Pedestrian Re-Identification Research Plan

Contents

Abstract—The research plan for pedestrian re-identification (ReID) focuses on advancing computer vision technology to accurately identify individuals across different camera views. Given the challenges in face recognition due to low-resolution images, ReID plays a critical role in security applications like surveillance and lost person rescue. The study is centered around two main methodologies: representation learning and metric learning. Representation learning employs convolutional neural networks (CNNs) for feature extraction, treating ReID as a classification or verification task. Metric learning, on the other hand, is designed to learn the similarity between images, aiming to minimize the distance between images of the same person while maximizing the distance between different individuals.

I. Backgroud

Pedestrian Re-identification, also known as Pedestrian Re-Identification, or ReID for short, is a technology that uses computer vision technology to determine whether there is a specific pedestrian in an image or video sequence. Widely regarded as a sub-problem of image retrieval. Given a surveillance pedestrian image, retrieve the pedestrian image across devices. Pedestrian re-identification is generally only applied when face recognition fails, but in the real world, the probability of face recognition failure is very high, because face recognition generally requires a 96*96 pixel face, we can get A more reliable performance. At least not less than 32*32 pixels. But usually we can't get such high-definition photos, so it is very necessary to study pedestrian re-identification. Pedestrian re-identification is generally used in the field of security, such as video surveillance, criminal investigation, danger warning, unmanned supermarkets, lost rescue and other scenarios.

