DOI: 10.23977/jaip.2023.060204
ISSN 2371-8412 Vol. 6 Num. 2

Improved Method for Pedestrian Recognition Based on Generative Adversarial Networks

Lu Jianheng

School of Data Science, Guangzhou Huashang College, Guangzhou, China

Keywords: Re-Identification; Unlabeled Sample; Semi-supervised; Generative Adversarial Networks

Abstract: Traditional supervisory person re-id technology learning methods mainly relies on pre-marked image data, but there are a lot of unlabeled data in actual security scene, which seriously limits the application of person re-id technology in security monitoring field. Therefore, it is very important to study the semi-supervised learning of unlabeled data generated by antagonistic network. In the process of using GAN to generate data, in this paper we use global and local information of pedestrian image to generate realistic pedestrian image conditionally, and trains robust feature representations for different intra-class changes of cameras, so as to improve the accuracy of person re-id. The experimental results show that this method is more effective than the benchmark method. The performance of dataset Market1501 and Duke MTMC-reID improved by 4% and 3% respectively.

1. Introduction

Pedestrian re-identification is a technique to determine whether two images from two non-overlapping coverage cameras are the same pedestrian [1], which is widely used in security surveillance, but it remains challenging due to the significant variation of the same pedestrian within different cameras, mainly indicating changes in clothing, large background differences, and severe occlusion leading to low recognition rate of pedestrian re-identification. As pedestrian re-recognition is mainly affected by factors such as occlusion, viewpoint, pose and background, the same pedestrian varies significantly from camera to camera, leading to false positives in pedestrian re-recognition. Therefore, designing or learning feature representations that are as robust as possible to intra-class variations is one of the main goals of pedestrian re-recognition [2-6].

The mainstream methods for pedestrian re-recognition mainly consist of feature extraction [2,3,4,5,6] (learning features that can cope with changes in pedestrians under different cameras) and metric learning [7,8,9,10,11] (mapping the learned features to a new space so that the same people are closer and different people are further away), represented early on by the HOG feature and KISSME metric algorithms [12]. By obtaining a new metric space to calculate the similarity between pedestrians to distinguish different pedestrians. In recent years, mainly deep learning methods have been used, and convolutional neural networks (CNNs) have recently become an important choice for feature extraction due to their strong expressive power and learning invariant depth embedding. To solve the occlusion problem in pedestrian reidentification, Liu et al [13] proposed a quality evaluation network, which assigns weights based on the quality of each image and eventually integrates pedestrian expressions, which can In order to further reduce the impact from intra-class variation, many existing methods use part-based matching or the whole to explicitly align and compensate for variation. Wei et al [14] divided the human body into head, upper limb and lower limb parts through human keypoint localisation and extracted features separately from the whole features. The accuracy of recognition was further improved by fusing the features with the overall features. To address the problem of too few training samples, Liang et al [15] quickly generated more training data for the adversarial generative network DCGAN by generating data and expanding the technique to assign class labels to outlier unlabeled data, and finally verified through design experiments that the newly generated data did improve recognition accuracy. For the pedestrian re-recognition problem with different poses, Hai et al [16] used a GAN to distill the features of pedestrians by removing redundant information, learning only the features related to identity information, removing the redundant feature information such as human pose, and using only the robust features extracted by the encoder without adding additional computation when making inferences. To address the various low resolution and scale mismatch problems in pedestrian re-recognition, Wang et al [17] proposed a cascaded super-resolution GAN (CSR-GAN) framework by enhancing the resolution of pedestrian images by cascading multiple SRGANs in series to improve scale adaptive scaling, followed by the insertion of a re-recognition network to supplement the image feature representation. With the development of Generative Adversarial Networks (GANs), generative models are enhanced by introducing additional data, mainly considering the quality of the generated images and the diversity of the generated images, to bridge the gap between synthetic and real scene data while adequately covering the unseen inner class variations.

