Research Advanced in Pedestrian Re-identification based on Deep Learning

Shibo Ren*

School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

* Corresponding Author Email: 2018021413024@std.uestc.edu.cn

Abstract. Pedestrian re-identification (re-ID) is always a research hotspot in the intelligent video surveillance community, which has been applied in various fields such as social security and criminal investigation. Though much effort has been devoted to boosting the accuracy and speed, person re-ID still faces numerous challenges and difficulties, including low-quality captured images, inconsistent camera angles, varied pedestrian postures, significant changes in background environments, and issues related to lighting and occlusion. This paper aims to summarize the recent studies of pedestrian re-ID to motivate the subsequent development. Specifically, this paper focuses on the technological development, application needs, and challenges of person re-ID from the aspects of occluded pedestrian re-ID, unsupervised person re-ID, virtual data generation, pose-aware person re-ID, and pedestrian search. It also identifies existing problems and offers a perspective on the future trends. The hope is that through the study of these methods, an understanding of the current development directions in person re-ID can be achieved and researchers can be provided with a reference.

Keywords: Pedestrian re-ID, pedestrian retrieval, deep learning, virtual data generation.

1. Introduction

Person re-identification (re-ID), which is also known as pedestrian re-identification, refers to determining whether a pedestrian appearing under different non-overlapping camera views is the same individual and ascertaining the appearance time and location to form the movement trajectory of the targeted person [1]. Thanks to the success of pattern recognition and deep learning technology, pedestrian re-identification technology has attracted increasing attention from academics and industry.

After more than 10 years of development, significant progress has been achieved and the mainstream person re-identification studies develop through three representative stages: re-ID based on multi-camera tracking, re-ID based on traditional features, and re-ID based on deep learning [2]. In the video surveillance system, Fig. 1 shows the classical framework of pedestrian re-identification, which first consists of multiple surveillance cameras with non-overlapping views. The position and number of cameras can be changed at any time, and the images and perspectives captured by different surveillance cameras are all different. In such a surveillance system, monitoring the movements of pedestrians is crucial. The monitoring of pedestrians is mainly based on pedestrian detection, trajectory tracking, and feature extraction modules. Feature similarity measurement is another core part of the task, which determines whether the system can accurately distinguish different pedestrians and recognize the images of the same one under different camera perspectives. There are usually two types of methods: distance-based measurement and probability model-based measurement. The distance-based measurement method mainly judges the similarity of two samples by calculating the distance between feature vectors, while the probability model-based method calculates the similarity by modeling the distribution of feature vectors.Fig. 1 shows a standard pedestrian re-identification system.

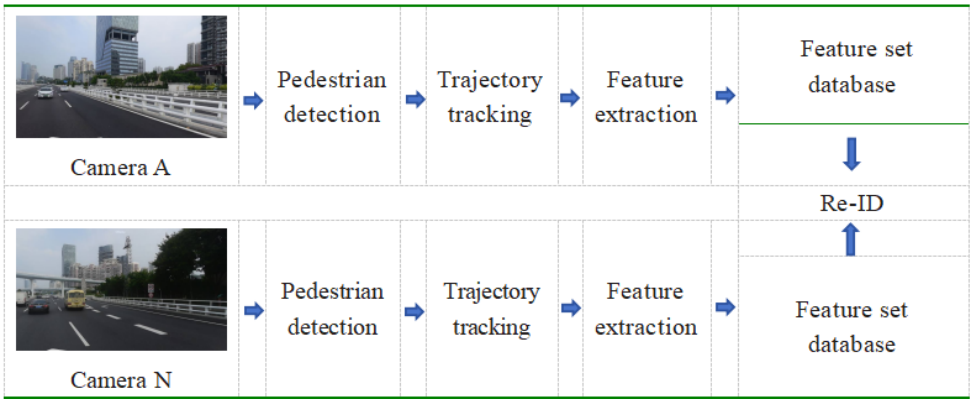

Fig. 1 Pedestrian Re-identification Framework in Video Surveillance Systems [1]

In real-world applications, person re-identification technology still faces numerous challenges, including but not limited to occlusion issues, high annotation costs, a lack of training data, significant differences between various scenarios and modalities, and insufficient cross-domain generalization capabilities. Comprehensive analysis indicates that the performance of individual re-identification algorithms in practical applications has not met expectations due to the combined effects of multiple factors. This suggests that there are still certain gaps to be overcome before this technology can be widely adopted. In view of this, this paper reviews the latest research findings in areas such as occluded individual re-identification, cross-domain individual re-identification, unsupervised individual re-identification, and virtual data generation, aiming to meet the needs of the development and practical application of individual re-identification technology. At the same time, this article also outlines the current research status, the problems faced, and future development trends.