II. Literature Review

There are many algorithms for person re-identification, among which representation learning and metric learning are the most basic. The key question is how the convolutional neural network is trained? To train a network, we know that a loss function is needed, and representation learning and metric learning are classified according to the difference of this loss function. Representation learning-based methods are a very common class of person re-identification methods. This is mainly due to the rapid development of deep learning, especially convolutional neural networks. Since CNN can automatically extract representational features from the original image data according to the task requirements, some researchers regard the pedestrian re-identification problem as a classification problem or a verification problem. The classification problem refers to the use of pedestrian IDs or attributes as training labels to train the model; the verification problem refers to inputting a pair of pedestrian images, and letting the network learn whether the two images belong to the same pedestrian. Using classification/identification loss and verification loss to train the network, the network input is several pairs of pedestrian images, including classification sub-network and verification sub-network. The classification sub-network performs ID prediction on the image, and calculates the classification error loss according to the predicted ID. The verification sub-network fuses the features of the two pictures to determine whether the two pictures belong to the same pedestrian. The sub-network is essentially equivalent to a two-class network. After training with enough data and inputting a test image again, the network will automatically extract a feature, which is used for the pedestrian re-identification task[1]. Only relying on the ID information of pedestrians is not enough to learn a model with sufficient generalization ability. Additional attribute features of pedestrian images, such as gender, hair, clothing and other attributes. By introducing pedestrian attribute labels, the model must not only accurately predict the pedestrian ID, but also predict the correct pedestrian attributes, which greatly increases the generalization ability of the model, and most experimental results also show that this method is effective. The features output by the network are not only used to predict the ID information of pedestrians, but also to predict various pedestrian attributes. By combining ID loss and attribute loss, the generalization ability of the network can be improved[2]. In general, there is still a lot of work based on representation learning, which has become a very important baseline in the ReID field, and the method of representation learning is relatively robust, the training is relatively stable, and the results are relatively easy to reproduce. However, representation learning is prone to overfitting on the domain of the dataset, and when the training ID increases to a certain extent, it becomes relatively weak. Usually additional FC layers are needed to tutor feature learning, and the FC layers are discarded during testing. The FC layer dimension of ID loss is consistent with the number of IDs. When the training set is too large, the network is huge and the training convergence is difficult. In the verification loss test, a pair of pictures needs to be input, and the recognition efficiency is very low. Distributed training for representation learning is generally more mature. Metric learning is a method widely used in the field of image retrieval. Unlike representation learning, metric learning aims to learn the similarity between two images through the network. In the problem of pedestrian re-identification, the similarity of different pictures of the same pedestrian is greater than that of different pictures of different pedestrians. The final loss function of the network makes the distance between the same pedestrian image (positive sample pair) as small as possible, and the distance between different pedestrian images (negative sample pair) as large as possible. Commonly used metric learning loss methods include contrastive loss, triplet loss, quadruple loss, hard sample sampling triplet loss, and boundary mining loss. The hard sample triplet loss is an improved version of the triplet loss. The traditional triplet randomly samples three images from the training data. Although this method is relatively simple, most of the sampled pairs are simple and easily distinguishable sample pairs. If a large number of training sample pairs are simple sample pairs, then it is not conducive to the network to learn better representations. A large number of papers have found that training the network with harder samples can improve the generalization ability of the network, and there are many ways to sample pairs of difficult samples. One of these methods is the batch-based online hard-to-sample sampling method. For each training batch, randomly select P pedestrians with IDs, and each pedestrian randomly selects K different pictures, that is, a batch contains P*K pictures, and then for each picture a in the batch, you can choose the most A hard positive example and a hardest negative example and a form a triplet. By defining the TriHard loss, the triple loss is calculated. Usually, the TriHard loss is better than the traditional triple loss[3]. Directly learn the similarity between pictures by constructing a network. No additional FC layers are needed to tutor feature learning, the FC layers in the testing phase are discarded. Network size is independent of training set size, but data sampler time consumption increases. TriHard Loss is the current benchmark for metric learning in the industry. Metric learning is usually trained randomly and requires a certain amount of training experience. Distributed training of metric learning is not mature, and usually needs to implement part of the code by itself. In the early ReID research, everyone mainly focused on the global global feature, which is to use the whole image to obtain a feature vector for image retrieval. But then everyone gradually found that the global feature encountered a bottleneck, so they began to gradually study local local features. Commonly used ideas for extracting local features mainly include image slicing, positioning using skeleton key points, and posture correction. The global feature refers to the feature extraction of the global information of each pedestrian image, and this global feature does not have any spatial information. In order to solve the problem of global features, local features are proposed, slices and poses are the most used, and the effect is the best. Image segmentation is a very common way to extract local features. Each image is characterized by