Automation and Machine Learning (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2024.050113
ISSN 2516-5003 Vol. 5 Num. 1

Review of Deep Learning-based Pedestrian Re-identification Research

Mingda Yang[1], Wenzhun Huang[1,*], Ruixiang Li[1], Chengyu Hu[1]

[1]School of Electronic Information, Xijing University, Xi'an, China
[*]Corresponding author

Keywords: Pedestrian Re-identification, Deep Learning, Convolutional Neural Networks, Attention Mechanisms

Abstract: Pedestrian re-identification is dedicated to recognizing the same individual across different cameras and fields of view. With the development of deep learning methodologies and computational capabilities, deep learning and related approaches have been widely applied in the research domain of pedestrian re-identification. By summarizing the latest advancements in the application of deep learning to pedestrian re-identification, this paper categorizes the research of scholars both domestically and internationally in recent years. It analyzes different algorithms, datasets, and performance evaluation metrics, compares the strengths and weaknesses of various methods, and thereby points out current research hotspots and the broad prospects for future development.

1. Introduction

Pedestrian re-identification, also known as Person Re-identification (Person ReID) or human re-identification, stands as a pivotal research direction within the domain of computer vision. It is dedicated to solving the identity recognition and matching issue of specified individuals across different fields of view within multi-camera systems. As illustrated in Figure 1, a quintessential application of pedestrian re-identification involves employing feature extraction and matching algorithms to compare a target individual in a single image with pedestrian images captured by a surveillance system, aiming to find the target individual with the highest feature similarity.

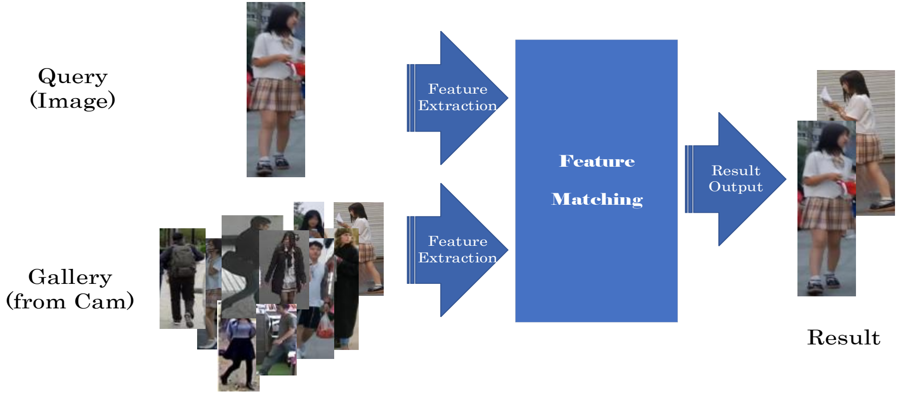


Figure 1: Example of person re-identification application scenarios

The task of pedestrian re-identification can be delineated into two main stages: feature extraction and feature matching. Feature extraction processes images or video clips through deep learning algorithms, such as Convolutional Neural Networks (CNNs), to obtain characteristics that can represent an individual's identity. Feature matching involves searching for the same individual across different times and spaces within large-scale data through distance metrics and ranking.

Owing to the diverse sources of data used for refining individual features, including static images, video sequences, mixed media information, and even text descriptions, coupled with the fact that the target individuals' data often originates from different surveillance cameras, numerous challenges may arise in practical research. These challenges encompass variations in lighting conditions, diversity in camera shooting angles, inconsistencies in image resolution, and changes in the posture of the target individuals. Moreover, environmental factors such as obstructions by other individuals or objects, as well as long-term appearance changes of the target individuals (e.g., clothing, hairstyles), further complicate the identification process. Consequently, extracting more robust and distinctive features through deep learning methods is beneficial for enhancing the accuracy of pedestrian re-identification systems when facing these challenges. Furthermore, by considering the impact of environmental factors and long-term appearance changes comprehensively, through multi-scale feature analysis and feature fusion strategies, the adaptability of pedestrian re-identification to complex scenes can be significantly improved, ensuring the generalization capability and practical application value of pedestrian re-identification technology in real systems.

