White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/174437/

Version: Accepted Version

Article:
Li, D., Huang, L., Ye, B. et al. (2020) FSRM-STS: Cross-dataset pedestrian retrieval based on a four-stage retrieval model with Selection–Translation–Selection. Future Generation Computer Systems, 107. pp. 601-619. ISSN: 0167-739X

https://doi.org/10.1016/j.future.2020.02.028

Article available under the terms of the CC-BY-NC-ND licence (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

FSRM-STS: Cross-dataset Pedestrian Retrieval Based on a Four-stage Retrieval Model with Selection-Translation-Selection

Daifeng Li[1*], ♣Lu Huang[1], ♣Biyun Ye[1], ♣Fangbin Wan[1], Andrew Madden[1], Xingjian Liang[2]

1. School of Information and Management, Sun Yat-Sen Uniersity, Guangzhou, Guangdong, China
2. School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, Guangdong, China

*Corresponding author: lidaifeng@mail.sysu.edu.cn
♣These authors contributed equally to this work.

ABSTRACT: Pedestrian retrieval is widely used in intelligent video surveillance and is closely related to people's lives. Although pedestrian retrieval from a single dataset has improved in recent years, obstacles such as a lack of sample data, domain gaps within and between datasets (arising from factors such as variation in lighting conditions, resolution, season and background etc), reduce the generalizability of existing models. Factors such as these can act as barriers to the practical use of this technology. Cross-dataset learning is a way to obtain high-quality images from source datasets and can assist the learning of target datasets, thus helping to address the above problem. Existing studies of cross-dataset learning directly apply translated images from source datasets to target datasets, and seldom consider systematic strategies for further improving the quality of the translated images. There is therefore room for improvement in cross-dataset learning. This paper proposes a four-stage retrieval model based on Selection-Translation-Selection (FSRM-STS), to help address this problem. In the first stage of the model, images in pedestrian retrieval datasets are semantically segmented to provide information for image-translation. In the second stage, STS is proposed, based on four strategies to obtain high quality translation results from all source datasets and to generate auxiliary datasets. In the third stage, a pedestrian feature extraction model is proposed, based on both the auxiliary and target datasets. This converts each image in target datasets into an n-dimensional vector. In the final stage, the extracted image vectors are used for cross-dataset pedestrian retrieval. As the translation quality is improved, FSRM-STS achieves promising results for the cross-dataset pedestrian retrieval. Experimental results on four benchmark datasets Market-1501, DukeMTMC-reID, CUHK03 and VIPeR show the effectiveness of the proposed model. Finally, the use of parallel computing for accelerating the training speed and for realizing online applications is also discussed. A primary demo based on cloud computing is designed to verify the engineering solution in the future.

KEYWORDS
Pedestrian Retrieval, Transfer Learning, Image Translation, Semantic Segmentation, Image Retrieval

1. INTRODUCTION

With the development of artificial intelligence, traditional ways of thinking in information retrieval have been greatly challenged. Data-driven information retrieval technology is playing an increasingly important role and has significantly promoted the development of information retrieval. This paper focuses mainly on pedestrian retrieval, also known as person re-identification, which is an important image retrieval technology and also a hot topic in artificial intelligence and deep learning. It uses computer vision technology to determine whether a specific pedestrian appears in images or videos taken by different devices. The technology can be further combined with pedestrian detection/tracking technology and is now widely applied in intelligent video surveillance, security systems and other related fields. In recent years, the performance of pedestrian retrieval models on a single dataset has greatly improved. However, the results of cross-dataset pedestrian retrieval are far from ideal. In fact, cross-dataset pedestrian retrieval is a more common problem than pedestrian retrieval on a single dataset. It is usual to train a model on a labeled dataset but the challenge is to test it on another, unlabeled dataset. For example, an ideal pedestrian retrieval model is trained based on datasets sampled in Sun Yat-Sen University, then it is applied to an unlabeled dataset based on a sample from Fudan University. Experimental results show that performance is generally poor. Therefore, further research is needed. The difficulties of dealing with cross-dataset pedestrian retrieval problems mainly arise for the following reasons.

1) Lack of sample data. This is a common problem in deep learning. For example, VIPeR [1], an early pedestrian retrieval dataset, only contains 1,134 images caught by 2 cameras. The best result of pedestrian retrieval reported for this

dataset is 56.3% [2]. However, in DukeMTMC-reID [3], which has 36,411 images from 8 cameras, and is currently the biggest pedestrian retrieval dataset in the world, the pedestrian retrieval accuracy is up to 81.4% [4]. Some of the discrepancy in accuracy will be because of differences in levels of recognition difficulty between the two datasets, but the huge difference in sample size will also significantly affect the accuracy of recognition. Collecting and labeling data manually is expensive, so there are clear advantages in being able to use samples from other pedestrian retrieval datasets to improve the training of new models.

2) Differences between datasets. Large differences due to variation in lighting conditions, resolution, season, background, clothing, posture, and occlusions can be seen in Figure 1. Such variations between databases can cause performance to drop significantly when the model is trained on one dataset then tested on others. In the experiment, the accuracy of the model trained on CUHK03 [5] was 81.9% based on Open_REID model [6][7][8], but it decreased to 66.6% when tested directly on Market-1501 [9]. Differences between datasets make it difficult for pedestrian retrieval models to be applied to more than one dataset, greatly limiting the extent to which the models can be generalized.

3) Images within a dataset also differ. (see Figure 1). These variations may interfere with the judgment of the neural network, reducing the accuracy of recognition. For example, DukeMTMC-reID [3] has a lot of background differences; the resolution of Market-1501 [9] is variable, and pedestrians' postures are diverse; and in VIPeR [1] the shooting angle and lightning conditions vary a lot. Such differences between and within different datasets present huge barriers to the successful application of pedestrian retrieval between datasets.

Recent research relating to cross-dataset learning has mainly focused on reducing the gaps between datasets [10, 11, 12, 13, 14, 15]. The main idea is to design algorithms that optimally translate images from source datasets so that they match the style of images in target datasets, allowing knowledge to be exchanged between datasets. However, previous research has tended not to consider whether all images need to be translated, or whether all the translated images are suitable for different target datasets. In another aspect, current models take the images as a whole to make translation, it is assumed that pedestrian information in an image should be paid more attention, while seldom researches concern that.

To address the above problems, a four-stage retrieval model with a selection-translation-selection mechanism (FSRM-STS) is proposed. To exclude the impact of differences arising from occlusions and background variation, both between and within datasets, pedestrian images are semantically segmented so that the model would not be affected by complex backgrounds and occlusions, and would focus on the translation of pedestrian information. Experimental results show that simple concatenation of segmented images and raw images can combine the advantages of both, and can provide more useful information for image-translation among different datasets.

In the second stage of the proposed model, a selection-translation-selection mechanism (STS) is used to generate auxiliary datasets from source datasets. These help to address problems arising from insufficient data and from the high cost of labeling. Previous studies directly translate images from source datasets to a target dataset, and seldom consider strategies for selecting across datasets to retain high quality images and remove images with different distributions. To deal with this problem, four combination strategies of TrAdaBoost-based transfer learning [46] and SPGAN-based image-translation models [11] were designed. The four STS strategies generate four auxiliary datasets at this stage.

In the third stage, four pedestrian image representation models will be learned from the four auxiliary datasets. The representation model can extract features from each image and represent the features as n-dimensional vectors. In the last stage, the learned four models from the third stage are used to extract features from each image in the target datasets. These provide inputs for the pedestrian retrieval model and finally realize pedestrian retrieval tasks. The main contributions can be summarized as follows:

1) A novel four-stage retrieval model with STS is proposed to employ semantic segmentation, transfer learning and image-translation into a unified framework to overcome the problems associated with systematically reducing differences between and within pedestrian retrieval datasets. All the stages contribute to cross-dataset pedestrian retrieval.
2) A Selection-Translation-Selection mechanism with four strategies is designed to obtain high-quality images from

a source dataset for model learning of target dataset. Unlike previous studies, the research found that certain strategies lead to a significantly improved performance in the final pedestrian retrieval task on one or several specific cross-dataset combinations. STS could integrate all the four strategies into a unified framework, helping to improve the transfer of knowledge from source datasets to target datasets.

3) Extensive experiments on four benchmark datasets show that, for all 16 cross-dataset combinations, the *proposed* four-stage retrieval model with STS (FSRM-STS) outperforms the four state-of-the-art approaches significantly. Besides, the average mAP and Hit@10 obtain 5% and 6% improvement, which surpasses the seven baselines by large margins.
4) The usage of Parallel computing to accelerate the training speed and realize online applications of the proposed model is discussed. A primary demo based on a distribution system, is developed to verify the solution in the future.

Figure 1. Difference analysis between and within pedestrian retrieval datasets.

2. RELATED WORK

The following five aspects will be introduced, four-stage learning, pedestrian retrieval, cross-dataset pedestrian retrieval, semantic segmentation, transfer learning to which our work is closely related.

2.1. Multi-Stage Learning

Multi-stage retrieval models are based on the perception that each component of the model is closely connected to the other components, so that a few changes in one part will affect the other parts simultaneously. In a multi-stage retrieval learning model, the output of the formal stage is usually the input of the next stage, thus make the whole model an

indivisible entity stage by stage. It is a method widely used in fast visual object detection and ranking applications. Wang et al [16] propose two temporally constrained ranking algorithms based on class probabilistic prediction models, in which one algorithm prunes the input ranked documents and the other refines the rank order. These researches gave some inspirations for the design of our model. Lee et al [17] propose a face alignment method that uses stage-wise Gaussian process regression trees (CG-PRT) constructed by combining Gaussian process regression trees (GPRT) in a stage-wise manner. Wu et al [18] improve upon the *multi-stage* regression framework and propose the Constrained Joint stage-wise Regression Framework (CJCRF) for simultaneous facial action unit recognition and facial landmark detection. In ranking applications, *multi-stage* learning is used for achieving high top-k rank effectiveness in an efficient manner. Liu et al [19] present novel research through studying the design and deployment of a *multi-stage* model in a Large-scale Operational E-commerce Search application, which deals with hundreds of millions of user queries per day with hundreds of servers.

2.2. Pedestrian Retrieval

Research of pedestrian retrieval can be divided into two directions, hand-crafted and model-based deep pedestrian retrieval. Handcrafted methods consist of finding unique feature extraction methods and metric learning models. The best known pedestrian feature extraction methods include color histogram [20, 21], spindle histogram [22] etc. Representative works on metrics learning [23, 24, 25, 26, 27, 28] are also proposed. Since Krizhevsky [29] won the ILSVRC-2012 competition, the CNN-based deep learning model has become popular. In recent years, some representative deep pedestrian retrieval models have included image slice [30], human body recognition [31, 32, 33, 34], a combination of long-term and short-term memory networks [35, 36, 63], and GAN to generate more samples [37, 38, 39] etc. Based on the existing researches, plenty of innovations have been proposed in recent years to continuously improve the performance on benchmark datasets. For example, Xiao et al [6][7] presented a pipeline for learning deep feature representations from multiple domains with CNNs, and proposed a domain guided dropout algorithm to improve the feature learning procedure. Lin et al [40] proposed an attribute-person recognition (APR) network, which learnt a re-ID embedding and at the same time predicted pedestrian attributes. Zheng et al [4] sought to improve learned re-id embeddings by better leveraging the generated data. They proposed a joint learning framework that couples re-id learning and data generation end-to-end to generate high-quality cross-id composed images. Unlike the above pedestrian retrieval models which focus on improving the performance on a single dataset, the research is aiming to solve a cross-dataset pedestrian retrieval problem.

2.3. Cross-dataset Pedestrian Retrieval

Due to the unique features of Pedestrian Retrieval task, data labeling is time consuming compared with other computer vision tasks. Thus, utilizing image transfer or translation to obtain new training data from cross-datasets is important. Zheng et al [41] designed a verification-identification model to combine two tasks: multi-classification and relevance identification into a unified framework, and proposed a Siamese Network. Based on the verification-identification framework, Geng et al [2] developed a two-stepped fine-tuning strategy to transfer knowledge from auxiliary datasets, and proposed a novel unsupervised deep transfer learning model based on co-training. Tian et al [10] proposed an unsupervised transfer learning model to bridge inter-dataset bias and intra-dataset difference via a proposed imitate Model simultaneously, and adopted a dual classification loss to learn a discriminative representation across domains. Deng et al [11] designed SPGAN, which considers both self-similarity and domain-dissimilarity based on cycleGAN [42, 43] to translate images between different benchmark datasets. Zhu et al [43] proposed a GAN based model for the situation of unpaired training data does not exist, and realized collection style transfer, object transfiguration, season transfer, photo enhancement, etc. Wang et al [13] developed a novel deep learning method for transferring the labelled information of an existing dataset to a new unseen (unlabeled) target domain for person re-id without any supervised learning in the target domain. Liang et al [15] proposed a novel Many-to-Many Generative Adversarial Transfer Learning method (M2M-GAN) that takes multiple source sub-domains and multiple target sub-domains into consideration and performs each sub-domain, transferring mapping from the source domain to the target domain in a unified optimization process.