The content of this paper is as follows. Section 1 is the introduction. Section 2 illustrates popular solutions to the existing challenges in the research field of person re-identification. In Section 3, the datasets, evaluation metrics and performance of representative methods will be introduced. Section 4 provides the discussion. The conclusion is shown in Section 5 .

2. Method

2.1. Obstructed pedestrian re-identification

In video surveillance scenarios of real-world, pedestrian images are often affected by occlusion issues, and it is significantly affects the performance of current pedestrian re-identification algorithms in practical applications. Many researchers have begun to focus on the topic of occluded pedestrian re-identification algorithms. Existing research effectively reduces the impact of occlusion by extracting visible pedestrian features, feature alignment, and supplementing missing pedestrian information, thereby improving the performance of occluded pedestrian re-identification.[3].

For instance, MLRAT [4] aims to delve into the embedding of each patch within an image to more accurately capture local features. This study employs unsupervised learning techniques to identify semantic key regions closely related to pedestrian representation and to master the structural correlations between these key regions through a two-layer graph convolutional network (GCN). Subsequently, these structural correlations are applied to guide the integration and learning process of local features. SLRA captures the interrelationships between a set of images through global features, which helps the model study discriminative features. It utilizes a RAT, relation-aware transformer, module to capture the associations between neighboring samples. Moreover, to narrow the gap between training phases and testing ones, SLRA adopts self-distillation techniques to incorporate sample-level relationships into the core network., promoting the core network to generate discriminative features.

The method employs an innovative strategy that combines the cross-entropy loss function with the ArcFace loss function. By applying the cross-entropy loss function, the predictions of model align with the true labels, thereby enhancing classification performance. Simultaneously, with the help of

the ArcFace loss function, the discriminability of classes within the feature space is strengthened, which in turn improves the recognizability of the features. The method increases the separation between features of different classes, making the boundaries clearer in the feature space. Consequently, the optimized model achieves a 70.4% Rank-1 accuracy and a mean average precision (mAP) at 61.8% on the Occluded-DukeMTMC dataset, verifying that this optimization strategy can significantly enhance the performance of the base network.

2.2. Domain Generalized Pedestrian Re-identification

The current pedestrian re-identification models have insufficient generalization performance across different scenarios, mainly due to the complexity of the task itself, involving numerous variables and challenges; at the same time, the lack of sufficient labeled training data is also a significant reason. Many researchers have long recognized the challenges of cross-scene problems and proposed transfer learning as a strategy to enhance the model's adaptability in new scenarios. This strategy has become a research hotspot in this field of pedestrian re-identification, but its implementation process is time-wasting and labor-intensive. Therefore, there is an urgent need for further research on plug-and-play human re-identification technology to promote its domain generalization capabilities in practical applications.

Yu et al. proposed a method of pedestrian re-identification that fuses multi-scale features with an attention mechanism [5]. This method extracts global and local features and uses stacked convolutional layers to strengthen the learning of the image’s overall and detailed information. To extract multi-scale features, the algorithm first designs a bottleneck layer with stacked convolutional ones, which cover different fields of view. Then, it uses a feature fusion attention module to dynamically weight and fuse these multi-scale features. Then a multi-level attention module is added to deeply mine the semantic information of detailed features. Finally, the detailed features rich in semantic information are input into the discriminator to perform the task of pedestrian re-identification. In addition, a style regularization module is designed to reduce the impact of image brightness variations between different datasets on the model's generalization performance.

Experiments were finished on two widely used pedestrian datasets, Market-1501 and DukeMTMC-reid, to verify the effectiveness of each module. MFEM (Multi-scale Feature Extraction and Fusion Module), SNM (Style Normalization Module), and Attention (Multi-level Attention Module) represent different modules. Through ablation studies, the results show that when the multi-scale feature extraction and fusion module, multi-level attention module, and style normalization module are used together, the model can achieve optimal performance, thus confirming the effectiveness of each module.