Prior to 2016, the main feature extraction methods in pedestrian re-identification research included texture features (LBP, Gabor features, etc.), shape features (HOG, etc.), and color features (RGB, HSV, etc.), focusing on extracting low-level visual information from images to distinguish different pedestrian identities through intuitive features.

Following 2016, the field of pedestrian re-identification witnessed a significant shift from traditional methods to deep learning approaches. This shift encompassed the introduction of deep learning methodologies such as region-based feature extraction methods, attention mechanisms, Generative Adversarial Networks (GAN), and unsupervised learning. Concurrently, as the quality and scale of related datasets significantly improved, pedestrian re-identification technology entered a new era of rapid development.

2. Deep Learning-based Methods for Pedestrian Re-identification

2.1. Supervised Learning

In pedestrian re-identification tasks, models obtain feature vectors through feature extraction from input images, automatically learning high-level feature representations from image data that effectively differentiate between individuals. These feature vectors are then used for feature comparison to identify and match the same individual across different images. Due to the requirement for large volumes of labeled samples for training deep learning models, a substantial amount of manual labor is typically required. Thus, based on the data labeling situation, feature extraction methods can be categorized into supervised learning, weakly supervised learning, and unsupervised learning.

In the field of computer vision, Convolutional Neural Networks (CNNs) are the most commonly used deep learning models for extracting features from static images. These models can automatically learn high-level features representing individual identities, such as appearance, clothing style, and color distribution. CNN models like ResNet, VGG, Inception, and AlexNet have been widely applied in pedestrian re-identification, primarily for processing images. In scenarios involving video data, pedestrian re-identification needs to consider temporal sequence information, i.e., the dynamic changes of an individual across video frames. For video data, besides using CNNs to extract features from each frame, Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs) are often combined to process temporal sequence data, capturing the dynamic characteristics of individuals over time.

Ahmed S.M et al.[1] utilized a Hypothesis Transfer Learning (HTL) model adaptation approach, leveraging limited labeled data and previously learned source models to transfer knowledge, a method that spans both supervised and weakly supervised learning.

Xuesong Chen et al.[2] introduced the Salience-guided Cascaded Suppression Network (SCSN) method, enabling the model to mine diverse salient features and integrate these features into the final representation in a cascading manner. This network employs a cascading suppression strategy to progressively uncover diverse potentially useful features that might be obscured by other salient features, integrating different feature embeddings at each stage to obtain a final pedestrian representation with high discriminability. The schematic diagram of SCSN is shown in Figure 2[2].

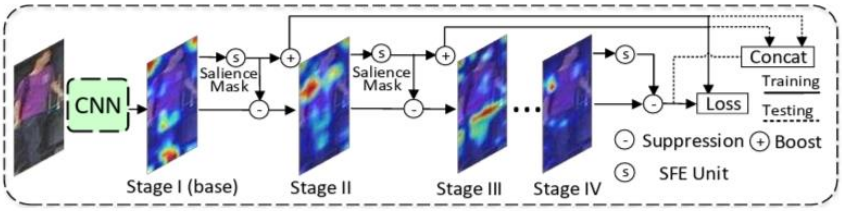


Figure 2: The insight of the Salience-guided Cascaded Suppres sion Network (SCSN).

2.2. Weakly Supervised Learning

While most existing pedestrian re-identification methods have achieved high accuracy in supervised settings, they require extensive annotated data for model training. In practical applications, especially in target domains (e.g., new surveillance scenarios) where identity labels are unavailable, this approach faces significant limitations.

Researchers have actively explored unsupervised or weakly supervised learning methods to address the challenge of requiring large amounts of annotated data in real-world applications of pedestrian re-identification. Weakly supervised learning leverages limited or imprecise annotation information to guide the model learning process, reducing the dependence on large-scale, precisely annotated data. Dengpan Fu et al. [3] applied a simple online multi-object tracking system to the existing unlabeled Re-ID dataset "LUPerson" and constructed a noisy-labeled variant dataset "LUPerson-NL." The ID labels automatically derived from trajectories inevitably contain noise. Dengpan Fu and colleagues developed a large-scale pretraining framework (PNL) that utilizes spatial and temporal correlations as weak supervision.