Though the above research obtained significant improvements on common benchmark datasets, the quality of

translated images could be further improved. It is found that for a certain translation algorithm, take SPGAN as an example, the performance on different cross-datasets are quite different, which will significantly influence the accuracy of the model on all benchmark datasets. Thus, unlike previous studies, the pedestrian retrieval task is treated as a four-stage optimization process, where each stage is assigned a sub-task to solve a certain problem which still exists in current research. In order to obtain a higher quality of auxiliary dataset, the research focus mainly on how to make use of the advantages of existing transfer learning and image translation methods by incorporating them into a unified framework. Four strategies are designed based on different ensemble modes of transfer learning and image translation models; each strategy could obtain higher quality auxiliary datasets on certain cross-dataset combinations. The research finds that by combining the four strategies into a unified optimization process, the generalizability of the proposed model is significantly improved on all cross-dataset combinations.

2.4. Transfer Learning

Transfer learning is used to transfer knowledge from domain A to domain B to improve the performance in domain B, which saves a lot of time labeling data in domain B. In addition, training a deep neural network with random weights is difficult. It requires large-scale datasets and can take a long time for the model to converge. It is therefore common to start with pre-trained weights instead of random weights [44, 45]. Because of these advantages, transfer learning is widely used in deep learning. The instance-based transfer learning is utilized in the research to better solve the problem of insufficient data. Regarding instance-based transfer learning methods, Dai et al [46] proposed an enhancement algorithm, TrAdaBoost, which is an extension of the AdaBoost algorithm. In TrAdaBoost, Boosting method is used to establish a weight adjustment mechanism to increase the weight of effective samples and to decrease the weights of samples with different distributions. Jiang et al [47] raised a method to remove miss-leading training examples from the source domain based on the difference between the conditional probabilities $P(y_T \mid x_T)$ and $P(y_S \mid x_S)$. TrAdaBoost has been widely applied in different domains [48][49][50], and exhibit better performance in cross-domain image classification.

A lot of pedestrian retrieval research also introduces transfer learning methods. Ma et al [51] put forward a new multitasking maximum folding metric learning model for person reidentification in a camera network. Pent et al [52] came up with an unsupervised transfer learning method using multitasking dictionaries, which can convert labeled individuals in the source dataset into unlabeled individuals in the target dataset, and these individuals never appear in the target dataset. Most pedestrian retrieval models which adopt transfer learning methods are based on fine-tuning. An instance-based transfer learning method is incorporated with image-translation algorithms, which effectively compensates for the lack of data in pedestrian retrieval, reduces the cost of manually labeling data samples, and improves the ability of knowledge models between datasets by transferring images with higher quality.

3. METHODOLOGIES

3.1. Model Framework

As is shown in Figure 2, the proposed model consists of four stages. There are, respectively, a semantic segmentation stage (Stage I), a selection-translation-selection stage (Stage II), a pedestrian image representation stage (Stage III) and a pedestrian retrieval stage (Stage IV).

STAGE I. Semantic Segmentation: The semantic segmentation model is used to separate pedestrians from their background and to concatenate the segmented image with the raw image to form new source datasets. The purpose of applying semantic segmentation is to reduce the influence of noise and environmental factors such as lighting conditions, resolution, seasonal variation and background etc., which may reduce the generalizability of existing models, and mainly focus on pedestrians' information. Suppose both source and target datasets contain training data that includes M and N persons described by $D_S=\{I_K,y_k\}_{k=1}^{M}$ and $D_T=\{I_K,y_k\}_{k=1}^{M}$, in which I_K and y_k represent the images of the kth person and ID respectively. The semantic segmentation model could get a segmented image from I_K as $I_k^{'}$ then

concatenate the two images to formulate a new image dataset as: $D_S=\{cat(I_k,I_k^{'}),y_k\}_{k=1}^{M}$ and $D_T=\{cat(I_k,I_k^{'}),y_k\}_{k=1}^{M}$.

STAGE II. Selection-Translation-Selection: The second stage uses a selection-translation-selection mechanism. Auxiliary datasets are used to support a source dataset training model and to address the problem of insufficient data and the high cost of labeling data in pedestrian retrieval, and to improve generalizability among different benchmark datasets. The basic functional components of the selection-translation-selection mechanism include an instance-based transfer learning model and an image-translation model. In our research, transfer learning is mainly used to filter and select high quality images in an auxiliary dataset or a translated dataset, the images of which is obtained through image-translation model. TrAdaBoost [46] is chosen as a transfer learning solution. The image-translation model was adopted to translate images in the auxiliary dataset into the same style as those in the source dataset. SPGAN [11] is chosen as an image-translation solution. In the selection-translation-selection mechanism, four strategies are defined for combining both transfer learning and image-translation. These are TrAdaBoost (STS1), SPGAN (STS2), TrAdaBoost + SPGAN (STS3), TrAdaBoost + SPGAN + TrAdaBoost (STS4). These strategies are integrated into a uniform framework to enhance the advantages of each strategy on all source and target dataset combinations (ex. CUHK03 -> DukeMTMC-reID, Market-1501 -> CUHK03). In the second stage, the four strategies generate four auxiliary datasets A1, A2, A3, A4 from the source dataset.

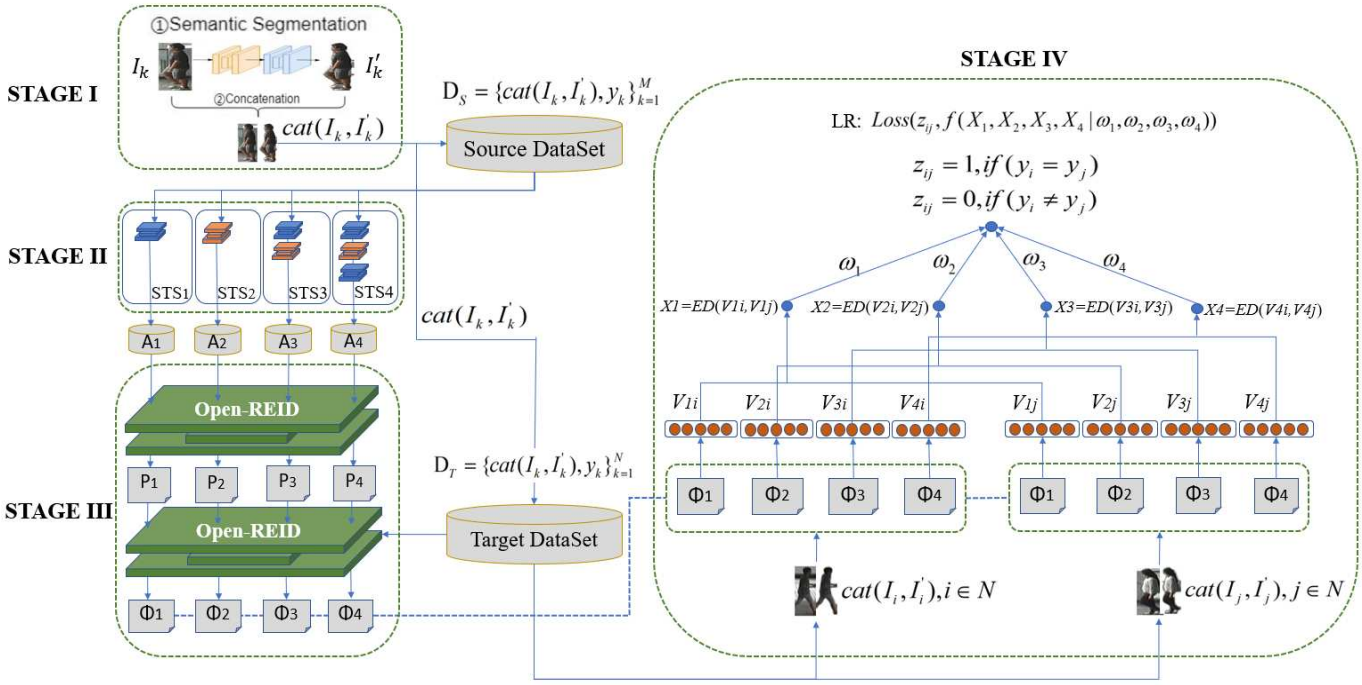


Figure 2. The framework of the proposed model.

STAGE III. Pedestrian Image Representation: The pedestrian Image Representation layer is used to train a cross-dataset pedestrian identification model for person image feature extraction using a n-dimension vector. The pre-trained models P1, P2, P3, P4 can be gotten from auxiliary datasets A1, A2, A3, A4 which have removed different distribution samples and contained new translated images based on different Selection-Translation-Selection strategies. Since the pre-trained models have already obtained some useful knowledge from the auxiliary dataset, we can continue training them on source datasets to get representation models Φ1, Φ2, Φ3 and Φ4 of each image $cat(I_k,I_k^{'})$.

STAGE IV. Pedestrian Retrieval: In this layer, a retrieval model is trained to realize Pedestrian Retrieval. We build training image pairs from target dataset for the model, two person images $cat(I_i,I_i^{'})$ and $cat(I_j,I_i^{'})$ are generated as a training sample, then the model could get their n-dimension vector representation $(V_{1i},V_{2i},V_{3i},V_{4i})$ and $(V_{1j},V_{2j},V_{3j},V_{4j})$ respectively (Each Φs generates a n-dimension vector Vsk towards image k. ex. $V_{1k}=\phi(cat(I_k,I_k')|P1)$, $V_{2k}=\phi(cat(I_k,I_k')|P2)$, $V_{3k}=\phi(cat(I_k,I_k')|P3)$, $V_{4k}=\phi(cat(I_k,I_k')|P4)$); after that, for each vector pair, the model will

calculate the vector Euclidean distance X1=ED(V1i,V1j), X2=ED(V2i,V2j), X3=ED(V3i,V3j) and X4=ED(V4i,V4j) between them. At the last step, a logistic regression model f is introduced to assign weight $\omega_1, \omega_2, \omega_3, \omega_4$ to X1, X2, X3 and X4.The loss function is defined as: $Loss(z_{ij}, f(X_1, X_2, X_3, X_4 \mid \omega_1, \omega_2, \omega_3, \omega_4))$, where z_{ij} =1, if $y_i = y_j$; $z_{ij} = 0$, if $y_i \neq y_j$.

3.2. Semantic Segmentation

In the research, Semantic Segmentation is used to extract person features from images. Given a 256*256 image, the pixel feature of which is described as { pixel 1 , pixel 2 , … , pixel 256∗256 }, a fully convolutional neural network is proposed to convert the fully connected layers of CNN into multi-channel convolutional layers. Thus, it can calculate the classification possibility of each pixel, thus the output is {y_1, y_2, …, $y_{256*256}$}, y_i in {0, 1 }, y_i = 1 means that the ith pixel feature of the image belongs to person object. In order to further improve the performance of semantic segmentation, Atrous Convolution is used in intensive forecasting tasks. Atrous Convolution allows a deep convolutional neural network to control the resolution at feature level, it can also combine more contextual information without increasing the number of parameters or the amount of computation, effectively expanding the view of filters. Atrous Spatial Pyramid Pooling (ASPP) is also proposed to segment objects more efficiently on multiple scales. ASPP probes multiple convolutional layers with filters using different sampling rates and effective fields, so that it can detect objects and image context at multiple scales. Furthermore, to improve the accuracy of the edge of the positioning object, the model combines a deep convolutional neural network (DCNN) and a probability map model. In general, the largest pooling and unsampled methods used in DCNNs affect positioning accuracy. To overcome this problem, the proposed model combines the final DCNN layer with a fully connected Conditional Random Field (CRF). The method shows qualitative and quantitative performance improvement.

3.3. TrAdaBoost based Transfer Learning

To address the problem of insufficient data, TrAdaBoost, an instance-based transfer learning algorithm, is utilized to reuse some data in the auxiliary datasets. TrAdaBoost employs a boosting method which uses a small amount of labeled data to establish a mechanism that increases the weight of valid data and reduces the weight of invalid data. This effectively expands the volume of data for model training. In recent researches, TrAdaBoost is proved to have good performance on image filtering [48][49][50]. Before applying TrAdaBoost, it was necessary to adapt the algorithm, as described in Algorithm 1.

Ts denotes a labeled source dataset, the images of which have the same distribution, Tt represents a labeled target dataset, the images of which have the same distribution. The distribution between Tt and Ts are different, and T is a dataset created by merging of Tt and Ts. m, n, m + n are the sizes of Tt, Ts and T respectively. hz is a classifier trained using T (T = {x_1, x_2, …, x_i, …, x_{m+n}}), and the number of iterations is N. ω_i^z is the weight of the ith data in the zth iteration. p_z represents the weight distribution of training data on T in the zth iteration. h_z (xi) represents the predicted value of data xi, and c(xi) is the actual value of data xi (whether xi is in target dataset Tt). β_1 and β_2 are the weights of samples having similar distributions and different distributions respectively.

The goal is to train a classifier h_z: X → Y (Y ∈ {0, 1}) to minimize the prediction error on the labeled dataset Tt, given a small number of labeled training data Ts from a source dataset, and a large amount of labeled training data Tt from a target dataset.

If the predicted value Y of a sample data is 0, it is a sample with different distribution from Ts, so its weight is reduced by reducing the weight of β2 (Formula 4). In the next loop, samples with different distributions will affect the learning process less than the current round. Samples with greater training weights β1 (Formula 4) will help the learning algorithm train a better classifier.