the CNN network, and the local features are input to the LSTM network in sequence, and the segmented image blocks are sent to a long and short-term memory network in sequence, and are automatically expressed as the final features of the image. The final feature fuses all Local features of image patches. However, this disadvantage is that the requirements for image alignment are relatively high. If the two images are not aligned up and down, then the phenomenon of head and upper body contrast is likely to occur, which will make the model judge wrong. Horizontal dicing is an early work and is rarely used now. Because vertical cutting is more in line with our intuitive feeling for human body recognition, horizontal cutting is rarely used in the field of pedestrian re-identification. We often find a phenomenon that two people are very similar, but these two people are not the same person. This gives us some inspiration. We hope to pay more attention to similar areas and go to similar areas to find out if there are any different things. This is also The reason we do normalization, I hope that the smaller the distance between horizontal blocks, the greater their gradients, the greater the gradients, the greater the contribution to the network, and the greater the network contribution, the network will focus on similar areas. In order to solve the problem of manual image slicing failure when the images are not aligned, some papers use some prior knowledge to align pedestrians first. These prior knowledge are mainly pre-trained human pose (Pose) and skeleton key point (Skeleton) models. The keypoints of pedestrians are estimated by the pose estimation model first, and then the same keypoints are aligned by affine transformation. A pedestrian is usually divided into 14 key points, and these 14 key points divide the human body results into several regions. To extract local features at different scales, the authors set three different PoseBox combinations. Afterwards, the three PoseBox corrected pictures and the original corrected pictures are sent to the network to extract features, which contain global information and local information. In particular, this affine transformation can be performed in preprocessing before entering the network, or after inputting into the network. If it is the latter, it is necessary to make an improvement on the affine transformation, because the traditional affine transformation is not steerable. In order to make the network trainable, a derivable approximate radiation variation needs to be introduced[4]. It feels a bit sloppy to use local features as independent features and use a small piece of information to calculate ReID loss, but the final experiment proves that this method is very effective[5]. Using a pose estimation model to obtain (14) key pose points of pedestrians. Obtain the part area with semantic information according to the pose points. Extract local features for each part region. Combining local and global features often yields better results. At present, single-frame ReID research is still the mainstream, because the data set is relatively small, even a single-GPU PC will not take too long to do an experiment. But usually the information of a single frame image is limited, so many works focus on the use of video sequences for the study of person re-identification methods. The main difference between the methods based on video sequences is that these methods not only consider the content information of the image, but also consider the motion information between frames. The main idea of the single-frame image-based method is to use CNN to extract the spatial features of the image, while the main idea of the video sequence-based method is to use CNN to extract spatial features and use Recurrent neural networks (RNN) to extract time series features. The network input is a sequence of images. Each image is passed through a shared CNN to extract image spatial content features, and then these feature vectors are input into an RNN network to extract the final features. The final feature fuses the content features of a single frame image and the frame-to-frame motion features. And this feature is used to train the network in place of the image features of the previous single-frame method. One of the representative methods of the video sequence class is the Accumulative Motion Context Network (AMOC). The AMOC input includes the original image sequence and the extracted optical flow sequence. Usually, traditional optical flow extraction algorithms are required to extract optical flow information, but these algorithms are computationally time-consuming and cannot be compatible with deep learning networks. In order to obtain a network that automatically extracts optical flow, the author first trains a motion information network (Motion network, Moti Nets). The input of this motion network is the original image sequence, and the label is the optical flow sequence extracted by traditional methods. The original image sequence is shown in the first row, and the extracted optical flow sequence is shown in the second row. The network has three outputs of optical flow prediction, Pred1, Pred2, and Pred3, which can predict three different scales of optical flow. Finally, the network fuses the optical flow prediction outputs at the three scales to obtain the final optical flow map, and the predicted optical flow sequence is shown in the third row. By minimizing the error of predicting and extracting the optical flow map, the network can extract more accurate motion features[6]. A very big problem with ReID is that it is difficult to obtain data. The largest ReID data set is only a few thousand IDs and tens of thousands of pictures. There is a big problem in the early papers using the GAN method to generate images. The images generated by the early GAN are uncontrollable and poor quality samples. This problem was later solved by CycleGAN. A typical CycleGAN paper, this paper proposes a method to generate very realistic new sample images. The process of this method mainly has two steps: the first step is to use a multi-branch network to decompose the image into three components (pose, foreground and background) and encode the features respectively; the second step uses an adversarial way to learn three corresponding mapping functions , which maps Gaussian noise to the learned embedding feature space. These three components of the image can be obtained by using the existing network, and new embedded features can be sampled to generate new target images, making the generation process more controllable. The paper has conducted experiments on both the Market1501 and Deepfashion datasets. The results show that the generated samples are not only very real, but also can improve the performance of the ReID model as training data[7].