2.3. Unsupervised Learning

To reduce the dependency on costly and time-consuming manual annotation processes and to enhance the deployability and scalability of pedestrian re-identification systems in new environments, unsupervised learning utilizes clustering techniques, self-learning methods, and domain adaptation to perform effective feature extraction and identity discernment. Yoonki Cho et al. [4] proposed a part-based pseudo-label refinement method that improves the learning efficiency and accuracy during the unsupervised learning process by refining pseudo-labels.

Haocong Rao and Chunyan Miao introduced [5] a generic Transformer-based skeletal graph prototype contrastive learning (TranSG) method, combined with structure-trajectory hint reconstruction, to comprehensively capture valuable spatio-temporal semantics in skeletal relationships and skeletal graphs for person re-ID. They also verified the universality of this method in unsupervised scenarios.The schematic diagram of TranSG applied to person re-identification (re-ID) is shown in Figure 3[5].

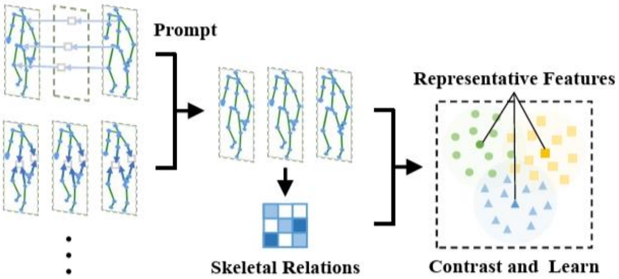


Figure 3: TranSG integrates into the contrastive learning of typical features for person re-ID.

Yutian Lin et al. [6] adopted an iterative training mechanism that forgoes clustering, which also enhances the robustness to hyperparameter variations.

Yunpeng Zhai et al. [7] introduced an enhanced discriminative clustering technique (AD-Cluster), which trains through iterative density-based clustering, adaptive sample augmentation, and discriminative feature learning. It learns an image generator and a feature encoder aiming to maximize intra-cluster diversity in the sample space and minimize intra-cluster distance in the feature space in an adversarial minimax manner.

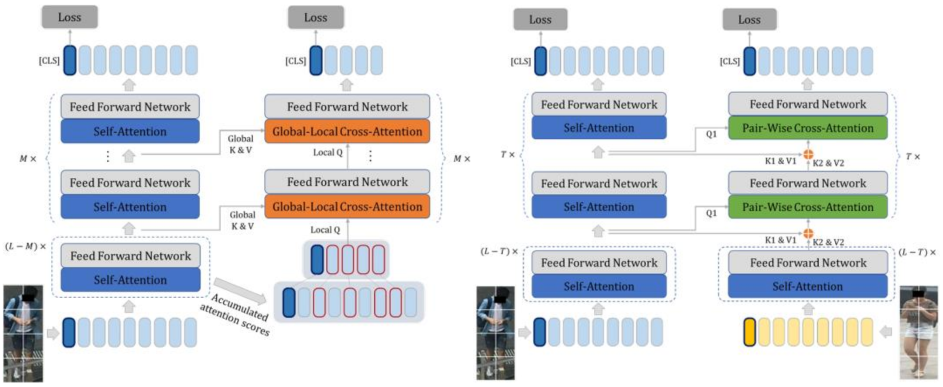
Guangcong Wang et al. [8] proposed a Smoothing Adversarial Domain Attack (SADA) method to reduce the gap between labeled source domains and unlabeled target domains, using a trained camera classifier to guide the alignment of source domain images with target domain images. To stabilize memory traces after initially acquiring cross-domain knowledge transfer from the source domain, a p-Memory Reconsolidation (pMR) method was introduced. During self-training in the target domain, source knowledge is reconsolidated with a small probability p.

2.4. Attention Mechanisms

Attention mechanisms, by emulating the human visual focusing mechanism, allocate different weights to differentiate the importance of input information. In the computational process, higher weights are assigned to more critical information, significantly enhancing the capability of pedestrian re-identification models to process key features.