Here two pedestrian retrieval datasets (the target dataset and the source dataset) are utilized after they have been semantically segmented in the first stage. Following training, a classifier is generated which can remove the data with different distributions in the source dataset and retain all the images with similar distributions in an auxiliary dataset. The

auxiliary dataset and target dataset will be further used to train the pedestrian retrieval model.

Algorithm 1. The Description of TrAdaBoost
Ts: source dataset.
Tt: target dataset.

INPUT:

1. Two labeled training datasets Ts and Tt, with n and m data respectively.
2. Combine Ts and Tt to obtain T = {x_1, x_2, …, x_i, …, x_{m+n}}.
3. A basic classifier, h_l: X↦Y (Y∈ {0, 1}). Here gradient decision tree is utilized.
4. Iteration number: N.

INITIALIZE:

1. Initialize the weight of each data in T: $\omega^1 = (\omega_1^1, \ldots\ldots, \omega_{n+m}^1)$,其中，

$$\omega_i^1 = \begin{cases} 1/n, & When\ i = (1, \ldots\ldots, n) \\ 1/m, & When\ i = (n+1, \ldots\ldots, m) \end{cases} \quad (1)$$

LOOP FOR: z= 1, … …, N iteration.

1. Set p^z as:

$$\omega^z = \omega^z / \sum_{i=1}^{n+m} \omega_i^z \quad (2)$$

2. Get a weak classifier h_z: X ↦ Y, based on p^z distribution of T.
3. Calculate the error rate ε_z of h_z on T_t:

$$\varepsilon_z = \sum_{i=n+1}^{n+m} \frac{\omega_i^z |h_z(x_i) - c(x_i)|}{\sum_{i=n+1}^{n+m} \omega_i^z} \quad (3)$$

4. Set $\beta_1 = \varepsilon_z/(1-\varepsilon_z)$, and $\beta_2 = 1/\left(1+\sqrt{2\ln n/N}\right)$
5. Set updated weight vector as：

$$\omega_i^{z+1} = f(x_i) = \begin{cases} \omega_i^z \beta_2^{|h_z(x_i)-c(x_i)|}, & i = 1, \ldots\ldots, n \\ \omega_i^z \beta_1^{-|h_z(x_i)-c(x_i)|}, & i = n+1, \ldots\ldots m \end{cases} \quad (4)$$

OUPUT: Final classifier h_N to filter images from Ts (From 1 to n).

$$h_N(x_i) = \begin{cases} 1, & \omega_i^N > 0,\ i = 1, \ldots\ldots n \\ 0, & \omega_i^N < 0,\ i = 1, \ldots\ldots n \end{cases} \quad (5)$$

3.4. SPGAN based Image Translation

In a given image domain, images were collected under the same conditions, for example by the same camera; using the same device configuration; at the same time etc. The images therefore exhibit a unique style. Different image domains have different image styles. Image translation changes the style of an image from one domain, to make it consistent with the style of another domain. to overcome problems arising from lack of training data in a sub-area.

In the research, SPGAN (Similarity Preserving Generative Adversarial Networks) is used for image translation [11]. SPGAN is an optimized version of cycleGAN [42, 43]. CycleGAN introduces two generator-discriminator pairs, {G_T, D_T} and {G_S, D_S}, which map a sample from one domain onto a second domain and produce a sample that is indistinguishable from those in the second domain. For generators G_T and G_S, and their associated discriminators D_T and D_S, the adversarial loss is:

$$L_T(G_T, D_T, t, s) = E_{t\sim P(T)}[(D_T(t)-1)^2] + E_{s\sim P(S)}[(D_T(G_T(s))^2] \quad (6)$$

$$L_S(G_S, D_S, t, s) = E_{s\sim P(S)}[(D_S(s)-1)^2] + E_{t\sim P(T)}[(D_S(G_S(t))^2] \quad (7)$$

where t and s are sample images from target and source domain dataset, and their distribution are P(T) and P(S) respectively. CycleGAN also introduce a cycle-consistent loss, which is used to handle the lack of paired training data. The loss function is shown as below:

$$L_{cycle}(G_T, G_S) = E_{t\sim P(T)}[\| G_S(G_T(t)) - t \|_1] + E_{s\sim P(S)}[\| G_T(G_S(s)) - s \|_1] \quad (8)$$

Researchers have found that using only L_T, L_S and L_{cycle} can lead to the generation of unreal images [97]. To prevent this, a target domain identity constraint was introduced:

$$L_{tdi}(G_T, G_S, t, s) = E_{t\sim P(T)} \| G_S(t) - t \|_1 + E_{s\sim P(s)} \| G_T(s) - s \|_1 \quad (9)$$

SPGAN integrates a SiaNet with cycleGAN to learn a latent space, which could constrain the learning of mapping function. The contrastive loss of SiaNet could be seen in formula (10):

$$\boldsymbol{L}_{con}(\boldsymbol{i}, x_1, x_2) = (1 - i) \times \{\max(0, m - d(x_1, x_2))\}^2 + \boldsymbol{i} \times d^2(x_1, x_2) \quad (10)$$

where x_1 and x_2 are vector representations of two images, $i \in \{0, 1\}$ is the label to indicate whether x_1 and x_2 are the same person (positive) or not (negative). $d(x_1, x_2)$ denotes Euclidean distance between x_1 and x_2, m is a parameter to control the weight between positive and negative pairs. The sampling of pairs is based on an unsupervised manner [11].

Thus the over-all objective function of SPGAN could be seen in formula (11):

$$\boldsymbol{L_{total}} = \boldsymbol{\lambda_1} \times \boldsymbol{L_T} + \boldsymbol{\lambda_2} \times \boldsymbol{L_S} + \boldsymbol{\lambda_3} \times \boldsymbol{L_{cycle}} + \boldsymbol{\lambda_4} \times \boldsymbol{L_{tdi}} + \boldsymbol{\lambda_5} \times \boldsymbol{L_{con}} \quad (11)$$

where $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ are parameter weights of loss functions L_T, L_S, L_{cycle}, L_{tdl} and L_{con}. For the generators, three convolutional layers is used for encoding, two deconvolutional layers and a convolutional layer is used for decoding, and 6 resNet layers for transformation. For the discriminators, the framework was constructed with 5 convolutional layers.

3.5. Strategies based Selection-Translation-Selection

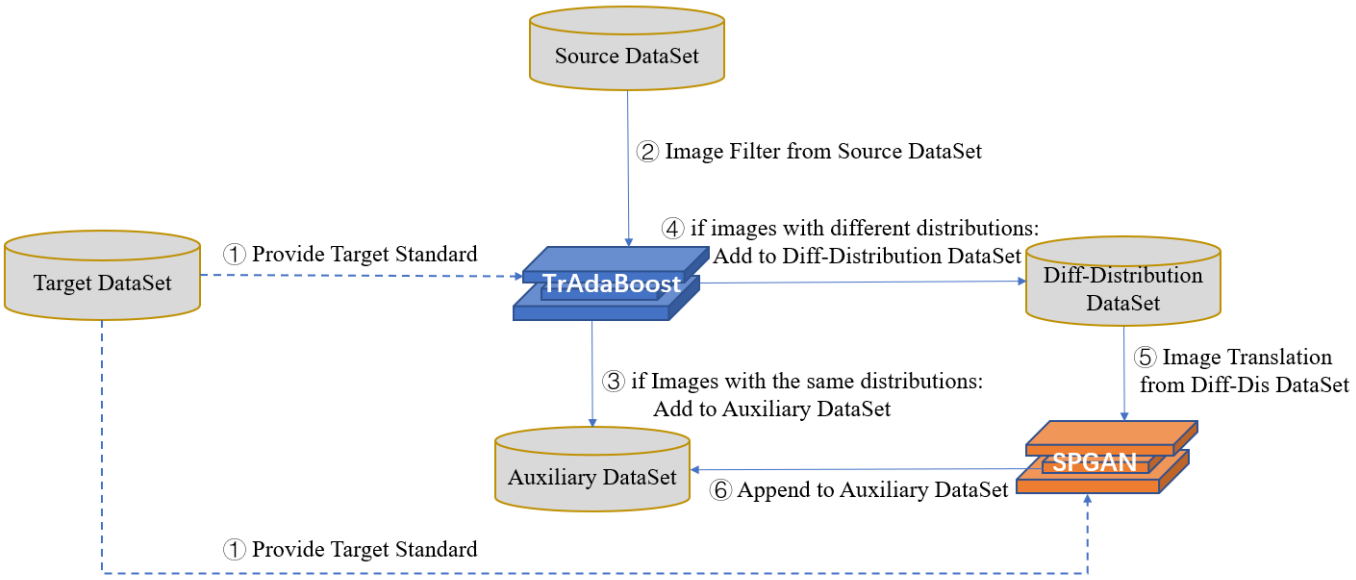
In the primary experiment, it is found that use of SPGAN alone for image-translation in multiple stage model could not obtain an expected performance (The improvement on some datasets is not significant compared with direct cross-dataset learning). One reason is that there are already images in the source dataset that satisfy the image distributions in the target dataset, so image-translation adds unexpected noises to translated images. Another is that there is a number of bad cases in the translated images, such as low quality, incomplete images with inconsistent distributions. To deal with these problems, TrAdaBoost is combined with SPGAN. It is found that different combination strategies had differing effects on different cross-dataset combinations. To obtain an optimal model, which performed well on all cross-dataset combination, a Selection-Translation-Selection (STS) mechanism is designed to improve the quality of the translated images. STS used the four strategies, STS1, STS2, STS3 and STS4, described below:

STS_1 uses only TrAdaBoost to identify and remain high-quality (similar distribution with images in target dataset) images from a source dataset as auxiliary dataset.

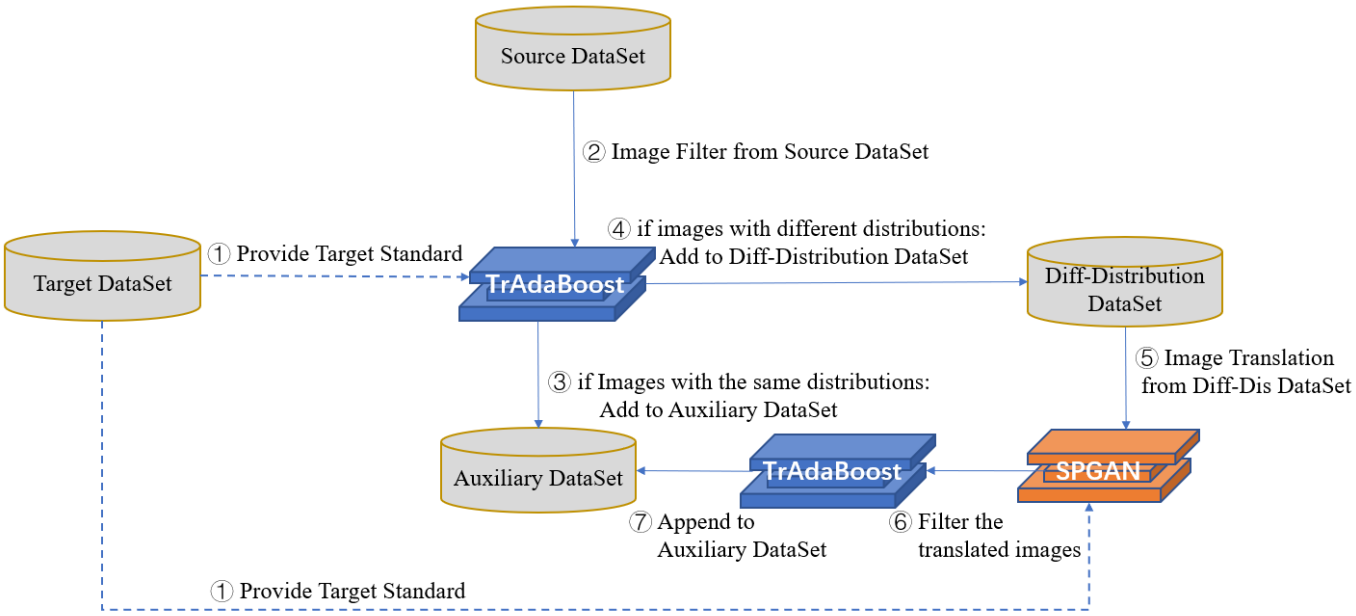
STS_2 uses SPGAN to translate images from a source dataset into new images with the same feature distribution as those of target datasets, and the new images are collected as auxiliary dataset.

STS_3 uses a combination of TrAdaBoost and SPGAN (TrAdaBoost-SPGAN) and is summarized in Figure 3 (a). The strategy has six steps. Step 1 - the target dataset provides images to both TrAdaBoost and SPGAN as target standard. Steps 2 & 3 - TrAdaBoost filters high-quality images from the source dataset to generate the primary auxiliary dataset. Step 4 - TrAdaBoost collects images with different distributions from source dataset to generate Diff-Distribution Dataset. Step 5 - SPGAN is used to translate images in Diff-Distribution Dataset into new images which match the style of the target dataset. Step 6 - the new images are added to the final auxiliary dataset.

