval problem. Re-Id has great application in intelligent video monitoring, intelligent security field and others. Current Re-Id research meets issues like video/image low resolution, same people show different viewpoints, object occlusion and illumination conditions [27], examples are shown in Figure 3.1.



(a) resolution (b) viewpoint (c) occlusion (d) illumination

Figure 3.1: Four issues Re-Id researches are confronting with. (a) Due to camera situation, some Re-Id datasets contain low resolution images; (b) same person appears in variant views in dataset; (c) people in some scene are occluded by other people; (d) people are shot in different scene with different light conditions.

3.1. Methodologies

Representation learning based method [14, 27] is conventional but essential, it depends on CNN extracting features for task requirements automatically. Beyond exploiting people's ID, other attributes like dresses and appearances are analysed to improve performance. Representation learning based Re-Id methods become research baseline commonly, with comparably high robustness and stability. **Metric learning** based methods aim to learn distance between mages on feature space. The general idea is to push images with different IDs as far as possible, mean while, the positive samples should stay closer. Euclidean distance between two images I_i, I_j is defined as equation 3.1, f_i, f_j are features extracted by feed-forward neural network.

$$d_{I_i,I_j} = ||f_{I_i} - f_{I_j}|| \tag{3.1}$$

Contrastive loss (equation 3.2, $y = 0$: different ID, $y = 1$: same ID, α is hyper-parameter and $(s)_+ = max(s, 0)$) [23] applying Siamese network to operating the distances in feature space;

$$L_c = y d^2_{I_i,I_j} + (1-y)(\alpha - d_{I_i,I_j})^2_+ \tag{3.2}$$

triplet loss (equation 3.3, a: anchor, p: positive sample and n: negative sample) [19] introduced Anchor to do such task.

$$L_t = (d_{a,p} - d_{a,n} + \alpha)_+ \tag{3.3}$$

Furthermore, information such as gait, pose, partial feature and temporal sequence are attempted which also achieved good performance. Dataset size is also a research interests, GAN (Generative Adversarial Network) [28, 29] made ways to deal this problem.

3.2. Evaluation Metrics

In current research, several metrics are commonly referenced for evaluating trained model, such like mAP , Rank-n ($n \in \mathbb{N}_+, 1, 2, \cdots, 10, \cdots$) and CMC.

- mAP (Mean Average Precision)

mAP is the mean value of AP (Average Precision), firstly, Precision and Recall are defined in equation 3.4 and 3.5

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.4}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.5}$$

where TP stands for True Positive, FP stands for False Positive and FN stands for False Negative. Simply, Precision is how many samples detected are accurate, and Recall is how many accurate items have been retrieved. An example of complex dataset's PR diagram is shown in Figure 3.2, which indicates the performance of a model.

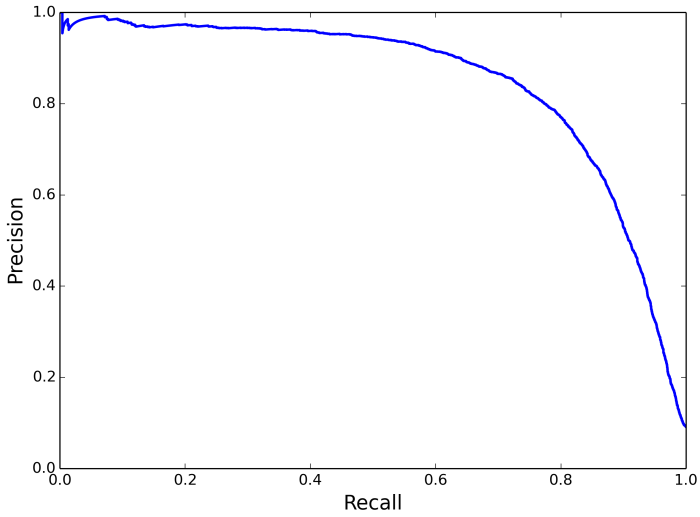


Figure 3.2: An example of PR diagram. The larger the area enclosed by the curve is, the better a model is.