The generation pipeline in existing methods is kept relatively separate from the discriminative re-learning phase. As a result, pedestrian re-identification models are usually trained on the generated data in a straightforward manner. In this paper, we seek to improve existing methods by making better use of the generated data, by proposing a new discriminative loss function based on a combination of global and local information about the pedestrian, combined with a generative adversarial network capable of conditionally generating pedestrian images. The proposed method in this paper generates better quality images compared to other typical pedestrian re-identification GAN generation methods, not only reducing the process. The proposed method not only reduces the computational effort, but also reduces the discriminative complexity.

2. Related work

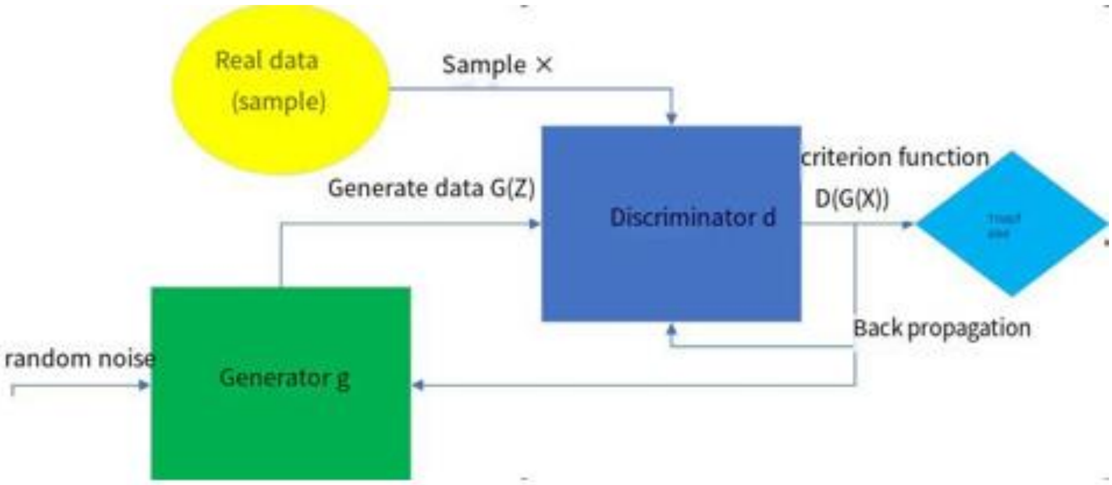

Figure 1: Schematic diagram of GAN network structure

Generative adversarial network, a probabilistic generative model in deep learning, was first proposed by Ian Goodfellow et al [18] in 2014, which consists of two mutual adversarial modules, a generator G, whose role is to capture the data distribution features, and a discriminator D, whose role is to estimate whether the images come from the training data or from the images generated by the generator G. The basic structure of the GAN network is shown in Figure 1. The corresponding data G (Z) is generated using generator G, and then the real sample data and the generated sample data are fed into discriminator D for true/false judgement.

When a generative adversarial network transforms an image, during training, the goal of the generative network G is to generate as many real images as possible to deceive the discriminative network D. The goal of D is to separate the images generated by G from the real images as much as possible, which corresponds to the maximum-minimum game, and finally to achieve dynamic equilibrium [18]. The following is the expression for the max-min game of the original generative adversarial network, which also represents the loss function of the generative adversarial network.

$$
\begin{aligned}
\min(\mathrm{G})\max(\mathrm{D})\mathrm{V}(\mathrm{D},\mathrm{G}) = \mathrm{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{data?}}(\mathrm{x})}[\log \mathrm{D}(\mathrm{x})] \\
+\mathrm{E}_{\mathrm{z}\sim \mathrm{p}_{\mathrm{z}}(\mathrm{z})}[\log(1-D(\mathrm{G}(\mathrm{z})))]
\end{aligned}
\tag{1}
$$

The right-hand two terms of Equation (1) denote the entropy of the discriminator and the entropy of the generator, respectively; Pdata, Pz denote the distribution of real data and the distribution of random noise, respectively; the generator minimizes V and the discriminator maximizes V. Both are jointly tuned. The original GAN does not require both G and D to be neural networks, as long as they satisfy the ability to fit the corresponding generating and discriminating functions. In practice, however, both G and D generally use deep neural networks. For pedestrian re-identification, G is a generative network and D is a discriminative network.