STS_4 adds a new TrAdaBoost based on STS_3 (TrAdaBoost-SPGAN-TrAdaBoost) and is summarized in Figure 3 (b). Images translated by SPGAN in STS_3 were found to contain noise. The addition of a new TrAdaBoost filtered the translated images (the sixth step in Figure 3 (b)) and helped to reduce this problem.



(a) STS3 (TrAdaBoost + SPGAN)



(b) STS4 (TrAdaBoost + SPGAN + TrAdaBoost)

Figure 3. The Framework of STS3 and STS4.

3.6. Two-step Open_REID based Pedestrian Feature Extraction

Pedestrian retrieval mainly solves problems of the following kind. For a training dataset that includes N persons described by $D=\{cat(I_k, I_k^{'}), y_k\}_{k=1}^{N}$, where I_k and y_k represent images of the k -th person and ID respectively, $I_k^{'}$ is the segmented image of I_k. In the training stage, the feature extraction function ϕ is obtained. A given image can be described by a feature vector using $V_k = \phi\left(cat\left(I_k, I_k^{'}\right)\right)$. During the testing stage, given a pair of images {Ii ,Ij}, where Ii ,I j∈ D. The model determines whether yi is equal to yj by calculating the vector Euclidean distance between yi and yj. In the research, Open_REID (https://github.com/Cysu/open-reid) model is used to extract Pedestrian features. Open_REID

mainly uses ResNet50 and triple loss to realize Pedestrian feature extraction. Unlike previous studies, there are two steps for using Open_REID. Firstly, it is applied to get four pre-trained models (P1, P2, P3, P4) from four auxiliary datasets (A1, A2, A3, A4) obtained using the STS mechanism. Secondly, Open_REID is trained on the target dataset with four different pre-trained models. This provides four feature extraction models for image $\mathrm{cat}(\mathrm{I}_k, \mathrm{I}_k^{'})$, which are Φ1, Φ2, Φ3 and Φ4. These four models help to increase the accuracy of the Pedestrian Retrieval stage.

4. EXPERIMENTS

4.1. Dataset and Evaluation Metrics

The proposed model is evaluated using four large-scale pedestrian retrieval datasets which have significant differences in lighting conditions, background, shooting angles and so on (see Table 1).

Market-1501 [9] was sampled in Tsinghua University. A total of six cameras, including five high-resolution cameras and one low-resolution camera, were used during the sampling process. The dataset contains 32,668 bounding boxes, of which 19,732 are gallery images, 3368 are query images and 12,936 are training images. In total, there are 1,501 different identities in the dataset, of which 751 are in the training dataset and 750 are in the test dataset.

CUHK03 [5] is composed of 1,467 identities and 28,192 bounding boxes. 26,264 bounding boxes of 1,367 identities are used for training and 1,928 bounding boxes of 100 identities for testing.

DukeMTMC-reID [3], one of the largest pedestrian retrieval datasets is a subset of the DukeMTMC dataset, consisting of 36,411 bounding boxes with 1,812 identities captured by 8 high-resolution cameras. 16,522 bounding boxes of 702 identities are used for training and the remaining serve as a test dataset. This dataset is very challenging because many pedestrians are wearing similar clothes and may be blocked by cars or trees.

VIPeR [1] has a variety of shooting angles and lighting conditions. Compared to the other datasets, VIPeR is quite small: it contains only 1,134 bounding boxes captured by 2 cameras. In the experiment, 200 identities are used for testing.

Evaluation Metrics Mean Average Precision (mAP) and Rank-1, Rank-5, Rank-10 accuracy are the criteria for measuring the quality of the model. mAP refers to the average accuracy rate of each relevant document retrieved by a query, which is an objective evaluation of the model. For a query, the Rank- i accuracy refers to whether the first i query results contain the correct image. If the correct matching image appears, Rank- i equals 1, otherwise, Rank- i equals 0. When there are multiple queries, Rank-i takes the mean value.

Table 1. The Description of four datasets

Dataset	Market-1501	CUHK03	DukeMTMC-reID	VIPeR
Bounding Boxes	32,668	28,192	36,411	1,264
Identities	1,501	1,467	1,812	632
Cameras	6	2	8	2
Scene	outdoors	indoors	outdoors	outdoors

4.2. Experiment Setup

Seven models from the cross-dataset image retrieval domain were selected as baselines. A description of each is given below:

- **Basic model I (BM-I)**: Images from the source dataset is trained by Open_REID (https://github.com/Cysu/open-reid) to provide a pre-trained model. The pre-trained model is directly used to train images from the target dataset.
- **Basic model II (BM-II)**: The proposed semantic segmentation model is incorporated in BM-I.
- **Person_REID (PRID)**: Images from the source dataset are trained by Person_REID to generate a pre-trained model, which is directly applied to train images from the target dataset. Person_REID [40] is a state-of-the-art model, which proposes a ResNet50 to learn a re-ID embedding and which predicts the pedestrian attributes simultaneously. This

multi-task method integrates an ID classification loss and a number of attribute classification losses, and back-propagates the weighted sum of the individual losses.

- **TrAdaBoost-based cross-dataset Pedestrian Retrieval (FSCM-STS$_1$)**: Images from the source dataset are filtered by TrAdaBoost to generate an auxiliary dataset (Strategy STS$_1$). Open_REID is then applied to obtain a pre-trained model through the auxiliary dataset. This model is then used to train images from the target dataset.
- **SPGAN based cross-dataset Pedestrian Retrieval (FSCM-STS$_2$)**: Images from the source dataset are translated by SPGAN. Firstly, an auxiliary dataset is generated(Strategy STS$_2$), then Open_REID is applied to obtain a pre-trained model through the auxiliary dataset. Finally, the pre-trained model is used to train images from the target dataset.
- **TrAdaBoost + SPGAN-based cross-dataset Pedestrian Retrieval (FSCM-STS$_3$)**: Images from the source dataset are filtered by TrAdaBoost, and the high-quality images are collected to form a primary auxiliary dataset. The remaining images, with different distributions, are processed by SPGAN to make style translations, and then appended to the auxiliary dataset (Strategy STS$_3$). Finally, the auxiliary dataset is used to help training of the target dataset by Open_REID.
- **TrAdaBoost + SPGAN + TrAdaBoost-based cross-dataset Pedestrian Retrieval (FSCM-STS$_4$)**: This model follows the same process as the previous one, up to the point where images are processed by SPGAN to make style translations, after which, they are again filtered by TrAdaBoost. Only translated images with the same distribution as images in the target dataset will be appended to the auxiliary dataset (Strategy STS$_4$). Finally, the auxiliary dataset is used to help with training of the target dataset by Open_REID.

The model proposed in this manuscript is a four-stage retrieval model with Selection-Translation-Selection mechanism (FSCM-STS). An experiment is also designed to obtain the optimal parameter sets for all models. The summary of all parameters can be seen in Table 2.

For **Semantic Segmentation**, Deeplab v2 [53] is adopted. It is a DCNN (Deep Convolutional Neural Network) + CRF (Conditional Random Field) model with ASPP (Atrous Spatial Pyramid Pooling). The learning rate is 2.5 * e^{-4}, and Momentum is set at 0.9.

For **SPGAN**, the parameter configuration is the same as the architecture released by its authors. an Experiment is also designed to obtain its optimal parameter sets (λ_1, λ_2, λ_3, λ_4) on different cross-datasets.

For **Person_REID**, a deep learning model with four convolution layers and one full-connection layer is adopted. It contains four local max pooling layers, and has an initial learning rate of 0.0001. The model is pre-trained on ImageNet for the fine-tune on the training dataset [40].

For **Open_REID**, ReNet-50 and triplet-loss are utilized. Mini-batch SGD is utilized to train the CNN model. Training parameters, such as batch size, momentum and gamma, were set to 32, 0.9, 0.1 respectively. The initial learning rate was set at 0.0002.

Besides, four most recent state-of-the-art models of cross-dataset pedestrian retrieval are selected for further validation. UMDL [52] and PUL [54] are mainly based on unsupervised metric learnng. SPGAN [11] and TJ-AIDL [13] are based on deep representative cross-dataset learning. According to the report of the researches above, the evaluations are mainly between two benchmark datasets: DukeMTMC-reID and Market-1501, and the well-trained models from source datasets are directly tested on target datasets (without continuing training on target datasets).

Table 2. The Configuration of all parameters

Model	Parameter	Best Value
TrAdaBoost	Epoches	3
SPGAN	Batch Size	1
	Learning Rate	$2 * e^{-4}$
	Pool Size	50
	Lambda	10
	Momentum	0.5
Open_REID	BatchSize	256
	LearningRate	0.002
	Margin	0.5
	WeightDecay	$5 * e^{-4}$
	Epoches	150
Person_REID	Batch Size	16
	Learning Rate	0.0001
	Momentum	0.9
	Gamma	0.1
	Stride	2
Deeplab V2	Learning Rate	$2.5 * e^{-4}$
	Momentum	0.9
	Weight Decay	0.0005
	Epoches	20000
	Batch Size	10

4.3. Evaluations

Semantic Segmentation Results

To semantically segment pedestrian retrieval datasets, a semantic segmentation model is used to trained on the PascalVOC dataset with a recognition accuracy rate of 90.8%. The model classifies each pixel, making it possible to separate people from their background according to classification results. The process is shown in Figure 4.

In most cases, the result of the semantic segmentation was satisfactory, and the model accurately classified the internal pixels of images. However, there were some deficiencies (see Figure 4). These are described below.

1) The results can easily be affected by extraneous objects or people. When an extraneous person appears in an image, the model cannot always distinguish him or her from the target person. This can adversely affect the model's performance.

2) The results are not precise enough, and some pixels relating to the environment are incorrectly classified. In addition, the division margin of the pedestrian is sometimes not smooth enough.

3) The performance in images with low resolution is less than ideal.

After experimenting, it was found that the accuracy of pedestrian retrieval using concatenated images combined with raw and segmented images was higher than that using only semantically segmented images. Therefore, concatenated images were used in the training models.

Transfer Learning Results

The concatenated images were further used for transfer learning. Each auxiliary dataset including Market-1501,CUHK03,DukeMTMC-reID, and VIPeR was matched with the other three source datasets respectively. The results of the transfer learning are shown in Table 3.

From Table 3, the number of remaining samples in auxiliary datasets is greatly reduced after transfer learning. This is because there are many images with different distributions existing within different datasets, and these images were removed during the transfer learning process, which facilitated better knowledge sharing between source datasets and auxiliary datasets. The further validation of transfer learning will be illustrated in Table 4.