Clearly, area enclosed by the curve and axis is related with model performance. The area is defined as Average Precision. From calculus we have the area is (equation 3.6):

$$AP = \int_0^1 p(r)dr \tag{3.6}$$

yet in Re-Id practice the result of PR diagram is discrete, $p(r)$ is unavailable, so it leads a discrete summation for N query results $(recall_k, precision_k), k \in 1, 2, \cdots, N$ (equation 3.7):

$$AP = \sum_{k=1}^{N} precision_k(recall_k - recall_{k-1}) \tag{3.7}$$

considering some neighbour points maintain same recall ($recall_k - recall_{k-1} = 0$), the only significant recall values have number of N_r, it is the number of images with same label with query image, equivalently, when recall is refreshed, it means a new images is correctly recognised, thus we have $\Delta recall = \frac{1}{N_r}$. So, equation 3.7 can be written in equation 3.8:

$$AP = \sum_{k} precision_k \Delta recall, k \in \Omega \tag{3.8}$$

Ω is the set of k_{th} images that have same label with query. The equation above is Average Precision for each query, thus mAP is mean value for all AP results (equation 3.9).

$$\text{mAP} = \frac{\sum_{i=1}^{N_q} \text{AP}_i}{N_q} \tag{3.9}$$

Figure 3.3: Demonstration of person Re-Id performance. From Rank-1 to Rank-10, the former a image are located, the higher probability it has that Re-Id system judged, where blue boxes indicate correct results and red ones are misclassified.

- CMC

CMC is the abbreviation of Cumulative Matching Characteristics, it reflects performance of Re-Id model. For calculating CMC, top K accuracy ($AccuK$) is defined (equation 3.10).

$$AccuK = \begin{cases} 1, \text{first K results contain target images} \\ 0, \text{first K results do not contain target images} \end{cases} \tag{3.10}$$

target images means that ones own same ID (label) with probe image. For each query, the $AccuK$ is like Heaviside step function, it is exactly a first hit problem. When it has the first matching as query image, ranked at F_{th}, $AccuK = F$, 0 goes to 1. Thus, for the system CMC is mean of summation of all $AccuK$. Based on the properties of step function, CMC must be a monotonically increasing function. Generally, when dataset is large enough, CMC curve is like Figure 3.4.

- Rank-n

Rank-n is the value on CMC curve when rank=n. For instance, Rank-1, Rank-10 are the value on CMC when rank=1 and 10.

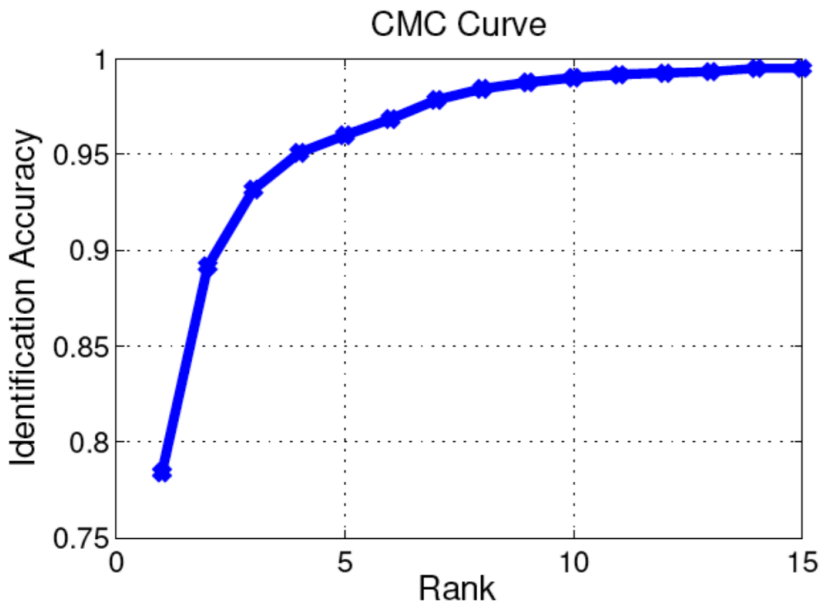


Figure 3.4: Example of CMC curve. The larger the area enclosed by curve and axis, the better the model is.

4

Transfer Learning

4.1. Transfer Learning

Transfer learning, just as its name implies, is to transfer the trained model parameters to new model to assistant model learning. Considering that some data or tasks have high relativity, it is possible to employ transfer learning to share the learned model parameters (also known as knowledge) to the new model in some way to speed up training process and optimise the learning efficiency of the model, preventing from starting from scratch.