Zhizheng Zhang et al. [9] introduced Multi-Granularity Reference-Aided Attentive Feature Aggregation (MG-RAFA), aimed at enhancing the performance of person re-identification in videos by considering features of different granularities and utilizing reference information. It effectively aggregates features from different frames of a video sequence, reinforcing useful features while suppressing irrelevant or distracting information, thereby obtaining a robust video-level representation of individuals.

Haowei Zhu et al. [10] proposed a Dual Cross-Attention Learning (DCAL) algorithm, complemented by self-attention learning. It employs Global-Local Cross-Attention (GLCA) to enhance the interaction between the global image and local high-response regions and uses Pairwise Cross-Attention (PWCA) to establish interactions between image pairs. This reduces misleading attention and spreads the attention response, uncovering more complementary parts for recognition.The overview of GLCA and PWCA is shown in Figure 4[10].



(a) Global-Local Cross-Attention (GLCA) (b) Pair-Wise Cross-Attention (PWCA)

Figure 4: Overview of GLCA and PWCA.

Haochen Wang et al. introduced the NFormer network, incorporating Landmark Agent Attention and Reciprocal Neighbor Softmax modules. By utilizing a few landmarks in the feature space for low-rank decomposition, it effectively simulates the relationship graph between images. This approach mitigates the interference from irrelevant representations and further alleviates computational burdens.

2.5. Addressing Occlusions

In practical applications, cameras often capture scenes where targets are obstructed, posing a significant challenge for pedestrian re-identification. Shang Gao et al. proposed a Pose-guided Visible Part Matching (PVPM) method capable of jointly learning discriminative features within an end-to-end framework. It utilizes a pose-guided attention mechanism that enables the model to focus on visible, unobstructed parts of the image, thereby reducing the interference of occluded sections on the feature extraction and matching process.

Guan’an Wang et al. employed a convolutional neural network (CNN) backbone and a keypoint estimation model to extract semantic local features, treating the local features of an image as nodes of a graph. They introduced an Adaptive Directional Graph Convolution (ADGC) layer to pass relational information between nodes. The proposed ADGC layer can automatically suppress the message passing of meaningless features by dynamically learning the direction and degree of links, introducing higher-order information processing to tackle occlusion issues, thus providing an effective solution for re-identification tasks involving occluded individuals.

Yingji Zhong et al., aiming to solve issues caused by occlusions and background, such as misaligned detection bounding boxes, proposed an Align-to-Part Network (APNet) for person detection and re-identification (reID). It refines detected bounding boxes to cover the estimated whole body area, from which discriminative part features can be extracted and aligned. Additionally, they contributed a Large-Scale dataset for Person Search in the wild (LSPS).

Yichao Yan et al. introduced the Multi-Granular Hypergraph (MGH), which builds hypergraphs of different spatial granularities through part-based features at various levels within video sequences. It connects hyperedges, which are a set of graph nodes (i.e., part-based features) over different time spans, capturing various temporal granularities to address misalignment and occlusion issues.

2.6. Cross-Modal Pedestrian Re-identification

Cross-modal pedestrian re-identification refers to the process of dealing with data from different modalities (visible light images, infrared images) in pedestrian re-identification tasks, addressing limitations that may arise from using single-modality data in specific environments, such as performance degradation due to significant lighting differences between day and night.

Jiawei Feng et al. introduced the Shape-Erased Feature Learning paradigm, which decorrelates modality-shared features across two orthogonal subspaces. It co-learns shape-related features in one subspace while learning shape-erased features in the orthogonal complement space. By maximizing the conditional mutual information between shape-erased features and identity, while discarding body shape information, this approach explicitly enhances the diversity of the learned representations.

Yukang Zhang et al. proposed the Modality Restitution and Compensation Network (MRCN) to narrow the gap between the two modalities of visible light and infrared images, extracting features that are both modality-independent and modality-specific. It restores normalized visible light and infrared features using modality-independent features, while compensating for features of one modality with those related to the other modality.

2.7. Clothing Variation

Addressing pedestrian re-identification over long time spans with clothing changes poses a significant challenge but is also a crucial step towards transitioning research to practical applications. Currently, methods such as gait recognition, head recognition, facial recognition, and silhouette feature extraction can to some extent handle pedestrian re-identification with clothing changes.