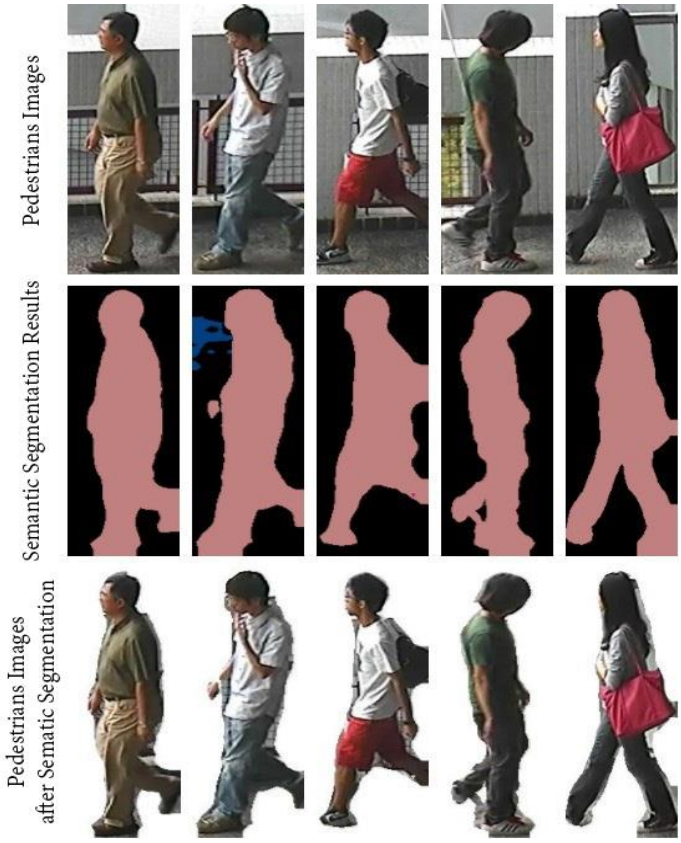


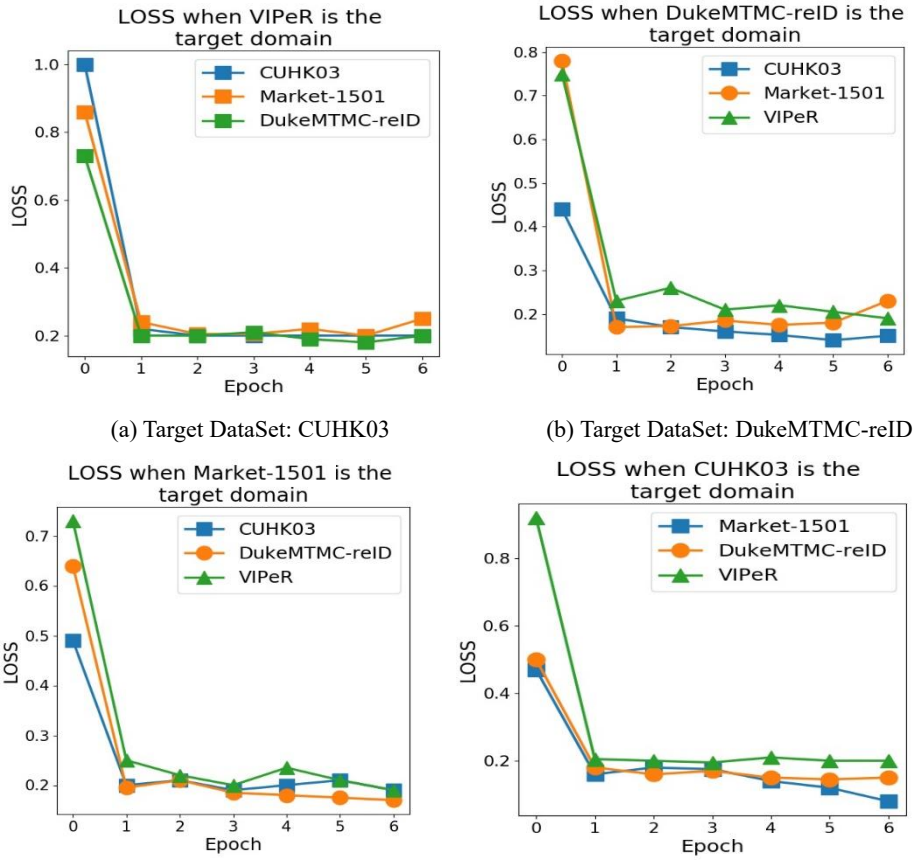
Figure 4. Semantic Segmentation Process

Table 3. Transfer Learning Results

Source Dataset	Number of Samples	Target Dataset	Remaining Samples
CUHK03	28,192	Market-1501	808
		DukeMTMC-reID	3,488
		VIPeR	14,885
Market-1501	32,668	CUHK03	4,290
		DukeMTMC-reID	2,319
		VIPeR	17,652
DukeMTMC-reID	36,411	Market-1501	3,018
		CUHK03	6,074
		VIPeR	16,400
VIPeR	1,264	Market-1501	477
		CUHK03	275
		DukeMTMC-reID	226

Image-Translation Results

The concatenated images from the source dataset could also be translated into new images which matched the style of images in target dataset by SPGAN. Translations were made on all cross-datasets, and the loss curve along the number of epochs as seen in Figure 5. After 6 epochs, the loss curves of all cross-dataset translations become steady, at around 0.2 (Figure 5). The translation effects of different cross-dataset combinations can be seen in Figure 6.



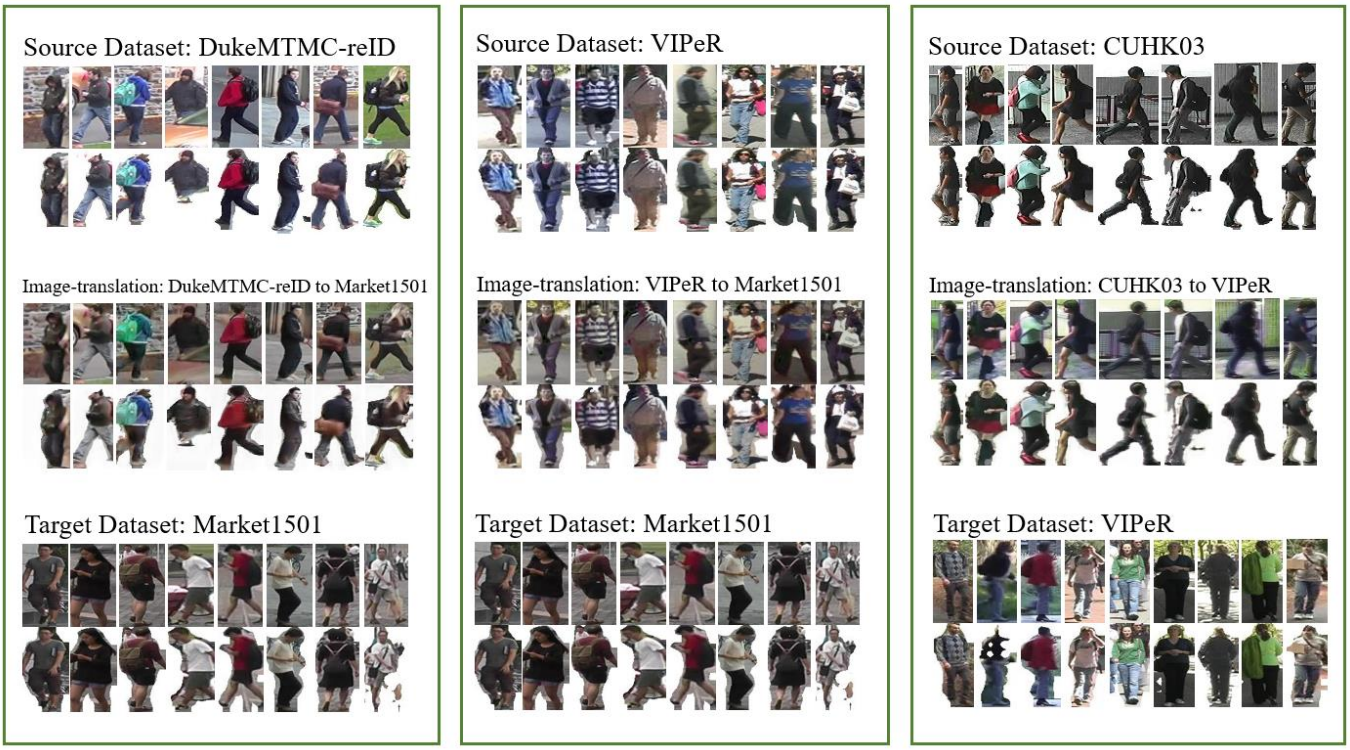
(a) Target DataSet: CUHK03

(b) Target DataSet: DukeMTMC-reID

(c) Target DataSet: Market-1501

(d) Target DataSet: VIPeR

Figure 5. Loss Curve of image-translation on all cross-dataset combinations.



(a) (b) (c)

Figure 6. Examples of image-translation between different cross-datasets.

As can be seen in Figure 6, the overall translation effects from different source datasets to target datasets are as good as expected. The translated images maintain a high consistency compared with images in target datasets: for example, images in DukeMTMC-reID are lighter and have greater clarity than those in Market-1501 (Figure 6 (a)), but the translated images from DukeMTMC-reID to Market-1501 have styles and qualities similar to the images in Market-1501 and seldom have the problem of missing content or color abnormalities for either original images or segmented images. The translated images from VIPeR to Market-1501 are significantly clearer because the number of pixels in VIPeR images is lower than in Market-1501, so the sharpeness is reduced, and the images are softer.

There are also bad cases amongst the translated images in which some image feature information is lost. For example, the color of some body parts is not normal, or the whole picture is less clear than expected. This may be due to differences in pixel values, lightness, background or tone. Because such cases can significantly influence the performance of Pedestrian Retrieval, a strategy using TrAdaBoost is also applied to filter all the images translated by SPGAN. This helped to remove bad cases, and improved the quality of the auxiliary datasets. The statistics of filtered results of all cross-datasets are shown in Table 4.

Table 4. Statistic of Filter Results of translated images from Source Dataset.

Source Dataset	Number of Samples	Target Dataset	Remaining Samples in Auxiliary Dataset
CUHK03	28,192	Market-1501	3,096
		DukeMTMC-reID	7,573
		VIPeR	2,1055
Market-1501	32,668	CUHK03	5,207
		DukeMTMC-reID	3,003
		VIPeR	21,677
DukeMTMC -reID	36,411	Market-1501	4,956
		CUHK03	7,414
		VIPeR	27,111
VIPeR	1,264	Market-1501	662
		CUHK03	375
		DukeMTMC-reID	361

In Table 4, images translated to VIPeR style have relatively high quality (on average, about 2/3 translated images from CUHK03, Market-1501 and DukeMTMC-reID pass the filter of TrdaBoost). Besides, translation between CUHK03 and DukeMTMC-reID also obtains relatively good results (averagely 1/5 translated images pass the exam of TrAdaBoost). While for other cross-data combinations, a certain proportion (from 12% to 25%) of translated images is also of high quality. Above all, image-translation can effectively provide a number of high-quality images for target datasets, helping to address problems arising from a lack of samples.

Effectiveness of FSRM-STS

In the second stage of the proposed four-stage retrieval model with STS (FSCM-STS), the four strategies described in section 3.5 are utilized. For a cross-dataset combination, these strategies generate four auxiliary datasets A1, A2, A3, A4. Open_REID is utilized to train the auxiliary datasets of all cross-dataset combinations to obtain the pre-trained model, and then use the pre-trained model to train the target datasets to get the final models of each cross-dataset with a specific strategy in the third stage.

For each strategy in the third stage, the training process (loss curve) on the target datasets are drawn. Taking the first sub figure of Figure 7 as an example, STS1 refers to the auxiliary datasets generated by strategy STS1, and the legend beneath the graph shows the combinations of cross-datasets. For example, DukeMTMC-reID_CUHK03 means

DukeMTMC-reID is the source dataset and CUHK03 is the target dataset. After 150 epochs, all the training process reached steady status with small loss values. STS2 with SPGAN had the lowest loss values for all four target datasets (average minimal loss value is 0.03). The average loss value for STS1 with TrAdaBoost and STS4 is around 0.05, and for STS3 it is around 0.04. When DukeMTMC-reID was the target dataset is, all strategies had relatively high loss values, ranging from 0.09 to 0.19, especially using STS3 between CUHK03 and DukeMTMC-reID. For other target datasets, the loss values were less than 0.03.

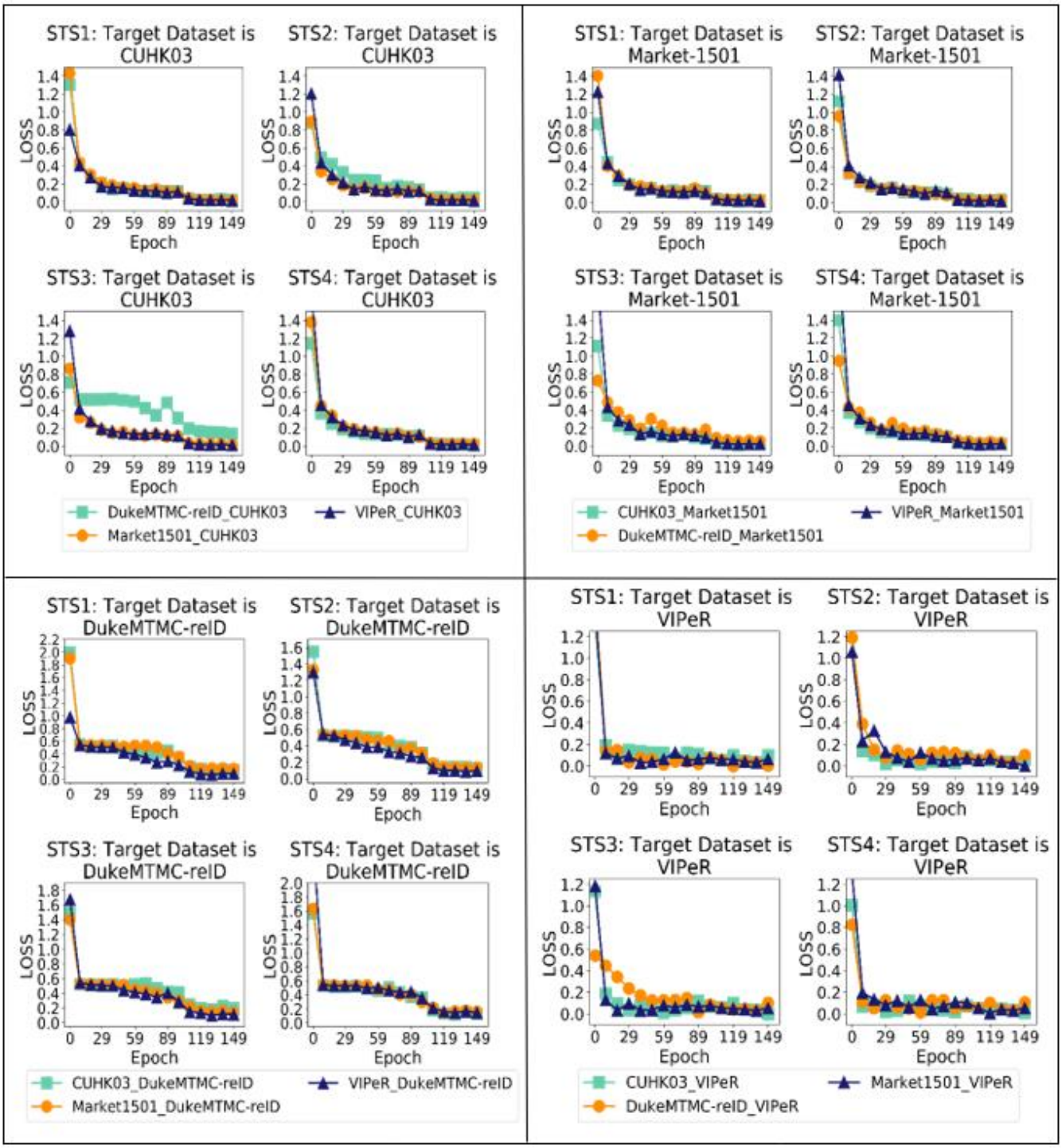


Figure 7. Training process (Loss curve) with Strategy STS1, STS2, STS3 and STS4.