Pan, Sinno Jialin and Yang, Qiang introduced transfer learning on [15], and in more detail, it is divided into the following directions in Figure 4.2. The mathematical definition is:

> "Given a source domain $D_S = X_S, f_S(X)$ and learning task T_S, a target domain $D_T = X_T, f_T(X)$ and learning task T_T, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$."

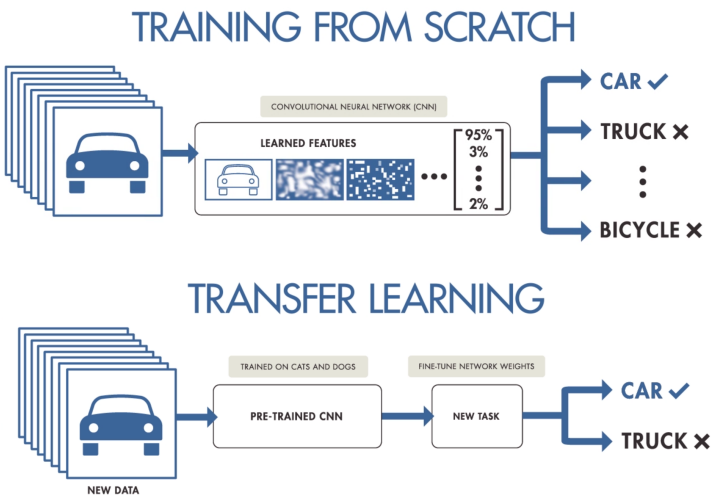


Figure 4.1: Demonstration of transfer learning.

4.1.1. Deep transfer model and fine-tune

With the development of deep learning, more and more task are solved via deep neural network, however, training a model from sketch is costly on both time and money perspective. Referring the idea of transfer learning from conventional machine learning, deep learning take actions to exploit the existed models.

The availability and principle of deep transfer learning is based on how neural network works. Take computer vision task as example, the former layers of neural networks are capable of learning low-order features such like edges and corners, then embedding to textures and shape and so on. Deep transfer learning aims to learn the

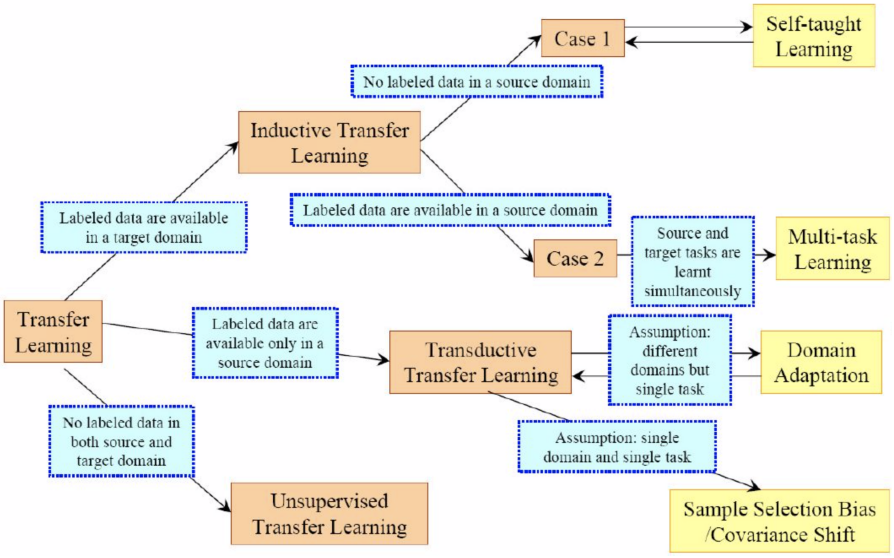


Figure 4.2: Various types of transfer learning.

object function $f_T(\cdot)$ from $f_S(X)$. The basic but common features mentioned above are reduplicate which would waste calculation resource on training them. Transfer model is to exploit those features and explore their own characteristics in latter layers. To keep the simple features, fine-tune technique initialises the task model by pre-trained model parameters obtained from the other datasets, .

In computer vision research, model pre-trained on ImageNet is deeply applied. The detailed explanation of ImageNet is located in Chapter 5.