Shijie Yu et al. introduced a dual-branch network named Biometric-Clothes Network (BC-Net), which effectively combines biometric and clothing features to accommodate re-identification under clothing variation.

Lijie Fan et al. explored the use of Radio Frequency (RF) signals that penetrate clothes and reflect off the human body, enabling the extraction of more persistent human identification features such as body shape and form. This research offers a new perspective in the field of pedestrian re-identification, showcasing the immense potential and application prospects of RF signals in ReID.

3. Datasets

3.1. Image Datasets

In the field of pedestrian re-identification, datasets composed of static images provide rich material for research. The VIPeR dataset includes images of 632 different pedestrians, each captured from two different angles, totaling 1264 images, primarily focusing on full-body images of pedestrians. The iLIDS dataset, with images sourced from multiple CCTV cameras in an airport's arrival hall, includes complex backgrounds, varying lighting conditions, and occlusions. It comprises images of 119 different individuals from 34 different cameras, with multiple images per individual, totaling over 600 images. The GRID dataset contains images of 250 pedestrians, covering multiple different perspectives, simulating the real surveillance system scenario where pedestrians can be captured from various directions and angles. The CUHK01 dataset includes 3,884 pedestrian images involving 971 different individuals, captured by multiple surveillance cameras within the campus of The Chinese University of Hong Kong. The CUHK02 dataset includes 7,264 images from 5 different camera views, involving 1,816 different individuals. The CUHK03 dataset introduces more images and higher challenges, containing 14,096 images involving 1,467 different individuals. In addition to providing manually annotated images, this dataset also includes pedestrian images automatically detected using the Deformable Part Model (DPM) detector. The RAiD dataset includes images from both indoor and outdoor environments with different lighting conditions and background complexities. The Market-1501 dataset, considered one of the important benchmarks for pedestrian re-identification research due to its large scale, diversity, and close-to-real surveillance scenario collection and annotation approach, contains 32,668 images from 6 different cameras, involving 1,501 different individuals.

The DukeMTMC-reID dataset, based on the DukeMTMC (Multi-Target, Multi-Camera) surveillance project, contains 36,411 images involving 1,404 different individuals. This dataset has been discontinued due to privacy concerns.

The Airport dataset focuses on pedestrian re-identification in airport scenarios, with images captured in different areas inside the airport, such as waiting halls, security checkpoints, and boarding gates. The MSMT17 dataset is renowned for its large data scale, diverse environmental conditions, and the high challenge of real-world application scenarios. It includes 126,441 images from 15 cameras (12 outdoor and 3 indoor), involving 4,101 different pedestrians. The COCAS dataset, designed specifically to address the clothing change issue in Re-ID, includes 5,266 identity labels with an average of 12 images per identity, captured by 30 cameras in various real-world scenes. The SYSU-30k dataset is a large-scale dataset designed for weakly supervised pedestrian re-identification research, containing identities of approximately 30,000 different individuals and over 2.9 million images. The Black-reID dataset focuses on pedestrian re-identification under low-light conditions or when individuals wear black clothing, containing a training set of 1,274 identities. This dataset aims to facilitate pedestrian re-identification by leveraging head-shoulder features instead of relying on clothing information, addressing the significant issue of missing clothing attribute information under these conditions.

3.2. Video Datasets

Video datasets offer dynamic information and temporal variations, placing higher demands on the real-time processing capabilities of algorithms. The CAVIAR dataset consists of video sequences captured by CCTV cameras in two shopping centers and a public square, providing diverse backgrounds and environmental conditions. The PRID-2011 dataset features two cameras located at different places, capturing pedestrian images with varying backgrounds and lighting conditions. It includes images of 400 pedestrians, 200 of whom were captured by both cameras. The iLIDS-VID dataset, recorded under multiple surveillance cameras at an airport, reflects the various challenges of the real world, including different lighting conditions, complex backgrounds, and various changes in pedestrian postures.