In the fourth stage of FSCM-STS, a logistic regression model was adopted to utilize the final models from four strategies to realize pedestrian retrieval. The performance of the proposed model and its comparison with seven baselines could be seen in Figure8.

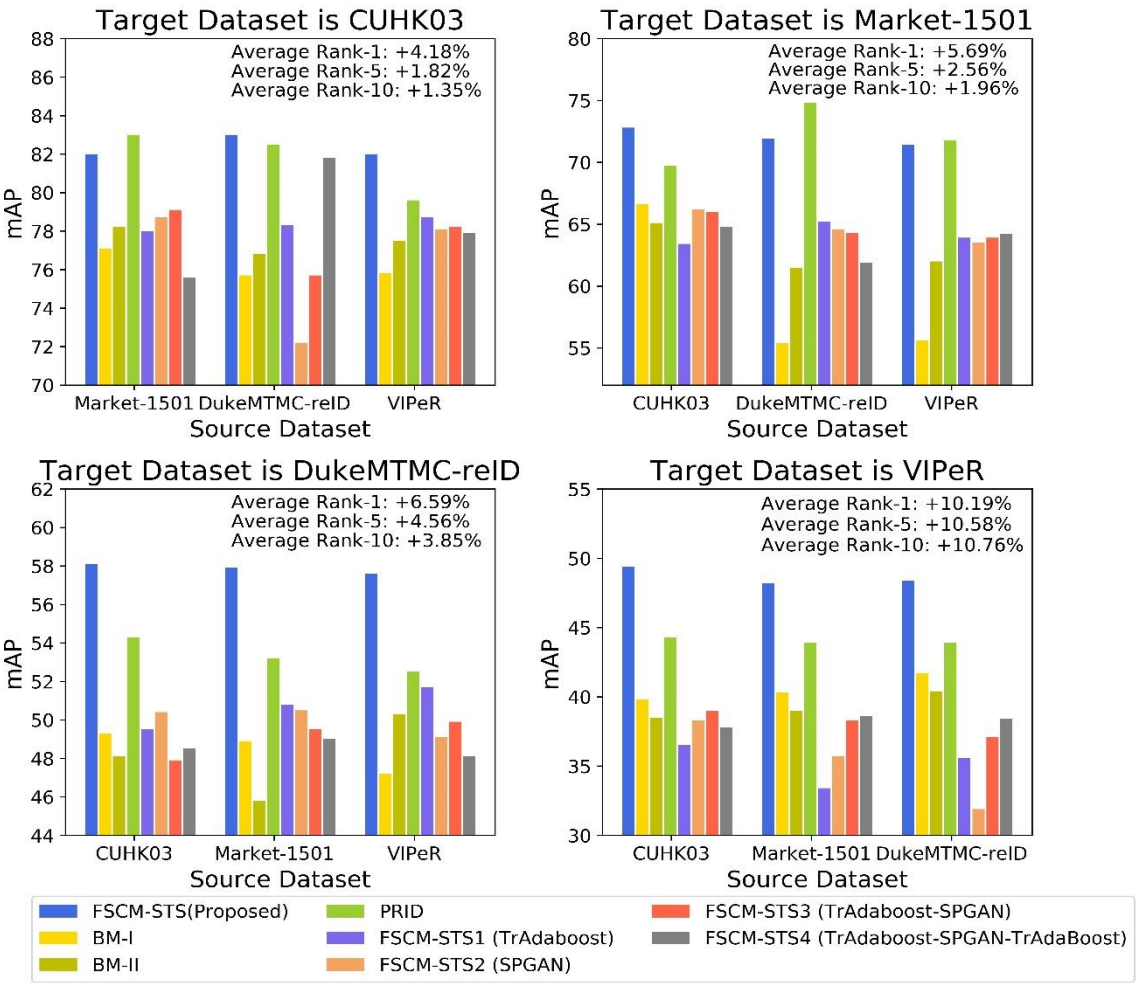


Figure 8. Comparisons of Proposed FSCM-STS

As is clear from the data in Figure 8, the proposed model significantly outperforms the other baselines in almost all the cross-datasets, though PRID is also very competitive, and has the best performance when the target dataset is Market-1501. For other baselines, performance is variable. BM-I (which takes source datasets directly as an auxiliary dataset) performs better when the target dataset is VIPeR; BM-II (incorporates segment model into BM-I) improves significantly on six cross-datasets compared with BM-I; and the maximal improvement of mAP is around 7% (From DukeMTMC-reID to Market-1501), which suggests that Semantic Segmentation is more useful for some cross-datasets than for others. One reason is occurrence of bad cases, for example, in the segmentation results of VIPeR, where occasionally, body information is missing or unclear. For other models, FSRM-STS$_1$ performs better on 4 cross-datasets, so using only TrAdaboost as the filtering strategy is a good solution. FSRM-STS$_2$ is good at learning from CUHK03 to DukeMTMC-reID; FSRM-STS$_3$ (TrAdaboost + SPGAN) suppresses other STS baselines on cross-dataset from Market-1501 to CUHK03; while FSRM-STS$_4$ (TrAdaboost + SPGAN + TrAdaBoost) obtains best scores on two cross-datasets (not including the proposed model and PRID).

The experimental results further illustrate the value of combining four strategies into a unified framework. Because each strategy has advantages on certain cross-dataset combination, none of them could obtain the highest scores on all datasets, while FSRM-STS$_1$ shows that using alone TrAdaboost produced better performances among the other three strategies. After utilizing STS to combine the four strategies, the average improvement of mAP is about 8%, hit@1 is about

7%, hit@5 is about 5% and hit@10 is about 4%. The state-of-art model PRID outperforms the four strategies separately on all cross-datasets because it considers the attributes of pedestrians. While FSRM-STS combines four strategies, and significantly surpasses PRID on 9 cross-datasets (especially when target datasets are DukeMTMC-reID and VIPeR), which further illustrates the effectiveness of the proposed model.

Comparison with the State-Of-The-Art Approaches

As introduced in Section 4.2, four state-of-the-art models are selected for comparison, which are UMDL [52], UL [54], SPGAN [11] and TJ-AIDL [13]. All the above models have the reported experimental results on two datasets: DukeMTMC-reID and Market-1501. So the two datasets are taken as experimental data, and the same experiment procedures are followed. Comparison results could be seen in Table 5. It is observed that the FSRM-STS outperforms all the competing methods. The main reason can be contributed the fact that the proposed model takes four strategies into consideration and optimized in a unified process.

Table 5. Comparison with the State-Of-The-Art Approaches

Model	DukeMTMC-reID -> Market-1501		Market-1501 -> DukeMTMC-reID	
	mAP	hit@1	mAP	hit@1
UMDL	12.4	34.5	7.3	18.5
PUL	20.5	45.5	16.4	30.0
SPGAN	26.7	57.7	26.2	46.4
TJ-AIDL	26.5	58.2	23	44.3
FSRM-STS	**30.2**	**61.3**	**29.1**	**51.5**

5. DISCUSSION

5.1. Analysis of the Accuracy of Pedestrian Retrieval

From Figure 8, it can be seen that, when the target dataset is VIPeR, the performance of the Baseline model BM-I is better than the other baselines without the proposed FSRM-STS model. This phenomenon indicates that knowledge could not be effectively transferred from other benchmark datasets to VIPeR, so the model performance did not to improve. Figure 9 shows the proportion distribution of images remaining in auxiliary datasets, which are selected by STS, and based on different source and target dataset combinations (Data is from Figure 8).

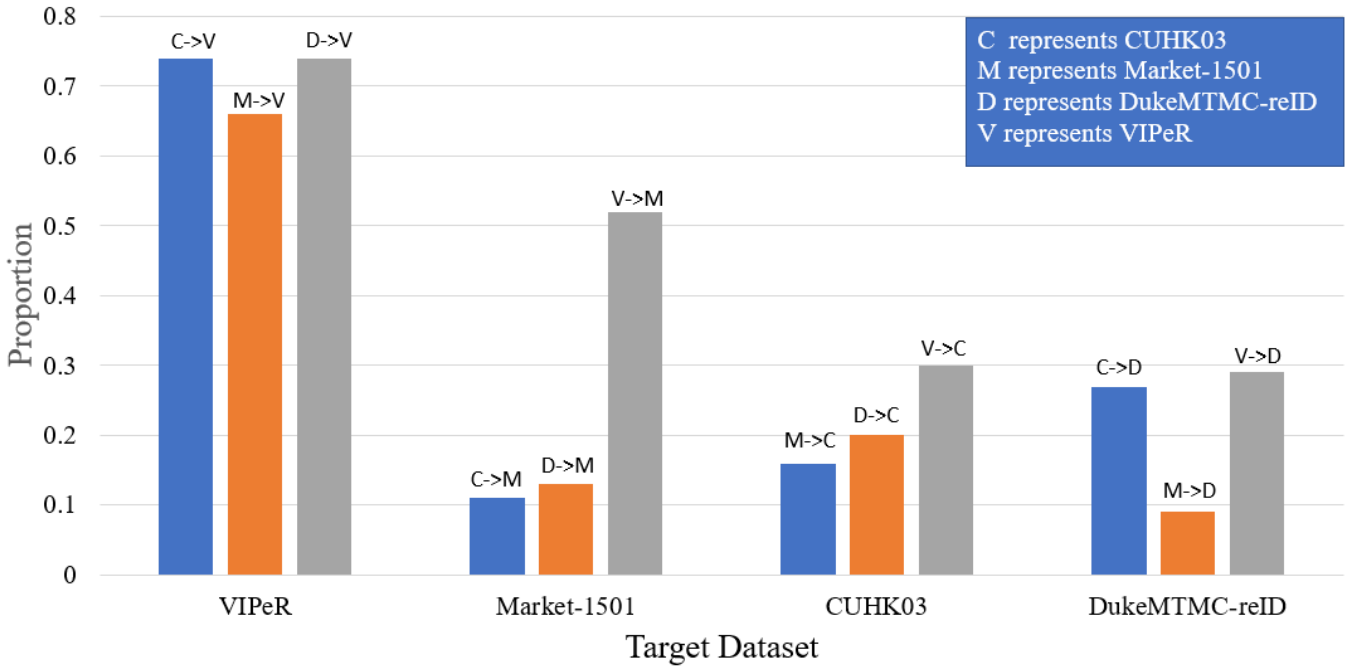


Figure 9. Image Proportion in auxiliary Dataset based on Different cross-dataset Combinations.

In Figure 9, C represents CUHK03, M represents Market-1501, D represents DukeMTMC-reID. "->" represents the direction from source dataset to target dataset, for example, C->V represents from CUHK03 to VIPeR. Unlike other cross-dataset combinations, the proportions of VIPeR (C->V, M->V, D->V) are all very high, suggesting that almost all the images in C, M and D could be transferred and translated to VIPeR style.

According to the results shown in Figure 8, the performance on VIPeR is the lowest (not better than Baseline model BM-I), which means that transferred and translated knowledge were not useful for pedestrian retrieval on VIPeR, and the existing proposed models could not identify the unique features of VIPeR and distinguish them from other benchmark datasets. In contrast, knowledge could be transferred from VIPeR to other datasets (The performances of V->M, C, D are significantly better than with BM-I). This phenomenon indicates that the unique features of VIPeR could not be learned by the current model, so the prediction results on VIPeR were not as good as other models. While some of the useful patterns could be learned from VIPeR to assist the model better understanding other benchmark datasets.

Because FSRM-STS combines the four strategies STS1, STS2, STS3 and STS4 into a unified framework, so even though each individual strategy may not be the best model on all benchmark datasets, the design of STS allows the four strategies to complement each other. According to the case study, for some images from different target datasets, at least one strategy could make a correct judgment, and contribute a higher weight in the final judgment stage. Figure 10 provides an example to illustrate how the mechanism works.