The current dataset for pedestrian re-identification requires baseline mapping and identity calibration, which is costly in terms of labour and time. With GAN, a larger number of pedestrian re-identification datasets can be generated quickly, and then the GAN-generated data can be processed and fused into the original training data, which can greatly increase the diversity of the original training dataset. Given the good results shown by CNN [19] in image processing, in order to improve the solution to the instability of GAN network training, Suarez et al [20] combined the CNN model to improve GAN and proposed DCGAN, which further improved the visual quality of the generated images.

There are two main innovations in this paper.

(1) A joint learning framework is proposed which combines re-learning and data generation end-to-end. The model involves a generation module that encodes each person as an appearance code and a structure code respectively, and a discriminator module that shares local features with the generation module, which conditionally generates pedestrian images based on the combination of global and local information about pedestrians in combination with generative adversarial networks, resulting in more realistic pedestrian images.

(2) A new discriminant loss function is proposed to improve the accuracy of pedestrian re-identification based on the information of the generated pedestrian images in the pedestrian retrieval process.

3. Improved generative adversarial network model

The core algorithm flow in this paper is shown in Figure 2, where our discriminative learning module D is embedded in the generation module G via a shared local feature encoder. The blue line indicates the extraction of local and global features of the pedestrian image, and the red line indicates the online feedback of the generated image to. Feature learning and fine-grained feature mining are primarily utilised in the generation module G in order to make better use of the generated data. The improved generative adversarial network model combines fine-grained feature information to conditionally generate high-quality images, and the discriminator module D involves primary feature learning and fine-grained feature mining, which are used against the generative module G to make better use of the generated data.

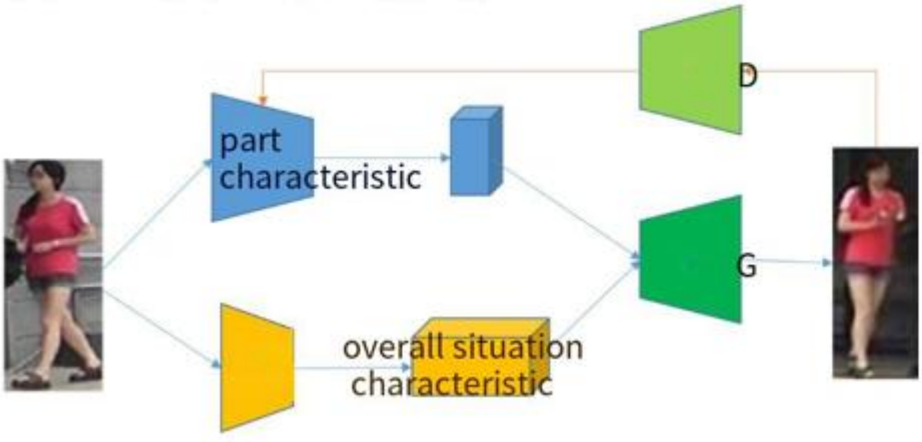

Figure 2: Flowchart of the algorithm in this paper

3.1 Generate modules

The real images and identity labels are denoted as $X = \{x_i\}_{i=1}^{N}$ and $Y = \{y_i\}_{i=1}^{N}$, where N is the number of images and $y_i \in [1, \mathrm{K}]$ and K denote the number of classes or identifiers in the dataset. Given two real images x_i and x_j in the training set, our generation module generates new pedestrian images from the local features and global features of the images. For the generated images, we use superscripts to denote the real images that provide the local feature codes and subscripts to indicate one that provides the global codes, while the real images only have subscripts as image indexes. Compared to the local features a_i; the global features s_j; maintain more spatial resolution to preserve geometric and positional properties. In practice, we convert the input image of E_s to greyscale to drive G in order to exploit a_i and s_j. We enforce the two goals of the generation module:

(1) self-identity generation to normalize the generator;

(2) cross-identity generation to make the generated images controllable and match the actual data distribution. We use the channel ×height ×width to indicate the size of the element map (i) E_a based on ResNet50 pre-trained on ImageNet by removing its global average pool and fully connected layers, and then attaching an adaptive maximum pool layer to output a local encoding in 2048 ×4 ×1. It maps to the global feature f_{prim} and the local feature f_{fine} via two fully connected layers, both of which are 512-dimensional vectors, and we use discriminators on three different input image scales: 64 ×32, 128 ×64 and 256 ×128. When updating D for stable training, we also adapt all input images to 256 ×128.