III. Research Problem

In the field of pedestrian re-identification, several key questions need to be addressed to improve the accuracy, robustness, real-time performance, and privacy protection. This research aims to investigate the following aspects:

a. Feature representation: How can pedestrian image features be effectively extracted to achieve accurate re-identification? Can deep learning methods be utilized to learn more discriminative feature representations?

The accurate extraction of informative and discriminative features from pedestrian images is crucial for achieving precise re-identification. This research will explore novel techniques, such as leveraging deep learning-based approaches, to extract features that capture unique characteristics of individuals while being robust to variations in appearance, pose, and lighting conditions. Deep learning methods have shown promise in learning more discriminative feature representations, and their potential will be further explored.

b. Video sequence modeling: How can useful spatiotemporal information be extracted from pedestrian video sequences to improve the accuracy and robustness of pedestrian re-identification?

By exploiting the temporal information present in video sequences, this research aims to develop methods that effectively model the motion and appearance variations of pedestrians over time. This will involve exploring techniques such as recurrent neural networks or 3D convolutional neural networks to capture the temporal dynamics and spatial correlations within video sequences. The goal is to leverage the rich spatiotemporal information in videos to enhance the accuracy and robustness of pedestrian re-identification.

c. Cross-domain/cross-camera re-identification: How can the performance of pedestrian re-identification be maintained across different domains, lighting conditions, viewpoints, and camera setups?

To address the challenges posed by domain shifts and camera variations, this research will investigate domain adaptation techniques and explore methods to learn domain-invariant representations. This will involve developing algorithms that can generalize well across diverse scenarios, ensuring robustness in cross-domain and cross-camera re-identification scenarios. The goal is to achieve consistent performance regardless of changes in domains, lighting conditions, viewpoints, or camera setups.

d. Real-time performance: How can the real-time performance of pedestrian re-identification algorithms be improved to handle high-speed video streams or large-scale pedestrian databases?

