Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/133/2025.20602

Clothes-changing Person Re-identification Based on Spatial Consistency

Mai Zhang[1,a,*]

[1]*Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), China*
a. zhangmai778@163.com
**corresponding author*

Abstract: Address clothes-changing person recognition of other key is to extract the characteristics associated with human nature, for example, face, hair, body size and gait. At present, most of the research work mainly focuses on processing multi-modal information and fails to make full use of the information related to human nature in the original RGB images. In this paper, a viewpoint-based adversarial loss algorithm is proposed to mine visually relevant features from the original RGB images by punishing the predictive ability of the ReID model. A lot of experiments show that our model has achieved good results.

Keywords: clothes changing, viewpoint, adversarial.

1. Introduction

The purpose of person recognition, especially in public safety and other critical areas, is to be able to accurately identify the same individual across different cameras and viewing angles. However, criminals often disguise their identity by changing clothing, which can greatly affect the effectiveness of traditional person identification methods. To mitigate the impact of clothing changes, some person recognition technologies utilize multimodal information or dual-flow frames to capture features other than appearance, such as body shape. However, these approaches also face challenges. For example, body shapes extracted through a body analysis network can be influenced by clothing sizes, introducing unwanted noise. In addition, integrating multimodal information to learn features that are not affected by clothing often leads to information redundancy and increased computational complexity. Existing methods often ignore the wealth of non-clothing and non-viewpoint-related information in raw RGB images. Despite the use of robust backbone networks to extract features from these images, the design of loss functions is often inadequate, resulting in feature maps focusing only on basic and non-discriminative information.

In order to mine non-clothing and non-viewing-angle related information in RGB domain more effectively, this paper introduces a viewing-angle based adversarial loss (VAL) algorithm on the basis of clothing based adversarial loss (CAL)[1] algorithm. Our approach involves developing an Angle of view calculation module that sits before the backbone network of the re-id model for calculating person Angle information. After the backbone network, a perspective classifier is added and VAL is defined as a polypositive class classification loss, where all perspective classes belonging to the same identity are treated as positive classes of each other.

DOI: 10.54254/2755-2721/133/2025.20602

By combining VAL with CAL, we force the backbone network of the re-id model to learn features independent of clothing and perspective. Through backpropagation, this combined loss function enhances the model's ability to highlight features independent of clothing and viewing angles, even if the same individual is wearing different clothing and presented from different viewing angles. A large number of experiments show that the performance of person recognition system is significantly improved by relying only on RGB images and using VAL.

2. Related Work

2.1. Person Re-ID

General person re-recognition aims to address challenges such as changes in person posture, lighting conditions, and occlusion, often assuming that an individual's clothing remains consistent over a brief period. Existing approaches employ contrastive learning strategies and region-level triplet loss to extract distinctive features associated with individual identities. Furthermore, interaction and aggregation Re-ID frameworks are utilized to model the interdependencies among spatial features and subsequently aggregate related features corresponding to the same body parts, thereby enhancing robustness against variations in posture and scale. Additionally, a block-based feature extraction strategy divides person images into multiple vertical segments, calculating classification losses for each segment individually to achieve fine-grained feature matching. These methods exhibit remarkable performance in person re-recognition tasks, primarily relying on the consistent appearance information provided by an individual's clothing. However, when persons change their attire, this reliance on appearance information becomes unreliable, resulting in a substantial decline in the effectiveness of these approaches.

2.2. Cloth-changing Person Re-ID

As the field of person re-recognition with clothing changes garnered increasing interest among researchers, several datasets were released, including PRCC, VC-Clothes, DeepChange, and LaST. These datasets featured the same persons in two or more different outfits and accessories, such as glasses, hats, and backpacks, significantly diversifying appearances and posing greater challenges. To tackle the CC-ReID (Clothing Change Re-Identification) task, some studies utilized auxiliary identity cues to extract appearance and body shape features from original images and body pose heatmaps, subsequently fusing them through cross-attention mechanisms. Additionally, measure-based learning methods demonstrated notable advancements in CC-ReID. Distinct from existing research, our proposed approach emphasizes the extraction of high-level identity-related semantic features. It directs the model to focus on semantically consistent regions that are independent of clothing and perspective, thereby more effectively addressing the challenge of re-recognizing persons with changed attire.

2.3. Alpha-Pose

Alpha-Pose is a powerful and efficient real-time multi-person pose estimation system based on a two-stage approach. The first stage is object detection, which first uses a target detector (such as YOLO, Faster R-CNN or Mask R-CNN) to detect the human body region in the input image and generate a human frame. The second stage is pose estimation. For each detected human body area, Alpha-Pose uses pose estimation forget (such as ResNet, HRNet) to detect key points of the human body (such as head, shoulders, elbows, knees, etc.) to build a human skeleton map.