The MARS dataset, proposed by the Chinese Academy of Sciences, automatically generates annotations using the DPM (Deformable Part Model) detector and GMMCP (Generalized Maximum Multi Clique Problem) tracker. It contains over 20,000 video clips from 1,261 different pedestrian identities. The LPW dataset emphasizes the capability of re-identifying pedestrians over a long time span, containing images of more than 7,000 pedestrians, with a total image count exceeding 590,000. The DukeMTMC-VideoReID dataset, built on the DukeMTMC dataset, provides a rich set of video sequence data to support cross-camera pedestrian identification research. The EgoReID dataset is designed for researching Re-ID issues in first-person viewpoint videos, featuring significant viewpoint changes, occlusions, and dynamic backgrounds. The LS-VID dataset is a large-scale video pedestrian re-identification dataset that includes 3,772 identity labels, 14,943 video sequences, and nearly 3,000,000 bounding boxes, covering about 200 different frames. It employs the Faster R-CNN detector for automatic bounding box annotations, which may enhance annotation efficiency and scene diversity coverage but could also introduce some noise.

4. Performance Evaluation Metrics

In the field of pedestrian re-identification, mAP (Mean Average Precision), Rank-1 accuracy, and CMC (Cumulative Matching Characteristic) curves are commonly used performance evaluation metrics .

mAP evaluates the quality of the ranked list returned by the model for each query. It is the average of the AP (Average Precision) values across all queries. The formula for mAP is represented as:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{1}$$

where Q is the total number of queries. AP is defined as:

$$AP = \frac{\sum_{k=1}^{n} (P_k \cdot rel(k))}{N} \tag{2}$$

Here, N is the total number of relevant retrievals, n is the total number of retrievals, P(k) is the precision at the top k retrieval results, and rel(k) indicates whether the kth retrieval result is relevant (1 for relevant, 0 for irrelevant).

Rank-1 accuracy is the proportion of correctly identified target pedestrians by the system among its top-ranked (i.e., first-ranked) match results.

CMC curves evaluate the identification accuracy of a pedestrian re-identification system at different ranks. It measures the probability that the correct match appears within the top N most similar list across all queries. The focus of CMC calculation is on the accuracy of ranking.

The formula for CMC is:

$$CMC(N) = \frac{1}{Q}\sum_{q=1}^{Q} 1(rank_q \leq N) \tag{3}$$

where Q is the number of query samples, $rank_q \leq N$ means the result is 1 if the correct match for query q appears within the top N positions; otherwise, it is 0.

5. Conclusion

With the rapid development of deep learning technologies, the field of pedestrian re-identification has made significant progress. However, practical research still faces multiple challenges, including the effects of illumination and angles, model generalization capabilities, handling heterogeneous data, and reducing the need for manual annotations. At the application level, challenges such as model generalization, handling heterogeneous data, rapid re-identification, model lightweighting, the impact of long time spans on features, and precision in processing surveillance videos remain to be addressed in the future. Future research directions include enhancing the capability to cross datasets and handle heterogeneous data, developing fast, effective, and lightweight models, cross-modal recognition, and exploring semi-supervised and unsupervised learning methods to reduce the dependency on extensive annotated data.

References

[1] Sk Miraj Ahmed, Aske R. Lejbølle, Rameswar Panda, et al. Camera on-boarding for person re-identification using hypothesis transfer learning[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020:12141-12150.

[2] Xuesong Chen, Canmiao Fu, Yong Zhao, et al. Salience-guided cascaded suppression network for person re-identification [J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020: 3297-3307.

[3] Dengpan Fu, Dongdong Chen, Hao Yang, et al. Large-Scale Pre-training for Person Re-identification with Noisy Labels[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022:01-11.

[4] Y. Cho, W. J. Kim, S. Hong and S. -E. Yoon.Part-based Pseudo Label Refinement for Unsupervised Person Re-identification[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022:7298-7308.

[5] H. Rao and C. Miao.TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification[C].2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023:22118-22128

[6] Yutian Lin, Lingxi Xie, Yu Wu, et al.Unsupervised person re-identification via softened similarity learning[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020:3387-3396.

[7] Yunpeng Zhai,Shijian Lu,Qixiang Ye,et al.AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-Identification[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2020:9018-9027.

[8] Guangcong Wang,Jian-Huang Lai,Wenqi Liang,et al.Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2020,10565-10574.

[9] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, et al.Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification[J].Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, 10404-10413.

[10] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian and Y. Shan.Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022: 4682-4692.