Efficient algorithms play a crucial role in real-time pedestrian re-identification systems. This research will focus on developing lightweight network architectures, optimizing feature extraction and matching processes, and exploring hardware acceleration techniques to achieve real-time performance without compromising accuracy. The aim is to enable the deployment of pedestrian re-identification systems that can handle high-speed video streams or large-scale pedestrian databases in real-time scenarios.

e. Privacy protection: How can personal privacy be protected during the pedestrian re-identification process to prevent misuse and infringement of individual rights?

Respecting privacy concerns is essential in the deployment of pedestrian re-identification systems. This research will investigate privacy-preserving methods, such as using anonymized representations, secure protocols for data sharing, or generative models for privacy-preserving feature generation. The aim is to develop techniques that balance the need for accurate re-identification with protecting the privacy of individuals. Ensuring privacy protection will help prevent misuse or infringement of individual rights when deploying pedestrian re-identification systems.

By addressing these research questions, this study aims to advance the field of pedestrian re-identification by proposing novel algorithms and techniques that improve the accuracy, robustness, real-time performance, and privacy protection aspects of existing methods. The outcomes of this research will have implications for video surveillance, public safety, and other applications requiring reliable and privacy-aware pedestrian re-identification systems.

IV. Research Purposes

The main objectives of pedestrian re-identification are as follows:

1. Cross-camera matching: Accurately identifying and matching the same individual across different cameras. This is crucial for tasks such as pedestrian tracking, security surveillance, and criminal investigations.

2. Cross-time matching: Recognizing and matching the same person in pedestrian images captured at different time points. This is important for tracking pedestrian trajectories, analyzing the temporal order of events, and conducting behavior analysis.

3. Robustness: Maintaining accuracy and robustness in the face of challenges such as varying lighting conditions, viewpoint changes, pose variations, occlusions, and low resolutions. This is essential for dealing with complex real-world scenarios and uncontrollable environmental changes.

4. Real-time performance: Conducting pedestrian re-identification rapidly in real-time video streams to accommodate applications that require quick responses, such as video surveillance and real-time security systems.

5. Privacy protection: Respecting individual privacy rights and implementing effective measures to prevent misuse and infringement of personal information during the pedestrian re-identification process.

In summary, the primary goals of pedestrian re-identification involve reliably identifying and matching pedestrian identities in diverse scenarios through accurate, robust, real-time, and privacy-aware approaches. This aims to enhance public safety, video surveillance, and other related applications.

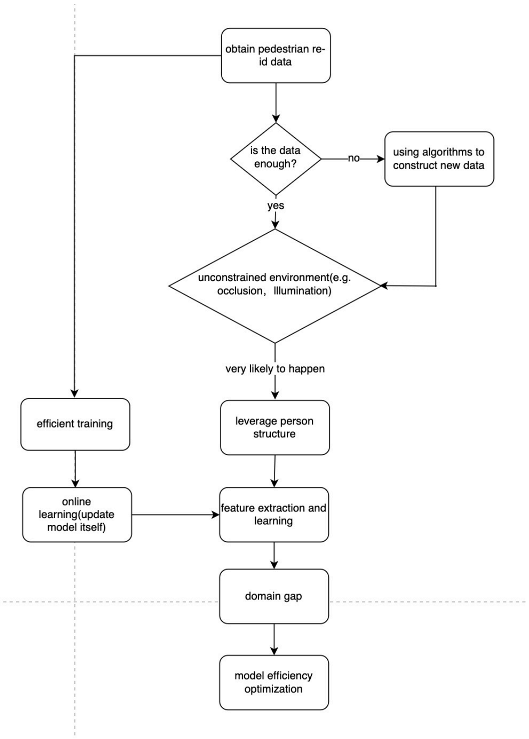
V. Research methods and steps

2022 Cvpr Research Directions Of Related Papers In The Direction Of Pedestrian Re-Identification By reading the related papers on the direction of pedestrian re-recognition in 2022cvpr, we can find that the following ideas are involved

1. Unsupervised clustering

2. Joint training
3. Generating synthetic samples with GAN
4. Large-scale distributed to improve generalization ability
5. Representation learning
6. Video sequence extraction, sequence re-identification, establishment of spatiotemporal patterns, multimodality
7. Cross attention mechanism

VI. Research design


obtain pedestrian re-id data
is the data enough?
no
using algorithms to construct new data
yes
unconstrained environment(e.g. occlusion, Illumination)
very likely to happen
efficient training
leverage person structure
online learning(update model itself)
feature extraction and learning
domain gap
model efficiency optimization