5

Dataset Description, Processing and Coding

5.1. Re-Id Dataset

5.1.1. Market-1505

The Market-1501[26] dataset was collected on the campus of Tsinghua University, shot in the summer and published in 2015. It includes 1501 pedestrians and 32,668 detected pedestrian bounding-boxes captured by 6 cameras. Each pedestrian was captured by at least 2 cameras, and there may be multiple images in one camera. The training set has 751 people, containing 12,936 images, and each person has an average of 17.2 training data; the test set has 750 people, containing 19,732 images, and each person has an average of 26.3 test data. The pedestrian bounding box of the 3368 query images are drawn manually, while the pedestrian boxes in the gallery are detected using DPM detectors.

- **Directory structure**
  - bounding_box_test
    - 0103_c5s1_017226_03.jpg
    - 0103_c5s1_017301_02.jpg
    - 0103_c5s1_017351_03.jpg
  - bounding_box_train
    - 0118_c3s3_070394_02.jpg
    - 0118_c6s3_094192_01.jpg
    - 0118_c6s3_094117_01.jpg
  - gt_bbox
    - 0001_c1s1_001051_02.jpg
  - query
    - 0478_c3s1_135658_00.jpg
    - 0478_c4s2_051923_00.jpg
  - readme.txt
- **Directory introduction**
  - "bounding_box_test": 750 people are used in the train set, containing 19732 images.
  - "bounding_box_train": 751 people are used in the train set, containing 12,936 images.
  - "query": Select an image from each camera randomly as a query for 750 people, so there are at most 6 queries for a person, with a total of 3,368 images.

  - "gt_query": Matlab format, it is used to determine which pictures of a query are good (the images of the same person from different cameras) and which ones are bad (images of the same person from the same camera or images of different persons).
  - "gt_bbox": Manually labelled bounding box, it is used to judge whether the bounding box detected by DPM is a good box.

- **Naming rules** take 0103_c5s1_017226_03.jpg as example:
  - 0103 represents the label of each person, from 0001 to 1501;
  - c5 represents the first camera (camera 5), there are 6 cameras in total;
  - s1 represents the first video segment (sequence 1), each camera has several video segments;
  - 017226 represents the 017226^{th} picture of c5s1, the video frame rate is 25fps;
  - 01 represents the first detection frame on the frame c5s1_017226. Due to the DPM detector, several bboxes may be framed for pedestrians on a single frame. 00 means manual marking box.

5.1.2. DukeMTMC-ReId

The DukeMTMC [17] dataset is a large-scale labelled multi-target multi-camera pedestrian tracking data set collected on Duke University campus in 2014. It provides a new large-scale high-resolution video data set recorded by 8 simultaneous cameras, with more than $7,000$ single-camera tracks and more than $2,700$ independent characters. DukeMTMC-reID is a pedestrian re-identification subset of the DukeMTMC dataset, and provides manually labelled bounding box.

- **Directory structure**
  - bounding_box_test
    - 0103_c5s1_017226_03.jpg
    - 0103_c5s1_017301_02.jpg
    - 0103_c5s1_017351_03.jpg
  - bounding_box_train
    - 0118_c3s3_070394_02.jpg
    - 0118_c6s3_094192_01.jpg
    - 0118_c6s3_094117_01.jpg
  - query
    - 0478_c3s1_135658_00.jpg
    - 0478_c4s2_051923_00.jpg
  - README.md
  - CITATION_DukeMTMC.txt
  - CITATION_DukeMTMC-reID.txt
  - LICENSE_DukeMTMC.txt
  - LICENSE_DukeMTMC-reID.txt

- **Directory introduction**
  - "bounding_box_test": 702 people are used in the train set, containing $17,661$ images.
  - "bounding_box_train": 702 people are used in the train set, containing $16,522$ images.
  - "query": Select an image from each camera randomly as a query for 702 people, so there are at most 8 queries for a person, with a total of $2,228$ images.
  - "gt_query": Matlab format, it is used to determine which pictures of a query are good (the images of the same person from different cameras) and which ones are bad (images of the same person from the same camera or images of different persons).

  - "gt_bbox": Manually labelled bounding box, it is used to judge whether the bounding box detected by DPM is a good box.