2.3. Unsupervised Pedestrian Re-identification

There are mainly two categories of methods in unsupervised pedestrian re-identification: unsupervised person re-identification based on domain adaptation, which assumes the existence of annotated source domain data and unannotated target domain data; and fully unsupervised person re-identification, which only involves unannotated data in the target domain and may include additional unannotated auxiliary information. When it comes to performing unsupervised pedestrian re-identification tasks, two core challenges are faced: one is to extract features that can distinguish individuals, and the other is to deal with the noise label problem that occurs during the generation of pseudo-labels. Researchers have proposed an unsupervised pedestrian re-identification method based on visual transformers. The method focuses on utilizing cameras and local similarity auxiliary modules to optimize similarity computation and address the noise label problem; at the same time, by improving image encoding, the robustness of the model is enhanced to more effectively extract features that distinguish individuals.

Yang and his collaborators [6] proposed an unsupervised person re-identification network based on the Vision Transformer. The network employs image encoding technology with a random sliding window, solving the problem of information loss, and uses the Vision Transformer's unique "dropped

features" as local variables to compute local similarity. By integrating local features, this method addresses the problem of information loss in unsupervised person re-identification and computes local similarity through "dropped features," while combining camera similarity to improve the quality of pseudo-label generation. In addition, the network integrates camera similarity to enhance the accuracy of overall sample similarity computation, thereby improving the quality of pseudo-label generation. The overall architecture of the network consists of three steps.: the first stage loads pre-trained Vision Transformer model weights to extract features; the second stage clusters the extracted features for camera and local fusion features and assigns pseudo-labels; and the third stage uses an assigned pseudo-labels for supervised training. Furthermore, a local similarity auxiliary module and an embedding strategy based on overlapping sliding windows are designed to optimize similarity computation, enhancing the model's robustness. To reduce the differences between categories and alleviate feature discrepancies, the loss function integrates both cross-entropy loss and triplet loss, while also enhancing intra-class similarity.

2.4. Virtual Data Generation

The largest publicly available datasets, such as MSMT17, have certain limitations in terms of data quality and scale. These datasets typically only include single-camera views and simple backgrounds, resulting in limited variation in the generated pedestrian images. This, in turn, poses challenges for training efficient deep learning models. However, in applications of daily life, the background is complex and variable, the number of images is vast, and lighting conditions are diverse and unpredictable. When models trained on open-source datasets are directly applied to these complex scenarios, their generalization ability is often limited, leading to a significant drop in performance. Therefore, the technology for generating and selecting pedestrian images has significant research value and importance.

Currently, virtual engine technology can generate scalable and customizable large amounts of virtual data. By mass-producing human models and simulating them walking along preset paths in a virtual engine, and then collecting and automatically annotating pedestrian images, a large number of virtual data with extremely low annotation costs can be obtained for model learning.

Cai et al. designed an algorithm for generating and selecting virtual data, which is tailored for cross-domain pedestrian re-identification [7]. Since the start of the algorithm, the target domain's foreground information is used, like the color distribution of pedestrian clothing, which guides the virtual 3D human models' generation, thereby producing virtual features similar to real human appearances. Then, by using foreground information to distinguish different pedestrians, the target domain background information on the generated virtual data is replaced to promote the quality of the source domain data. By analyzing the video stream of each camera, it can be observed that the pixel areas where pedestrians pass will change significantly due to different clothing. Based on this observation, the variance of each pixel and each channel of all unannotated images collected by each camera in the target domain is calculated, and the pixel areas with variance exceeding a specific threshold are used as background images for subsequent use, narrowing the background candidate range through these pedestrian-active areas. Finally, to select the most similar training subsets from the source domain to the target domain, distribution metrics such as the Wasserstein distance were used for model training. The experimental results showed that this method has more advantages compared to other existing pedestrian data generation algorithms, significantly improving the cross-domain generalization performance of the pedestrian re-identification model.