From the above figure, you can get a general understanding of the general process of the pedestrian re-identification system. Next, I will explain each module in detail.In real life, there are roughly three steps required to implement ReID. The first step is to obtain the original picture from the surveillance video of the camera;The second step is to detect the position of the pedestrian based on these original images;The third step is to use ReID technology to calculate the distance of the image based on the detected pedestrian image, but our research is based on common data sets. The previous two parts of image collection and pedestrian detection have been done. ReID mainly studies the first part three phases.The general realization idea of reid In fact, it is to extract the features first, and then calculate and compare the feature distances according to the retrieval map and the images in the image library, such as Euclidean distance, and then sort according to the calculated distances. The higher the ranking, the higher the similarity rate.I just mentioned that the core is the process of abstracting images into features. During training, we generally design a certain loss function, try to minimize the loss function in the training stage, and the minimization process reversely trains the features to be more meaningful, and the loss function is not considered in the evaluation stage.Because many research topics of ReID are now modified based on the Resnet50 structure. So it is necessary to understand the network structure of Resnet50, resnet is generally divided into five layers, the length and width of the feature map output by each layer will be reduced by half compared to the previous layer, pooling at the end, The so-called pooling is to take a maximum or average value in each feature map. Finally, classify based on this feature.

VII. Data collection and specific solutions

Common Datasets And Dataset Characteristics: Three mainstream public datasets used for ReId algorithm training and evaluation in academic research, They are Market1501, DukeMTMC-reID, CUHK03. After reading these data sets, you should be able to have an intuitive feeling, that is, in the ReID research, the number of picture sets is about tens of thousands, and the number of IDs is basically less than 2000, and the number of cameras is about 10 or less, and most of these photos are from schools, so their identities are mostly students. This can be compared with the current face datasets. The face datasets are often millions or tens of millions of photos. A dataset with many face IDs can be in the millions, and the identities are very diverse. This is actually a more realistic situation when ReID faces such complex problems as before, but with so little data. Solutions With Limited Data: In person re-identification system, most of the cases have to face the situation of missing data this leads to the basic treatment plan: 1.3D Game Engine 2.Generative Adversarial Network. Unconstrained Environment: In real-world situations, occlusion problems often occur in unconstrained environment: The simple idea can divide the picture into occluded part and non-occluded part, and perform similarity matching on these two parts respectively. Introduce spatial information Mapping the human body into 3D space, incorporating geometric structures, and dealing with occlusion issues. Introduce multimodal input to improve robustness. After data preprocessing, we will enter the step of feature extraction and learning, this is the most core step, roughly divided into three core ideas, representation learning, metric learning, local feature learning: To achieve representation learning, I will use the following method to take the output features of the upper and lower parts of the classification loss as input, input them into the contrast loss, and implement the supervised learning of the network by designing the classification loss and contrast loss. For metric learning, I will use the common ternary loss to do it. Generally, three pictures are selected. Two of the three pictures are of the same person, and the other picture is not of the same person. The purpose is to make similar distances closer, the distance between different classes is farther. Regarding the algorithm for selecting pictures, I will use the Batchhard algorithm to randomly select p individuals from the data set, each person's k pictures form a batch, and each person's k pictures form a k*(k-1) ap pair, Then take a negative that is closer to the ap from the rest of the people to form an apn group, and then this model makes the loss composed of apn as small as possible. For local feature learning, I will use a multi-granularity network to combine global features with multi-granularity local features. The global features are responsible for the overall macro-common features. Then we divide the image into different blocks, each with a different granularity. It is responsible for the extraction of features at different levels or at different levels. Part-based pixel alignment at the pixel level of the image in the most direct way based on local part feature learning. Based on local component feature learning and feature alignment, not at the pixel level, but at the feature level, a more interpretable feature representation method is formed. Different loss functions are used to identify pedestrian identities from high-dimensional redundant features and form feature representations with stronger discriminative ability. The previous methods are all based on a single image, but there are still continuous frames of video in the actual application scene. Continuous frames can obtain more information and are closer to the actual application. I will use the classic idea in this part of the experiment, cnn combined with rnn, each time Each image goes through a shared CNN to extract image spatial content features, and then these feature vectors are input to an RNN network to extract the final features. The final features fuse the content features of a single frame image and the motion features between frames. In the sorting optimization stage, after obtaining the initial search sorting results with the learned Re-ID features, the similarity relationship between pictures is used to optimize the initial search results, mainly including reordering and sorting fusion, etc. In the actual scene, the accuracy of the model is often reduced due to the camera angle, clothing, lighting and weather, which tests the generalization ability of the model. Usually, when a trained model is directly applied to a new scene or a new dataset, the recognition accuracy is generally low. The data of the new scene is not manually labeled, but pseudo-labels are assigned to the new data by clustering, and finally the existing model is fine-tuned. Use prior knowledge and soft labels to further mine information in the target domain. We all want the model to be efficient and fast to get the final result To improve model performance Generally there are the following methods Model pruning. Using auto-ml to automate the design of neural network architectures for pedestrian re-identification problems. Efficient training from millions of data trying to decide which data is important, we often run into long tail problems, A small number of categories occupy the vast majority of samples, and a large number of categories have only a small number of samples. The same samples from different camera styles make up positive samples Learning camera invariance. The samples of different domains form negative sample pairs, learn the potential relationship of the domain in the feature space, and improve the domain connectivity. The solutions to the long tail problem are generally divided into four types:Re-sampling Re-weighting Learning strategy use a combination of the above 3 strategies. Online learning, we want to update the model itself. we hope that each time a part of the data is input, there is no need to retrain, which can save a lot of time and improve efficiency FTRL algorithm can be considered. Evaluation metrics for pedestrian re-id, there are two evaluation indicators that are used more in ReID: Rank1, MAP. ReID is still a sorting problem after all, and rank is the core index of sorting hit rate. rank1 is the first hit rate, that is, whether the first picture hits him, and rank5 is whether at least one of the 1-5 pictures hits him. A more comprehensive indicator for evaluating ReID technology is the mean mAP mean precision. Because Rank1 only needs the first hit, there are a series of accidental factors in it, and there are some fluctuations during model training or testing. But mAP measures ReID more comprehensively, why? Because it requires the retrieved person to rank all the pictures in the base library at the top, then the mAP index will be high. Therefore, mAP is an indicator that can comprehensively reflect the true level of this model.