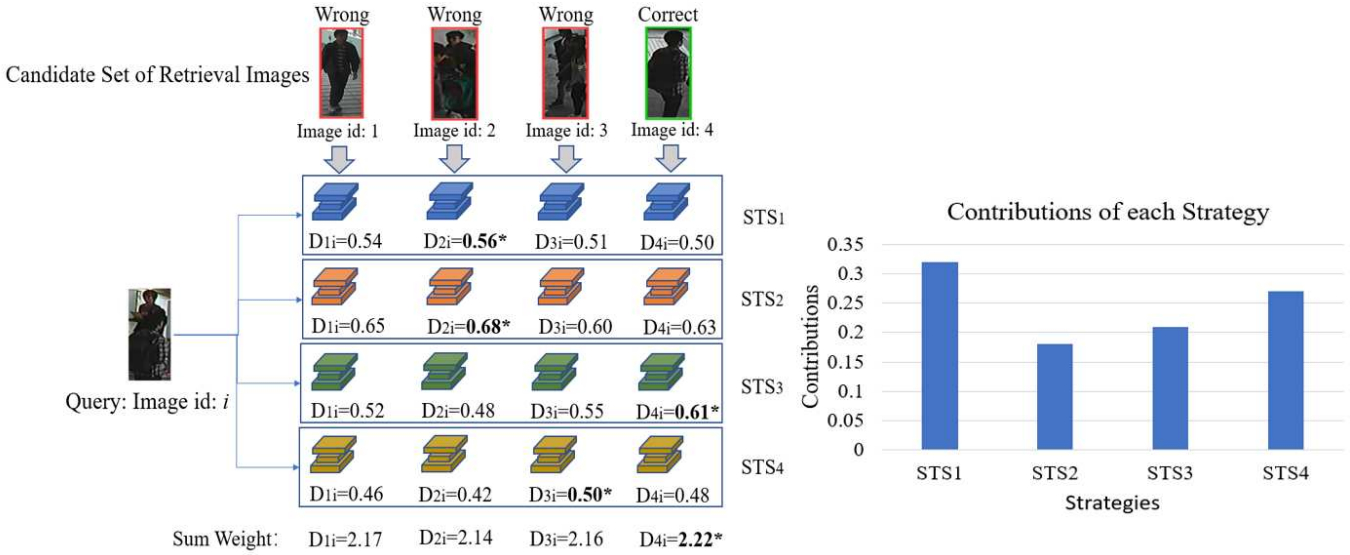


Figure 10. An Example of STS mechanism and the contribution distribution.

For a query image i, a candidate set with four images 1,2,3,4 has been retrieved from the proposed model, where Dki ($k \in \{1,2,3,4\}$) indicates the distance between images k and i. For strategy STS1, STS2 and STS4, the distance could not identify the correct image (4) from other wrong images. Strategy STS3 performed better in the task, and the distance of D4i is significantly bigger than the other strategies (average 9% increase). Due to the incorporation of STS3, the final sum weight shows the maximal value is from D4i, which allows the correct image to be retrieved and ranked in 1st place. The right sub-figure of Figure 10 shows the total contribution of each strategy towards the prediction results. STS1 (TrAdaBoost) contributes the most (32%) and STS2 (SPGAN) contributions the lowest (18%). The reason could be that translating images bring more noisy information int target datasets.

5.2. Case Study

To better analyze the experimental results, the model using DukeMTMC-reID is chosen as the source dataset and CUHK03 as the auxiliary dataset. This made it possible to compare the performances of baseline models and the FSRM-STS model. In Figure 11, query images of baseline models and the proposed model are on the left and the top 9 query results are on the right. Regardless of method, the results are not ideal, mainly because there is an additional person in the query image, which increases the difficulty of recognition. In the baseline model BM-I, the true positive image only appears twice. In the false positives the person shown has similar characters to the additional person in the query image. It can be inferred that BM-I was confused about which person was the target. The proposed model selected five true positive images. Of the four other baselines with different strategies, FSRM-STS4 performed best. It contributed a higher weight to FSRM-STS in the first ranking, and the second weight, FSRM-STS3 had the lowest performance, with only two true positive cases in top 10. The ranking results of both FSRM-STS1 and FSRM-STS2 are better than FSRM-STS3. As can be seen in Figure 11, the optimal results of FSRM-STS are based on a combination of the four strategies (STS1, STS2, STS3 and STS4).

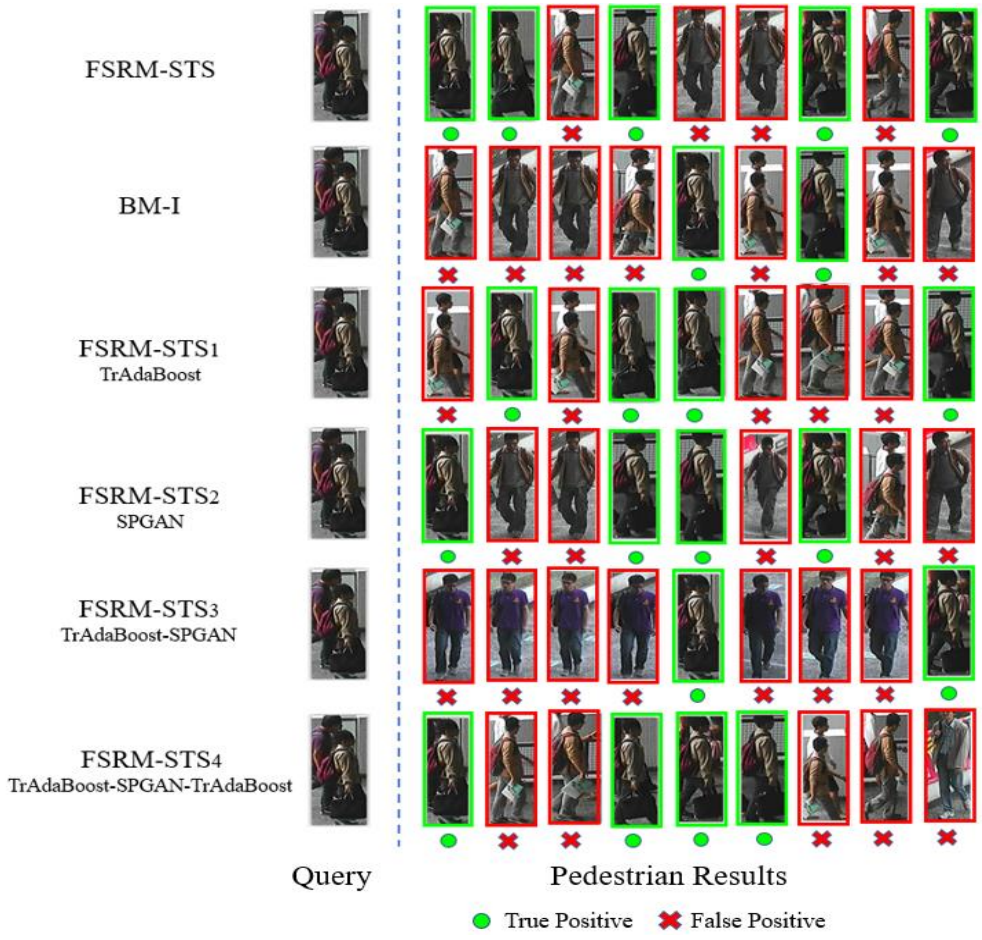


Figure 11. Case Studies of DukeMTMC-reID -> CUHK03 cross-dataset.

5.3. Discussion on Engineering Optimization

The proposed model has more than 10 sub-models should be trained and evaluated during the optimization process (semantic segmentation, image transfer and translation based on different strategies, person retrieval and etc). Thus, the training speed and the application of the trained model in reality is another problem that needs to be solved. The solution from an engineering perspective is discussed. The proposed framework can be seen in Figure 12.

Firstly, a parameter server (Pserver) framework [55, 56] is introduced to accelerate the training speed of the model. The parameter server is designed to be more suitable for engineering applications, where large-scale machine learning models are often deployed on server clusters with multiple CPUs and GPUs. This is a Horizontal Expansion Strategy, which is more cost-saving than a Vertical Extension Strategy. In current experiment, the code is mainly developed by Pytorch, which runs on a GPU server (two CPU Intel Xeon E5-2678 v3 and two GPU RTX 2080 Ti 24G). The current version of Pytorch is not good at parallel calculations on server clusters. For a parameter-based training framework, the main difference between Pytorch and Pserver is that all the sub models can be deployed on multiple GPU servers, and the most time-consuming parts (parameter updating based on gradient methods) are delivered to Pserver through the High-speed network. Paddle (http://www.paddlepaddle.org.cn/) is based on Baidu cloud (the largest search engine in China) and adopts the Pserver framework. The model could apply multiple GPU servers from Baidu cloud to generate GPU clusters, and then deploy Pserver on the clusters. The proposed FSCM-STS model could be deployed on Server Manager, and the Server Manager deployed the model on different GPU servers with the assistance of Resource Manager. Each GPU server could be assigned a specific number of workers, each of which was scheduled to manage several tasks, where each task contained the running code and a subset of training data.

The final models trained through Pserver framework were loaded in the model server (MS) of an online framework

designed to provide online services. The model server loads the well-trained models for online pedestrian image parsing and analysis. All the images in the source and target datasets are stored in a distribution database, and MS processes them offline to generate an index of each image. For the ith image, its key-value indexing format is {key: i, value:{V_{1i}, V_{2i}, V_{3i}, V_{4i}}}, where V_{ki} ($k \in \{1,2,3,4\}$), which means the representation embedding of image i derived from strategy k. Indexes are stored in a distribution memory database, such as Redis or HBase. The online framework provides a visualization portal to receive online queries (pedestrian images), and the portal sends the image to the Central Server (CS), CS invokes MS to get four representation vectors of the query image, and communicates with the Rank Server (RS) to retrieve the top K most related image indexes from the memory database in parallel. At the last step, the RS obtains the final list of images by searching the distribution database and returning the list to the portal.

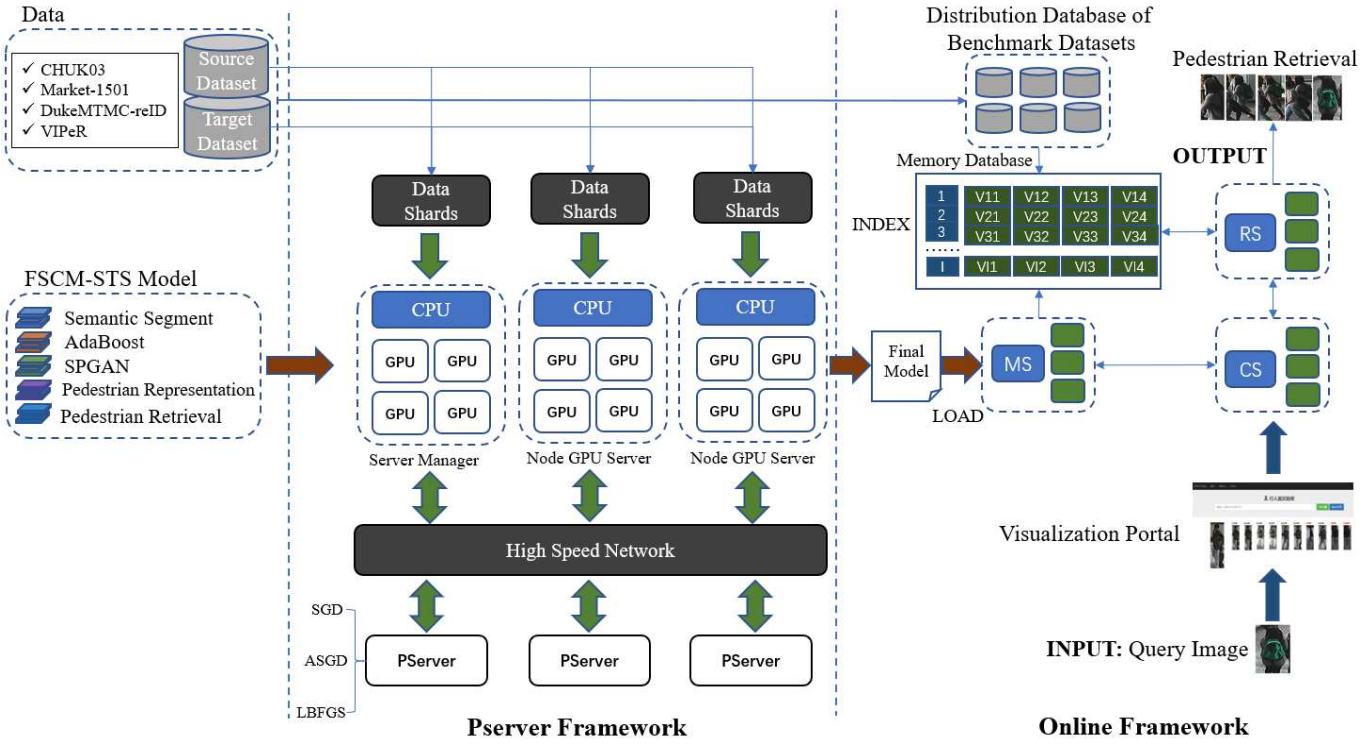


Figure 12. Engineering Framework of FSCM-STS.

In the future, the research will further verify the parallel calculation capability of the proposed model on Paddle. In another aspect, a Pedestrian retrieval demo is developed based on Online Framework (Figure 13 shows the demo interface). A hundred thousand images from 4 benchmark datasets (30G) are stored in a distribution database, and the index of each image and its four representation vectors are stored in a distribution Memory Database. The whole process of one retrieval task takes about 500 ms in its current version. The challenge of the proposed Online Framework for online applications will arise from the increasing number of images. The questions of how to efficiently store and manage the large number of images using advanced cloud storage methods and devices [57, 58, 59, 60, 61], and of protecting Data Security and Privacy [62] in a cloud environment will be two of future research directions.

Figure 13. Interface of Pedestrian Retrieval System Demo.

.

6. CONCLUSION

Pedestrian retrieval technology has a broad range of potential applications. The cross-dataset pedestrian retrieval model based on the four-stage retrieval model with a selection-translation-selection mechanism (FSCM-STS) proposed in this paper offers a potential solution to some of the problems in cross-dataset pedestrian retrieval (such as insufficient data, and differences between and within different datasets). Semantic segmentation and transfer learning are adopted in this model. Semantic segmentation reduces the influence of complex environments and occlusion in raw images and provides additional information. The selection-translation-selection mechanism provides four strategies to reduce the cost of labeling new data, and help pedestrian retrieval models to learn new knowledge, improving the accuracy of cross-dataset pedestrian retrieval.