- **Naming rules** take 0103_c5s1_017226_03.jpg as example:
  - 0103 represents the label of each person, from 0001 to 1501;
  - c5 represents the first camera (camera 5), there are 6 cameras in total;
  - s1 represents the first video segment (sequence 1), each camera has several video segments;
  - 017226 represents the 017226^{th} picture of c5s1, the video frame rate is 25fps;
  - 01 represents the first detection frame on the frame c5s1_017226. Due to the DPM detector, several bboxes may be framed for pedestrians on a single frame. 00 means manual marking box.

5.1.3. CUHK03

CUHK03 [13] is the first large-scale pedestrian re-identification dataset sufficient for deep learning research. The images of this dataset are collected on the campus of Chinese University of Hong Kong (CUHK). The data is stored in the `.mat` file format of `cuhk-03.mat`, containing $1,467$ different pedestrians, collected by 5 pairs of cameras.

- **Directory structure**
  - detected: 5×1 cell
    - 843×10 cell
    - 440×10 cell
    - 77×10 cell
    - 58×10 cell
    - 49×10 cell
  - labeled: 5×1 cell
    - 843×10 cell
    - 440×10 cell
    - 77×10 cell
    - 58×10 cell
    - 49×10 cell
  - testsets: 20×1 cell
    - 100×2 double matrix
  - README.md
  - CITATION_DukeMTMC.txt
  - CITATION_DukeMTMC-reID.txt
  - LICENSE_DukeMTMC.txt
  - LICENSE_DukeMTMC-reID.txt

- **Directory introduction**
  - "detected": 5×1 cells, it was labelled by machine, each cell contains a pair of photos collected by a camera group, as shown below:
    Each camera group is composed of $M \times 10$ cells, M is the pedestrian index, the first 5 columns and the last 5 columns are from different cameras in the same group. Each element in cell is a pedestrian frame image of $H \times W \times 3$ (uint8 data type), some may be vacant and an empty set.
  - "labelled": 5×1 cells, the pedestrian bounding boxes are manually labelled, the format and content are the same as "detected".
  - "query": Select an image from each camera randomly as a query for 702 people, so there are at most 8 queries for a person, with a total of $2,228$ images.

  - "testsets": 20×1 cells, test protocol, composed of 20 100×2 double type matrices (repeated 20 times). 100 × 2 double, 100 rows represent 100 test samples, the first column is the camera pair index, and the second column is the pedestrian index.

- **Naming rules** take 0103_c5s1_017226_03.jpg as example:
  - 0103 represents the label of each person, from 0001 to 1501;
  - c5 represents the first camera (camera 5), there are 6 cameras in total;
  - s1 represents the first video segment (sequence 1), each camera has several video segments;
  - 017226 represents the 017226^{th} picture of c5s1, the video frame rate is 25fps;
  - 01 represents the first detection frame on the frame c5s1_017226. Due to the DPM detector, several bboxes may be framed for pedestrians on a single frame. 00 means manual marking box.

5.1.4. PersonX_v1

PersonX_v1 [21] is a dataset designed for viewpoint research in re-Id field, the engine is generated by Unit [16]. The dataset is comprised of six backgrounds, three of them are pure colour and the rest three contain scene backgrounds. There are 1266 manually marked identities, including 547 females and 719 males, each of them are consisted of 36 images, those 36 angles apply 36 viewpoints with interval 10°, such as 10°, 20°, ..., 350° and 0°(360°). In our work, we chose the 6th subset, one with background, as the external synthetic dataset to be applied. The viewpoint description and an illustration of the subset of PersonX_v1 are shown in 5.1.