In addition, Alpha-Pose employs several techniques to improve accuracy and robustness: (1) Multi-scale reasoning: By running the network at different resolutions, the detection accuracy of

DOI: 10.54254/2755-2721/133/2025.20602

different sized targets is improved. (2) Online data enhancement: dynamically adjust images in the training process to enhance the generalization ability of the model. (3) Pose Flow: An online pose tracker that correlates all poses of the same person to achieve more continuous and accurate pose estimation.

Overall, Alpha-Pose is a leading human posture recognition tool with high accuracy, robustness, and ease of use. It has wide application prospects in many fields and is an important technology in the field of computer vision and artificial intelligence.

3. Methodology

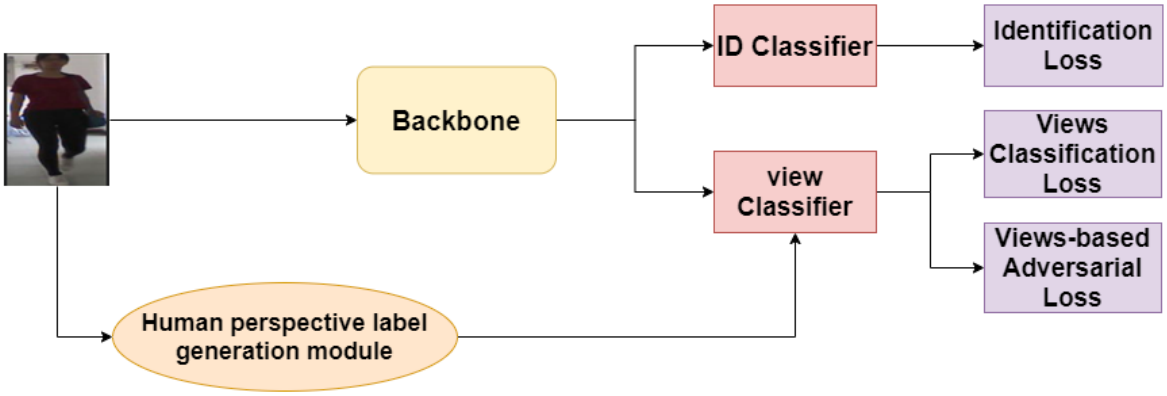

Figure 1: Overall architecture

In this section, we will introduce in detail the proposed human perspective label generation module and viewpoint-based antagonistic loss. An overview of our approach is shown in Figure 1.

3.1. Human perspective label generation module

The human perspective label generation module is a method to automatically assign perspective labels to person images by processing and analyzing the coordinates of key points of human bones. Firstly, advanced pose estimation algorithm (such as Alpha-Pose, etc.) is used to obtain the precise coordinates of the key points of human bones in the original person image, including the nose, left eye, right eye, left ear, right ear, left foot and right foot.

After obtaining the coordinates of these key points, the module further calculates the average of the coordinate points of the nose, left eye, right eye, left ear, and right ear to determine the center point of the body. This central point represents the approximate position of the human body in the image and is the basis for subsequent calculations. At the same time, the module also calculates the difference between the coordinate points of the left foot and the right foot, and obtains the vector between the feet, which reflects the posture and orientation of the human body in the image. Next, the module uses a series of mathematical transformations to obtain a converted vector based on the vector between the center point of the body and the foot. This converted vector more accurately reflects the orientation of the human body relative to the image plane. The module then uses a specific function to calculate the radian Angle between the converted vector and the center point. This radian Angle is a key parameter that directly reflects the perspective of the human body in the image. Finally, based on the calculated radian Angle, the module maps it to a preset set of view labels to determine the final view label. This perspective label can be a discrete classification label (front, side, back, etc.) or a continuous Angle value, depending on the needs of the application and the complexity of the scene.

DOI: 10.54254/2755-2721/133/2025.20602

In general, the human perspective label generation module can automatically assign accurate perspective labels to person images through accurate coordinate processing and analysis of human bone key points, which provides strong support for subsequent image analysis, behavior recognition and other tasks.

3.2. Viewpoint-based Adversarial loss

Current Re-ID and gait recognition techniques do not fully leverage the view angle information available in the RGB channel. In this paper, motivated by CAL, we enhance the Re-ID model's backbone network to extract view-independent information by penalizing its predictive capabilities. To achieve this, we introduce a new perspective classifier $C_{\varphi}^{V}(\cdot)$ with parameters φ appended to the backbone network. Each training iteration comprises a two-step optimization process.