VIII. Expected outcome

The expected outcomes of pedestrian re-identification are as follows:

Pedestrian identity recognition: Accurately identifying the identity of the same individual in different scenarios, ensuring consistency and uniqueness of pedestrian identities.

Cross-camera matching: Achieving accurate matching across different cameras, associating images of the same person in different scenes for pedestrian tracking and behavior analysis.

Robustness: Maintaining accuracy and robustness in the face of challenges such as lighting variations, viewpoint changes, pose variations, occlusions, and low resolutions, to adapt to complex real-world environments.

Real-time performance: Conducting pedestrian re-identification rapidly in real-time video streams to meet the requirements of applications that demand quick responses, such as real-time surveillance and security systems.

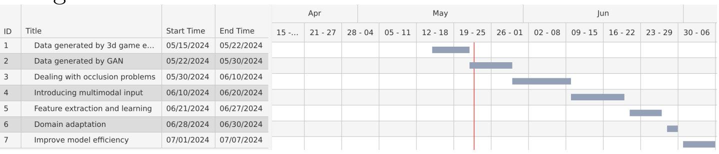
Privacy protection: Respecting individual privacy rights and ensuring the security and confidentiality of personal information during the pedestrian re-identification process.

A. Clear research plan

From the above process, we can understand the modules required to build a person re-identification system. Next, I will develop a research plan for each module. Limited training data for the problem of missing samples, I will use the following steps to practice: Generate human data from different perspectives through a 3D game engine. Use GAN to simulate the distribution of training data, generate images with GAN, alleviate the problem of limited data set data, allow the model to see more samples, and improve robustness. Learn about GAN networks, basic network structures, including generators, discriminators, and subsequent evolutions, such as CycleGAN, and loss functions, including discriminative losses, Generation loss, and some commonly used GAN methods, including CGAN, Pix2pix, their respective characteristics, GAN (uncontrollable, randomly generated sample images), CGAN (can Generate images with conditional constraints on GAN), Pix2pix (can convert paired images of A and B domains), CycleGAN (can convert any of A domain and B domain image for conversion). And in the case of small samples, the GAN network is used to generate images. Unconstrained environment dealing with occlusion problems in the real world, training part uses the attention map generated by keypoints to build a global representation, similar to a PCB, to build a local part representation, test part uses key points to distinguish occlusion and non-occlusion local representations, use non-occlusion representations for similarity matching, and ignore occluded parts. Because people live in a 3D world, the human body is mapped to the three-dimensional space, and then the point cloud is integrated into the geometric structure to learn the human body expression, obtain robust features, and deal with the occlusion problem. Because people live in a 3D world, the human body is mapped to the three-dimensional space, and then through the point cloud, the geometric structure is integrated to learn the human body expression, obtain robust features, deal with occlusion problems, and enhance the speed of reasoning. Introducing multimodal input, make the model closer to practical, such as searching for people through language description. Feature extraction and learning learn and practice in the following order: Understand the causes of misalignment of horizontal slices and solutions, in particular, various methods of feature alignment can indeed play a great role in improving the matching accuracy. such as AlignedReId. Starting from the most basic, perform pixel alignment of components at the pixel level of the image. Start from the feature level, read and practice the classic algorithm PCB. Form a more explanatory feature representation method, or use different loss functions to identify pedestrian identities from high-dimensional redundant features For example, using two loss combinations, verification + identification, it would be much better than using only one loss function. Domain adaptation: use unsupervised means obtaining pseudo-labels using clustering, to improve model performance. use homogeneous learning and heterogeneous learning to zoom in on positive samples and zoom out from negative samples. Model efficiency: understand the principles of network compression and acceleration, including front-end compression, such as knowledge distillation, compact model design, filtering-level pruning, back-end compression, such as low-rank approximation, unlimited pruning, parameter quantization and binarization.For example, the channel attenuation soft pruning model using the channel local correlation of the pretrained network, for the most commonly used ResNet-50 main network, compress the model size. use auto-ml to automatically design the neural network structure for the pedestrian re-identification problem, greatly reduce the amount of ResNet parameters, reduce the running cost, and automatically design the visualization of the pedestrian re-identification neural network structure. neural networks: understand the process of traditional neural network, initialize from network, provide training samples, forward calculation, reverse calculation, overfitting and underfitting solutions, deep convolutional network Network composition, including convolutional layers, pooling layers, fully connected layers, auxiliary layers, practical cases of convolutional neural networks, such as LeNet, try to implement CNN networks manually.

B. Conclusions

In general, in order to implement pedestrian re-identification, the following issues need to be addressed:
1. Data generation and refinement common tricks, GAN network, 2D feature to 3D feature conversion, feature alignment, multimodal input, improving loss function to obtain high-dimensional features.
2. Feature extraction and feature training, roughly divided into representation learning and metric learning.
3. Domain adaptation, common tricks, Leveraging pseudo-label clustering and prior knowledge.
4. Model efficiency, common tricks, Including front-end compression, back-end compression, Using auto-ml to find efficient networks.



ID	Title	Start Time	End Time
1	Data generated by 3d game e...	05/15/2024	05/22/2024
2	Data generated by GAN	05/22/2024	05/30/2024
3	Dealing with occlusion problems	05/30/2024	06/10/2024
4	Introducing multimodal input	06/10/2024	06/20/2024
5	Feature extraction and learning	06/21/2024	06/27/2024
6	Domain adaptation	06/28/2024	06/30/2024
7	Improve model efficiency	07/01/2024	07/07/2024

References

[1] C. Cho, W. J. Kim, S. Hong, et al., ”Part-based Pseudo Label Refinement for Unsupervised Person Re-Identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7308-7318.

[2] C. Wu, W. Ge, A. Wu, et al., ”Camera-Conditioned Stable Feature Generation for Isolated Camera Supervised Person Re-Identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20238-20248.

[3] Fu D, Chen D, Yang H, et al. Large-scale pre-training for person re-identification with noisy labels[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 2476-2486.

[4] D. Fu, D. Chen, H. Yang, et al., ”Large-Scale Pre-Training for Person Re-Identification with Noisy Labels,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2476-2486.

[5] X. Zhang, D. Li, Z. Wang, et al., ”Implicit Sample Extension for Unsupervised Person Re-Identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7369.

[6] S. Liao and L. Shao, ”Graph Sampling Based Deep Metric Learning for Generalizable Person Re-Identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7359-7368.

[7] H. Wang, J. Shen, Y. Liu, et al., ”Nformer: Robust Person Re-Identification with Neighbor Transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7297-7307.