$$\begin{aligned} L_{\mathrm{recon}}^{\mathrm{code}_1} &= E\left[a_i - E_a\left(G\left(a_i, s_j\right)\right)_1 \right] \\ L_{\mathrm{recon}}^{\mathrm{code}_2} &= E\left[\mathrm{s}_j - E_s\left(G\left(a_i, s_j\right)\right)_1 \right] \end{aligned} \tag{2}$$

With the generation mechanism, we enable the generation module to learn appearance and global features in an explicit and complementary sense, and generate high-quality pedestrian images based on latent codes, which largely reduces the complexity of generation.

3.2 Confrontation model

The discriminator module in this paper is embedded in the generation module by sharing local features as a backbone. By switching between images generated by local or global features, we propose primary feature learning and fine-grained feature mining to make better use of the images generated online. Also after fine-grained feature mining, in addition to using the generated data directly to learn the main features, an interesting alternative implemented through our specific generation process is to simulate clothing variations of the same person. When training images organised in this way, the discrimination module is forced to learn fine-grained id-related attributes not related to clothing (e.g. hair, hat, bag, body size, etc.). We combine an image generated from a global feature with different local features as the same class as the real image that provides the global feature.

High quality synthetic images can be considered essentially 'internal' (as opposed to 'outlier'), as the generated image maintains and recombines the visual content from the real data. With the two feature learning tasks described above, the discriminator module can use the generated data in a way that is specific to the way we manipulate local and global features. Rather than using a single supervision as in almost all previous approaches, we process the generated images from two different perspectives through primary feature learning and fine-grained feature mining, and we jointly train local and global encoders, decoders and discriminators to optimise the overall objective.

4. Experimental results and analysis

4.1 Evaluation metrics

Qualitative and quantitative comparisons of conditional-based generative adversarial network generation and discrimination results were performed on a benchmark dataset. The m-AP metric is a measure of the probability that the correct result of the detection algorithm occurs first, and is usually obtained by taking the mean of the AP of multiple queries (mean) to represent the accuracy of the query system, and the Rank1 metric is a measure of the probability that the top graph in the search results is the correct result, and is usually obtained by experimenting with multiple It is generally obtained by averaging the results over several trials.

4.2 Model Experiment Results

A comparison of the generated and real images on Market-1501 with the FD-GAN method on GAN generated images shows from the experimental results that the PN-GAN method, our method generates better results and retains more texture details.

Table 1: Results of pedestrian re-identification into migration tests on a publicly available dataset

	Market1501—DukeMTMC-reID	DukeMTMC-reID -market1501
Rank@1	42.62%	56.12%
Rank@5	58.57%	72.18%
Rank@10	64.63%	78.12%
mAP	24.25%	26.83%

To verify the generalisability of our method, we trained the model on dataset A and tested the model directly on dataset B (without adaptation). We denote direct transfer learning as A → B. Table 1 represents the migration test results of our method on pedestrian re-identification into migration on the public dataset, trained on Market1501 and tested on DukeMTMC-reID, rank1 is 42.62% and mAP is 24.25%, trained on Market1501 for testing, rank1 was 56.12% and mAP was 26.83%, as shown table 1.

4.3 Comparison of results and evaluation

Extensive experiments show that our proposed improved pedestrian re-identification method based on generative learning and joint discrimination can generate more realistic and diverse pedestrian images, while consistently outperforming the latest available algorithms for maximum accuracy in all benchmark tests.

The experimental results from Table 2 show that our method achieves the best performance. The improved pedestrian re-identification method based on generative learning and joint discrimination achieves significant gains of 8.3% and 10.3% for mAP [21] on Market-1501 and DukeMTMC-reID compared to the method using separately generated images, demonstrating the advantages of the proposed method. On the recently released large-scale dataset MSMT17 [22], our method significantly outperformed the FN-GAN approach with 9.0% Rank 1 and 11.9% mAP.