The UnrealForDuke dataset generated through the above data creation experiments surpasses the quality of existing publicly available real-world and virtual datasets. However, given that the process of learning pedestrian re-identification models requires extracting pedestrian features from image pixels and that the models ultimately determine whether the identity ID belongs to the same person based on the similarity of pedestrian features, the original training set can be filtered at the feature level to further obtain a set of pedestrian images that better match the target domain feature distribution.

3. Experiment

3.1. Common datasets

Pedestrian re-identification research before 2017 mainly focused on pedestrian feature representation and similarity measurement, while research combining attention mechanisms since 2018 has promoted the pedestrian re-identification performance. The common pedestrian re-identification datasets by public include Market-1501[8], DukeMTMC-reID (Duke multi-target, multi-camera for person re-identification) [9] and MSMT17(multi-scene multi-time) [10], as shown in Table 1.

Table 1. Characteristics of the most commonly used re-ID datasets

Data set	Year	Population	Number of images	Number of views	Descriptions
Market-1501 [8]	2015	1501	32668	6	Contains interfering images, close to real scenarios
DukeMTMC-reID [9]	2016	1812	36411	8	Pedestrian images are manually selected by boxes.
MSMT17 [10]	2018	4101	126441	15	Large scale, close to real-life scenarios

3.2. Evaluation Metrics

When evaluating the performance of re-identification algorithms, indicators commonly used include Rank-N metrics, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP). The Rank-N metrics involve ranking the feature distances between query images and registered images to retrieve images, and then calculating the probability of the correct target pedestrian image appearing within the top N ranked results. The higher this probability, the better the performance of the re-identification algorithm. Rank-N metrics provide an intuitive measure of performance, with common specific values including Rank-1, Rank-5, Rank-10, and Rank-20. For example, Rank-20 refers to the probability of the correct target pedestrian image appearing within the top 20 images in the registered database.

To reflect the recall performance of query results, mean average precision (mAP) is introduced. In mAP, AP stands for average precision, calculated by searching the part under the precision-recall curve (PRC) for each query, taking into account both precision and recall performance. mAP comprehensively evaluates the performance of an algorithm by calculating the mean of the average precision over multiple queries [8]. The specific calculation is as follows:

$$mAP = \frac{1}{m}\sum_{i=1}^{m} AP_i \tag{1}$$

where m is the number of query graphs in the query set, AP_i is the average precision for the i-th query image.

An ideal re-ID algorithm should ensure that the retrieved images for all query set images are ranked as high as possible in the final results, which mAP cannot reflect. To address this, Ye et al. proposed the mean inverse negative penalty metric [2], as:

$$mINP = \frac{1}{m}\sum_{i=1}^{m} \frac{|G_i|}{R_i^{hard}} \tag{2}$$

where m is the number of query graphs in the query set, R_i^{hard} refers to the index rank value of the most difficult positive sample in the matching list, which is the index position of the last correct matching result, while |Gi| refers to the total number of correctly matched samples.

3.3. Performance comparison

Table 2 presents the state-of-the-art (SOTA) statistical data currently in the field of pedestrian re-identification on the previously mentioned datasets Market-1501 and DukeMTMC, which focus on unsupervised pedestrian re-identification techniques. The "Source" column reveals whether source annotated data was utilized during the target Re-ID model training process. The "Generation" column indicates whether an image generation step was involved. The report lists Rank-1 accuracy (%) and mAP (%). The TJ-AIDL method requires additional attribute annotation information. The PAUL, MAR, and CASC use MSMT17, which is their source dataset.

Table 2. Results of the SOTA unsupervised pedestrian Re-ID on two mentioned datasets.

Methods	Gen	Source	Market-1501		DukeMTMC	
CAMEL ICCV'17	No	Model	mAP	R1	mAP	R1
PTGAN CVPR'18	Yes	Data	54.4	26.3	/	/
TJ-AIDL CVPR'18	No	Data	58.2	26.9	46.8	26.3
MAR CVPR'19	No	Data	58.2	26.6	44.3	23.1
PAUL CVPR'19	No	Model	67.8	40.1	67.1	48.1
CASC ICCV'19	No	Data	68.6	40.1	72.1	53.1
HCT CVPR'20	No	Model	65.4	35.5	59.3	37.8
SpCL NeurIPs'20	No	Data	80.1	56.5	69.7	50.6

It can be seen that as the years progress, the algorithms used for pedestrian re-identification generally become increasingly refined, and their performance tends to gradually improve. There is a lack of comparison under the same conditions to determine the impact of using synthetic virtual humans on performance.