Experimental results show that, compared with state-of-the-art pedestrian retrieval methods, FSCM-STS can improve cross-dataset pedestrian retrieval accuracy by an average of 6%. The four strategies only performed better on specific cross-data combinations, but after integrating them into a unified framework, the improvement on all cross-data combinations was significant. Case studies further indicate the value of applying FSCM-STS: the proposed models appear better at handling images with a complex background where, for example, the image has more than one pedestrian, or part of the pedestrians is blocked by trees or buildings.

The use of parameter server and parallel computing is discussed for accelerating the training speed, fast engineering iterations and to improve the online application values of the proposed model. A primary demo based on distribution environment was designed to confirm the value of the proposed framework.

The generalizability of the proposed model in other scenarios was also considered. Pedestrian retrieval is an important direction for image retrieval. It requires more accurate models with stricter constraints than other image retrieval models. In this paper, three mainstream methodologies are mainly discussed in Pedestrian retrieval, which are representation learning, metric learning and GAN based image translation. The proposed model combines the above three methodologies, and improves significantly on all cross-dataset combinations, especially for CUHK03 dataset, where the mAP reached as high as 83%. The methodologies are also relevant to many other image retrieval tasks, and similar researches have been verified in many scenarios, such as buildings, vehicles, devices, and face identification because the methodologies use representation learning to automatically extract image features based on different retrieval tasks. In future, more technologies will be attempted to incorporate, such as local feature learning [4, 28, 31, 33, 34] and attribute learning [13, 40], to further enhance the capability of the proposed model in Pedestrian Retrieval, and to extend the generalizability to other scenarios.

7. ACKNOWLEDGMENTS

This research is supported by Chinese National Youth Foundation Research (Grant No: 61702564), Soft Science Foundation of Guangdong Province (Grant No: 2019A101002020), Talent Scientific Research Foundation of Sun Yat-sen University (Grant No: 20000-18841202).

8. REFERENCE

[1] D. Gray, S. Brennan, H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking, in: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS). pp: 1-7. 2007.
[2] M. Geng, Y. Wang, T. Xiang, Y. Tian, Deep Transfer Learning for Person Re-identification, arXiv preprint arXiv:1611.05244 (2016). https://arxiv.org/abs/1611.05244.
[3] R. Zou, R. Cucchiara, E. Ristani, F. Solera, C. Tomasi, Performance measures and a data set for multi-target, multicamera tracking, in: European Conference on Computer Vision, Springer, 2016, pp. 17-35. https://doi.org/10.1007/978-3-319-48881-3_2.
[4] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang and J. Kautz. Joint Discriminative and Generative Learning for Person Re-identification. CVPR 2019.
[5] W. Li, R. Zhao, T. Xiao, X. G. Wang. DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 152-159. https://doi.org/10.1109/CVPR.2014.27.
[6] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In CVPR, 2016.
[7] T. Xiao, S. Li, B Wang, L. Lin and X. Wang. Joint Detection and Identification Feature Learning for Person Search. In CVPR, 2017.
[8] A. Hermans, L. Beyer, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. arXiv:1703.07737, 2017.
[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2015, pp. 1116-1124. https://doi.org/10.1109/Iccv.2015.133.
[10] J. Tian, Z. Teng, R. Li and Yan. Li. Imitating Targets from all sides: An Unsupervised Transfer Learning method for Person Re-identification. CoRR abs/1904.05020. 2019.
[11] W. Deng, L. Xheng, Q. Ye. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. CVPR 2018.
[12] J. Lv, W. Chen, Q. Li. Unsupervised Cross-dataset Person Re-identification by Transfer Learning of Spatial-Temporal Patterns. CVPR 2018.
[13] J. Wang, X. Zhu, S. Gong and W. Li. Transferable joint attribute-identify deep learning for unsupervised person re-identification. In CVPR 2018.
[14] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero-and homogeneously. In Proceedings of the European Conference on Computer Vision (ECCV), pages 172–188, 2018.
[15] W. Liang, G. Wang, J. Lai, J. Zhu. M2M-GAN: Many-to-Many Generative Adversarial Transfer Learning for Person Re-Identification. arXiv:1811.03768v1 . 2019.
[16] J. Lin, L. Wang, D. Metzler. Ranking under temporal constraints. in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010, pp. 79-88. https://doi.org/10.1145/1871437.1871452.
[17] D. Lee, H. Park, C. D. Yoo, Face Alignment using Cascade Gaussian Process Regression Trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 4204-4212. https://doi.org/10.1109/CVPR.2015.7299048.
[18] Y. Wu, Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 3400-3408. https://doi.org/10.1109/Cvpr.2016.370.
[19] S. Liu, F. Xiao, W. Ou and L. Si. Cascade Ranking for Operational E-commerce Search, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 1557-1565. https://doi.org/10.1145/3097983.3098011.
[20] I. Kviatkovsky, A. Adam, E. Rivlin. Color Invariants for Person Rei-dentification, IEEE Transactions on pattern analysis and machine intelligence, 35(7) (2012) 1622-1634. https://doi.org/10.1109/TPAMI.2012.246.
[21] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S. Z. Li. Salient color names for person re-identification. In European

conference on computer vision. Springer. pp. 536-551. https://doi.org/10.1007/978-3-319-10590-1_35. 2014.
[22] W. Hu, X. Zhou, T. Tan, J. Lou, S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28(4). 663-671. 2006. https://doi.org/10.1109/TPAMI.2006.80.
[23] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof. Large Scale Metric Learning from Equivalence Constraints, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2288-2295. https://doi.org/10.1109/CVPR.2012.6247939.
[24] D. Tao, L. Jin, Y. Wang, Y. Yuan, X. Li. Person re-identification by regularized smoothing kiss metric learning. IEEE Transactions on Circuits and Systems for Video Technology, 23(10) 1675-1685. 2013. https://doi.org/10.1109/TCSVT.2013.2255413.
[25] S. Liao, Y. Hu, X. Zhu and S. Z. Li. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. CVPR 2015.
[26] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang. Joint Detection and Identification Feature Learning for Person Search. CVPR 2017.
[27] Y. Chen, X. Zhu and S. Gong. Person Re-Identification by Deep Learning Multi-Scale Representations. ICCV2017.
[28] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A. C. Kot, G Wang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification. CVPR 2018.
[29] Z. Li, S. Y. Chang, F. Liang, T. S. Huang, L. L. Cao, J. R. Smith. Learning Locally Adaptive Decision Functions for Person Verification, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4639. 3610-3617. https: //doi.org/ 10.1109/CVPR. 2013.
[30] D. Cheng, Y. H. Gong, S. P. Zhou, J. J. Wang and N. N. Zheng. Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1335-1344. 2016. https://doi.org/10.1109/Cvpr.2016.149.
[31] L Wei, S Zhang, H Yao, W Gao, Q Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 420-428. 2017. https://doi.org/10.1145/3123266.3123279.
[32] X Zhang, H Luo, X Fan, W Xiang, Y Sun, Q Xiao, J Sun. Aligned reid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184. 2017. https://arxiv.org/abs/1711.08184.
[33] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1077-1085. 2017. https://doi.org/10.1109/Cvpr.2017.103.
[34] L. Zheng, Y. Huang, H. Lu, Y. Yang. Pose Invariant Embedding for Deep Person Re-Identification. arXiv preprint arXiv:1701.07732. 2017. https://arxiv.org/abs/1701.07732.
[35] R. R. Varior, M. Haloi, G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In: European conference on computer vision. Springer. pp. 791-808. 2016. https://dor.org/10.1007/978-3-319-46484-8_48.
[36] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan. End-to-End Comparative Attention Networks for Person Re-Identification, IEEE Transactions on Image Processing. 26(7). 3492-3506. 2017. https://doi.org/10.1109/Tip.2017.2700762.
[37] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, X. Xue. Pose-Normalized Image Generation for Person Re-identification. In Proceedings of the European Conference on Computer Vision (ECCV 2018). Springer. pp. 650-667. 2018. https://doi.org/10.1007/ 978-3-030-01240-3_40.
[38] Z. Zheng, L. Zheng, 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. pp. 3754-3762. 2017. https://doi.org/10.1109/iccv.2017.405.
[39] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang. Camera style adaptation for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5157-5166. 2018. https://doi.org/10.1109/CVPR.2018.00541.
[40] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang. Improving person re-dentification by attribute and identity learning. Pattern Recognition. 2019. https://doi.org/10.1016/j.patcog.2019.06.006.
[41] Z. Zheng, L. Zheng, Y. Yang. A discriminatively learned CNN embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 14(1). 13. 2018. https://doi.org/10.1145/3159171.
[42] J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative Adversarial Networks. In Proceedings of the 27th International Conference on Neural Information Processing System. (2). Pp: 2672-2680. Montreal Canada. Dec 08-13. 2014.
[43] J. Y. Zhu, T. Park, P. Isola, A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision. pp: 2223-2232. 2017. https://doi.org/10.1109/ICCV.2017.244.

[44] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and Transferring Mid- Level Image Representations using Convolutional Neural Networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, 2014, pp. 1717-1724. https://doi.org/10.1109/Cvpr.2014.222.
[45] E. Ahmed, M. Jones, T. K. Marks, An Improved Deep Learning Architecture for Person Re-Identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3908-3916. 2015. https://doi.org/10.1109/CVPR.2015.7299016.
[46] G. R. Xue, Y. Yu, W. Dai, Q. Yang. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine learning (ICML'07). 193-200. 2007. https://doi.org/10.1145/1273496.1273521.
[47] J. Jiang, C. Zhai. Instance weighting for domain adaptation in NLP. In Proceedings of the 45th annual meeting of the association of computational linguistics. ACL'07. pp: 264-271. 2007. https://www.aclweb.org/anthology/P07-1034.
[48] X. Liu, Z. Liu, G. Wang, Z. Cai, H. Zhang. Ensemble Transfer Learning Algorithm. IEEE ACCESS. (6). pp: 2389-2396. Dec. 2017.
[49] N. Li, H. Hao, Q. Gu, D. Wang, X. Hu. A Transfer Learning Method for Automatic Identification of Sandstone microscopic images. Computer & Geosciences. (103). 2017.
[50] L. Yan, R. Zhu, Y. Liu, N. Mo. TrAdaBoost Based on Improved Particle Swarm Optimization for Cross-Domain Scene Classification with Limited Samples. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 11(9). Sept. 2018.
[51] L. Ma, X. Yang, D. Tao. Person Re-Identification Over Camera Networks Using Multi-Task Distance Metric Learning. IEEE Transactions on Image Processing. 23(8). 3656-3670. 2014. https://doi.org/10.1109/Tip.2014.2331755.
[52] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, Y. Tian. Unsupervised Cross-Dataset Transfer Learning for Person Re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'16). pp. 1306-1315. 2016. https://doi.org/10.1109/CVPR.2016.146.
[53] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE transactions on pattern analysis and machine intelligence. 40(4). 834-848. 2017. https://doi.org/10.1109/TPAMI.2017.2699184.
[54] H. Fan, L. Zheng, Y. Yang. Unsupervised Person re-identification: Clustering and Fine-tuning. CoRR abs/1705.10444. 2017.
[55] M. Li, L. Zhou, Z. Yang, A. Li, F. Xia, D. Andersen, A. Smola. Parameter Server for Distributed Machine Learning. Big Learning NIPS Workshop. 6.1-10. 2013.
[56] M. Li. Scaling Distributed Machine Learning with the Parameter Server. In Proceedings of the 2014 International Conference on Big Data Science and Computing. No.3. August 04-07. NY. USA. 2014.
[57] K. E. Psannis, C. Stergiou, B.B. Gupta. Advanced Media-Based Smart Big Data on Intelligent Cloud Systems. IEEE Transactions on Sustainable Computer. 4(1). 77-87. 2019.
[58] Z. Fan, D. Park. Extending SSD Lifespan with Comprehensive Non-Volatile Memmory-based Write Buffers. Journal of Computer Science and Technology. 34(1):113-132. 2019.
[59] C. Wu, V. Sreekanti, J. M. Hellerstein. Autoscaling tiered cloud storage in Anna. In Proceedings of the VLDB Endowment. 12(6). 624-638. 2019.
[60] L. Tu, S. Liu, Y. Wang, C. Zhang, P. Li. An Optimized cluster storage method for real-time big data in Internet of Things. The Journal of Supecomputering. 1-17. 2019.
[61] R. Nayak, S. Choudhary. A Survey of Data Protection in Cloud Computing. International Journal of Modern Engineering & Management Research. 6(4). 2018.
[62] I. B. Mao, Y. Yang, S. Wu, H. Jiang, K. Li. IOFollow: Improving the Performance of VM live storage migration with IO following in the Cloud. Future Generation Computer System. 91. 167-176. 2019.
[63] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang. A siamese long short-term memory architecture for human re-identification, In European conference on computer vision. pp. 135-153. Springer. 2016. https://doi.org/10.1007/978-3-319-46478-7_9.