Table 2: Pedestrian re-identification results on publicly available datasets

Methods	Market1501 Rank1 mAP	DukeMTMC-reID Rank1 mAP	MSMT1 Rank1 mAP
Baseline	89.6 74.5	82.0 65.3	68.8 36.2
FD-GAN	94.0 84.4	85.6 72.7	76.0 49.7
Ours	**94.3 85.2**	**86.3 73.8**	**76.6 51.5**

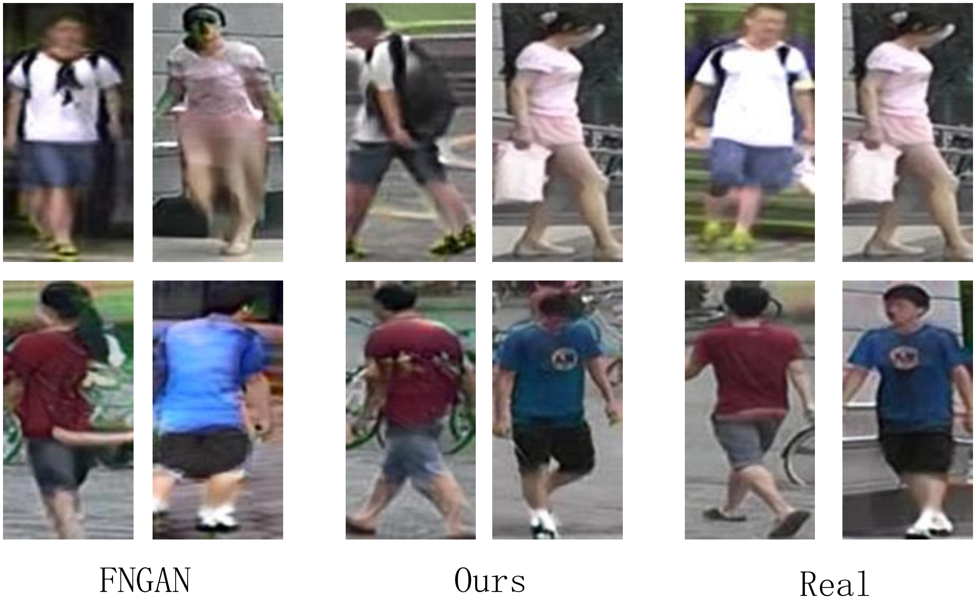

Figure 3: Generating results on the Market-1501 dataset

In the qualitative evaluation, we can see from Figure 3: Generating results on the Market-1501 dataset and Table 3 that our joint discriminant learning is beneficial for image generation. Pose conditioned PN-GAN produces relatively good visual results but still contains visible blur and artifacts, especially in the background. In contrast, we generate images that are more realistic and close to true in both foreground and background. In the quantitative evaluation, our qualitative observations above are confirmed by the quantitative evaluation. We applied SSIM to calculate intra-class similarity, compared to SSIM [23] (the higher the better) and FID [24] (the lower the better) to measure how close the distribution of the generated images is to that of the real images. It is sensitive to visual artefacts and therefore reflects the realism of the generated image. FID is more robust to noise, but still does not address the problem of overfitting on large-scale datasets, and feature extraction-based methods can only evaluate features based on their presence or absence, not their relative spatial location. Fid has better evaluation performance, but has the same drawbacks as SSIM, such as being unsuitable for use on internally varying large datasets and the inability to distinguish overfitting. As shown in Table 3, our method significantly outperforms other methods in terms of realism and diversity, indicating the high quality of the images we generate. Notably, we obtained a higher SSIM than the original training set due to the various poses, backgrounds, etc. introduced by the global features used to indicate its availability to reflect generative diversity [25].