4. Discussion

The current issues faced by this research direction typically include the following three aspects:

(1) Difficulty in annotating individuals: In video surveillance systems, a scene may contain many pedestrians with mutual occlusions, making annotation a huge task. Data annotation is not only time-consuming and costly, but the operation itself is also very difficult.

(2) Variations in perspective, background, and lighting conditions across different scenarios: Changes in background, different lighting conditions, and variations in shooting perspective can significantly alter the visual appearance of the same person, causing substantial interference to the computer's effective feature representation of pedestrians. This is also the reason why overcoming the challenge of scene transitions, which are inevitable in cross-domain surveillance systems, is necessary.

(3) The quality of visual data is highly sensitive to interference factors, such as significant differences and difficulties in controlling variables like resolution, frame rate, imaging conditions, and imaging angles, which pose challenges to ensuring high-quality visual data. Pedestrian images extracted from video footage often have a low pixel count and appear blurry, especially when the background color is similar to the clothing color of the pedestrian, sometimes making it difficult to discern the outline of the pedestrian. Additionally, in real-world scenarios, pedestrian images that are marked and annotated are often partially obscured by obstructions. Research on pedestrian re-identification generally focuses on these three challenges. These issues can be illustrated by various real-life situations shown in Fig. 2



Fig. 2 Pedestrian Re-identification Challenges in Video Surveillance Systems

5. Conclusion

In the area of computer vision, pedestrian re-identification technology, as a key research topic, has seen new development opportunities with the vigorous deep learning technology's development. This article tries to explore the challenges faced by pedestrian re-identification and introduces commonly used datasets and evaluation criteria from the perspective of deep learning. The article focuses on analyzing the progress and practical application needs of person re-identification technology, outlining the latest research findings in four advanced areas: person re-identification under occlusion conditions, person re-identification with domain generalization capabilities, person re-identification in unsupervised learning environments, and virtual data generation technology. In addition, the article illustrates the development status and trends of pedestrian re-identification technology nowadays through specific cases. By summarizing various methods, the article clarifies the current situation of pedestrian re-identification technology developed by deep learning and looks forward to possible breakthroughs in pedestrian re-identification tasks in our future, in the hope of forming a comprehensive perspective for observing and describing more people comprehensively.

References

[1] Zhang Y L. Research on Pedestrian Re-identification Algorithms. Shanghai Jiao Tong University, 2019. DOI:10.27307.

[2] Ye M, Shen J B, Lin G J, Xiang T, Shao L and Hoi S C H. 2022. Deep learning for person re-identification: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2872-2893.

[3] Zhang Y F, Yang H Y, Zhang Y J, et al. Research Progress in Pedestrian Re-identification Technology. Journal of Computer-Aided Design & Computer Graphics, 2023, 28(06): 1829-1862.

[4] Guorong Lin, Zhiqiang Bao, Zhenhua Huang, Zuoyong Li, Wei-shi Zheng, Yunwen Chen.

[5] Yu M, Li X B, Guo Y C. Domain Generalization for Person Re-identification with Attention Mechanism Fusion. Control and Decision, 2022, 37(07): 1721-1728. DOI:10.13195.

[6] Yang W Q, Ding Z F, Song Z G. Research on Unsupervised Pedestrian Re-identification Integrating Local Features with ViT. Fujian Computer, 2023, 39(12): 8-14. DOI:10.16707.

[7] Cai Y W, Zhang Y J, Zhang Y F. Virtual Data Generation and Selection for Cross-Domain Pedestrian Re-Identification. Journal of Graphics, 2023, 44(04): 775-783.

[8] Zheng L, Shen L Y, Tian L, Wang S J, Wang J D and Tian Q. Scalable person re-identification: a benchmark//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1116-1124.

[9] Ristani E, Solera F, Zou R, Cucchiara R and Tomasi C. 2016. Performance measures and a data set for multi-target, multi-camera track-ing//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 17-35.

[10] Wei L H, Zhang S L, Gao W and Tian Q. 2018. Person transfer GAN to bridge domain gap for person re-identification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 79-88.