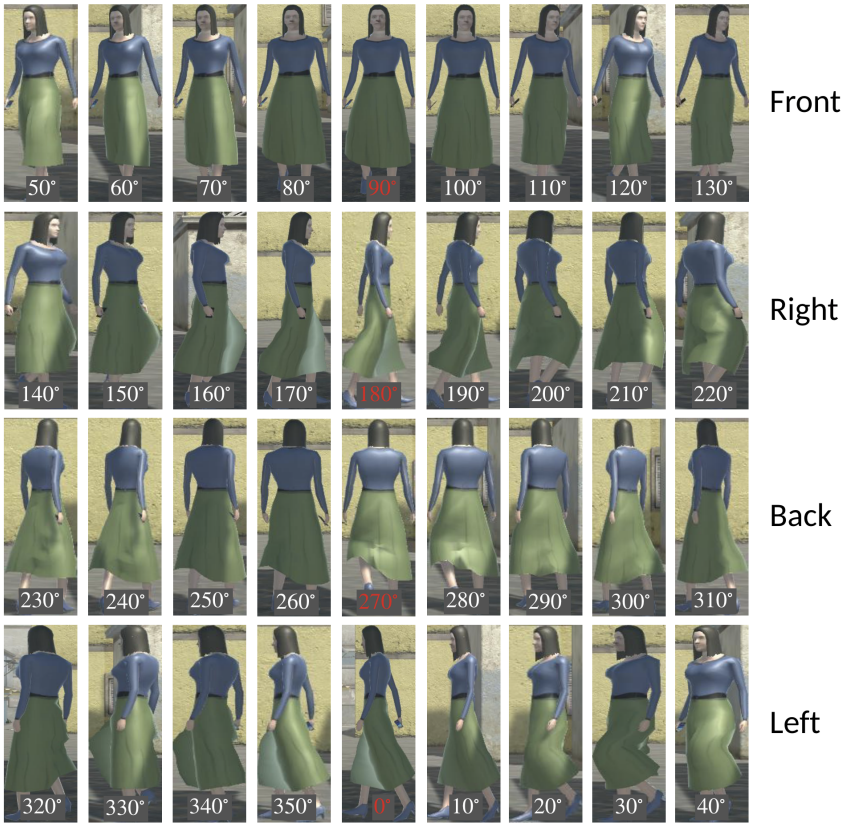


Figure 5.1: Different viewpoints of PersonX dataset. Each image of every identity in this dataset are sampled with interval 10°. Front views represent degree 50°-130°, similarly, 140°-220° stand for right, 230°-310° stand for back and 320°-40° stand for left. The angles which are marked in red are the due views of these four views respectively.

5.1.5. Imagenet

ImageNet [3] is the largest dataset of computer vision research domain. It was established by scientists from Stanford University, simulating human recognition system.

5.2. Pre-Processing

For Market-1501, DukeMTMC-reID, CUHK03 and PeronX_v1, the original data was divided into pytorch directory as pytorch/train, pytorch/val, pytorch/query and pytorch/test respectively, this operation is to load data via "Dataloader" package in Pytorch.

5.3. Coding

The experiments are mainly done based on Python, PyTorch framework are used specifically for deep learning; few data, which is stored in `.mat` file, such as CUHK03, were processed by Matlab. The data obtained during the experiment, such as data distribution, is stored in `.pyn` format.

Bibliography

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[2] M. Bertero, C. De Mol, and G. A. Viano. *The Stability of Inverse Problems*, pages 161–214. Springer Berlin Heidelberg, Berlin, Heidelberg, 1980. ISBN 978-3-642-81472-3. doi: 10.1007/978-3-642-81472-3_5. URL https://doi.org/10.1007/978-3-642-81472-3_5.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 262–275, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-88682-2.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[7] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[8] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In Anders Heyden and Fredrik Kahl, editors, *Image Analysis*, pages 91–102, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21227-7.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[12] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

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

[14] Tetsu Matsukawa and Einoshin Suzuki. Person re-identification using cnn features learned from combination of attributes. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2428–2433. IEEE, 2016.

[15] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[16] John Riccitiello. John riccitiello sets out to identify the en- gine of growth for unity technologies (interview). January, 18, 2015.

[17] 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 workshop on Benchmarking Multi-Target Tracking*, 2016.

[18] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. URL `http://arxiv.org/abs/1801.04381`.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, 2019.

[22] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016. URL `http://arxiv.org/abs/1607.08022`.

[23] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016.

[24] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. *CoRR*, abs/1711.08565, 2017. URL `http://arxiv.org/abs/1711.08565`.

[25] Yuxin Wu and Kaiming He. Group normalization. *CoRR*, abs/1803.08494, 2018. URL `http://arxiv.org/abs/1803.08494`.

[26] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[27] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[28] 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*, pages 3754–3762, 2017.

[29] 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*, pages 5157–5166, 2018.