Table 3: Pedestrian re-identification results on publicly available datasets

Mthods	Realism	Diversity
	(FID)	(SSIM)
Real	7.22	0.35
PN-GAN	257.00	0.247
Ours	**16.35**	**0.38**

5. Conclusion

In this paper we present a framework for an improved approach to pedestrian re-identification based on generative learning and joint discriminations, end-to-end, to re-learn and create images in a unified network. An online interaction loop exists between the discriminative model and the generative module, and a new discriminative loss function is proposed by combining global and local information about pedestrians in order to mutually support the two tasks. Our two modules are jointly designed to allow re-learning to make better use of the generated data, rather than simply training them, with the aim of improving pedestrian re-identification accuracy. Experiments on three benchmark tests show that the method proposed in this paper brings substantial improvements to the image generation quality and pedestrian re-recognition accuracy improvement compared to other typical pedestrian re-recognition GAN generation methods, generating better quality images while reducing the process computation and discrimination complexity.

References

[1] Bedagkar-Gala A, Shah S K. A survey of approaches and trends in person re-identification [J]. Image & Vision Computing, 2014, 32(4):270-286.

[2] Gray D, Tao H. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features[C]//Computer Vision - ECCV 2008, European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings. DBLP, 2008:262-275.

[3] Zheng W S, Li X, Xiang T, et al. Partial Person Re-Identification[C]//IEEE International Conference on Computer Vision. IEEE, 2016: 4678-4686.

[4] Zhao R, Ouyang W, Wang X. Unsupervised Salience Learning for Person Re-identification[C]//Computer Vision and Pattern Recognition. IEEE, 2013:3586-3593.

[5] Zhang L, Xiang T, Gong S. Learning a Discriminative Null Space for Person Re-identification[C]//Computer Vision and Pattern Recognition. IEEE, 2016:1239-1248.

[6] Van d W J, Schmid C, Verbeek J, et al. Learning Color Names for Real-World Applications[J]. IEEE Transactions on Image Processing, 2009, 18(7):1512-1523.

[7] Liu Y. Distance Metric Learning: A Comprehensive Survey [D]. Michigan State Universiy, 2006.

[8] Martin Köstinger, Hirzer M, Wohlhart P, et al. Large Scale Metric Learning from Equivalence Constraints[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012.

[9] Davis J V, Kulis B, Jain P, et al. Information-theoretic metric learning[C]// Icml 07: International Conference on Machine Learning. 2007.

[10] Hirzer M, Roth P M, Martin Ko ̈stinger, et al. Relaxed Pairwise Learned Metric for Person Re-identification[C]//

European Conference on Computer Vision (ECCV). Springer, Berlin, Heidelberg, 2012.

[11] Chen D, Yuan Z, Hua G, et al. Similarity learning on an explicit polynomial kernel feature map for person re-identification[C]// Conference on Computer Vision & Pattern Recognition. 2015.

[12] Martin Köstinger, Hirzer M, Wohlhart P, et al. Large Scale Metric Learning from Equivalence Constraints[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012.

[13] Liu Y, Yan J, Ouyang W. Quality Aware Network for Set to Set Recognition [J]. 2017.

[14] Wei L, Zhang S, Gao W, et al. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification [J]. 2017.

[15] Z. Zheng, L. Zheng, Y. Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]//IEEE International Conference on Computer Vision, 2017: 3774-3782

[16] Ge Y, Haiyu Zhao. FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification [J]. 2018.

[17] Wang Z, Ye M, Yang F, et al. Cascaded SR-GAN for Scale-Adaptive Low Resolution Person Re-identification [C] // IJCAI. 2018: 3891-3897.

[18] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks [J]. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.

[19] Shin H C, Roth H R, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning [J]. IEEE transactions on medical imaging, 2016, 35(5): 1285-1298.

[20] Suarez P L, Sappa A D, Vintimilla B X. Infrared Image Colorization Based on a Triplet DCGAN Architecture[C]// Computer Vision & Pattern Recognition Workshops. 2017.

[21] Chen B, Cha YF, Li YQ, et al. Pedestrian re-identification algorithm for translational variability similarity learning [J]. Journal of Electronics and Information, 2018, 40(10):100-106.

[22] Xu Yang. Research on pedestrian re-identification algorithm based on convolutional neural network [D]. East China Normal University, 2018.

[23] Liu Xiaokai. Research on pedestrian re-identification method in intelligent surveillance system [D]. Dalian University of Technology, 2017.

[24] Yao Wanchao. Pedestrian re-identification algorithm based on convolutional neural network [D]. Zhejiang University, 2017.

[25] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1116-1124.