3.2.1. Train the view classifier

In the first step, we optimize the view classifier by minimizing the view classification loss L_V, which is the predicted cross-entropy loss between the view label $C_{\varphi}^{V}(g_{\theta}(x_i))$ and the view label y_i^V. The company responsible for this process can be represented as

$$\min_{\varphi} L_V(C_{\varphi}^{V}(g_{\theta}(x_i), y_i^V) \tag{1}$$

When we denote $g_{\theta}(x_i)$ after l_2-normalization as f_i and denote the weights of j-th view classifier after l_2-normalization as φ_j, L_V can be expressed as:

$$L_V = -\sum_{i=1}^{N} \log \frac{e^{\left(f_i \cdot \varphi_{y_i^V}/\tau\right)}}{\sum_{j=1}^{N_v} e^{(f_i \cdot \varphi_j/\tau)}} \tag{2}$$

where N is the batch size, N_V is the number of view classes in the training set, and $\tau \in R^+$ is a temperature parameter.

3.2.2. Learning views-irrelevant features

In the second step, we freeze the parameters of the perspective classifier and encourage the backbone network to learn features that are not dependent on the perspective. To accomplish this, we need to penalize the predictive capability of the Re-ID model. Specifically, we introduce a multi-positive class classification loss L_{VA}, where all view classes associated with the same identity are considered positive classes relative to each other. For example, given a sample x_i, all views classes belonging to its identity class y_i^{ID} are defined as its positive views classes. Therefore, L_{VA} can be formulated as:

$$L_{VA} = -\sum_{i=1}^{N} \sum_{v=1}^{N_v} q(v) \log \frac{e^{(f_i \cdot \varphi_v/\tau)}}{e^{(f_i \cdot \varphi_v/\tau)} + \sum_{j \in S_i^-} e^{(f_i \cdot \varphi_j/\tau)}} \tag{3}$$

$$q(v) = \begin{cases} \frac{1}{K}, v \in S_i^+ \\ 0, v \in S_i^- \end{cases} \tag{4}$$

$S_i^+(S_i^-)$is a collection of clothing classes that have the same identity (different identity) as f_i. K is the number of classes in S_i^+ and $q(v)$ is the weight of the cross entropy loss of the v-th view class. Positive classes $(c = y_i^C)$ with the same clothes and positive classes $(c \neq y_i^C$和$c \in S_i^+)$ with different clothes have equal weight, i.e. $\frac{1}{K}$. The punishment mechanism makes the model pay more attention to features that are not related to views. The punishment mechanism means that a view

DOI: 10.54254/2755-2721/133/2025.20602

classifier is trained first, and then deprives the classifier of the ability to recognize views, thus shifting the model's attention to features that are not related to views. The loss function is used to extract the inherent features of human body efficiently. Person features are obtained by using common identity loss functions.

4. Experiments and discussion

4.1. Datasets

Our experiment used two datasets, both PRCC and LTCC, assembled using images taken from real surveillance cameras.

The PRCC includes 33,698 pictures from 221 people, from 3 different angles, and also provides outline sketch images of people to facilitate the extraction of people's outline information. The LTCC includes 17,138 images of 152 people, providing security camera images from 12 cameras in different lighting conditions.

4.2. Competitors

Re-ID task is a new and challenging topic, which has attracted the interest of researchers in related fields in the past 2-5 years. In our experiments, we used the latest and popular references as our competitors. Including HACNN(CVPR 2018), PCB(ECCV 2018), IANet(CVPR 2019), ISP(ECCV 2020), GI-ReID(CVPR 2022).

4.3. Implementation Details

We adopt ResNet-50 as the backbone for our Re-ID model, and to enhance feature granularity, we omit the final downsampling layer of ResNet-50. For image-based datasets such as LTCC and PRCC, we utilize both global average pooling and global max pooling to aggregate the feature maps output by the backbone. These pooled features are then concatenated and normalized using Batch Normalization. The input images are resized to 384*192 pixels, and we apply random horizontal flipping, random cropping, and random erasing for data augmentation. The batch size is set to 64, with each batch containing 8 individuals, each represented by 8 images. We train the model using Adam optimizer for 60 epochs, introducing a specific technique L_{VA} after the 25th epoch. The initial learning rate is set to a predefined value $3.5e^{-4}$ and is reduced by a factor of 10 every 20 epochs. The parameter τ in Equation (3) is set to 1/16. Notably, we directly apply these optimal parameter settings to other datasets without further tuning.

4.4. Experimental result

We compare our approach to some advanced methods on the PRCC and LTCC datasets. The results are shown in Table 1. It can be seen that the effect of our method is significantly improved in the case of clothing changes, thanks to the fact that we apply perspective information to the model to obtain more robust features to the appearance changes.

Table 1: Performance comparison

	PRCC		LTCC	
	Rank-1	mAP	Rank-1	mAP
HACNN[2]	21.8	23.2	21.6	9.3

DOI: 10.54254/2755-2721/133/2025.20602

Table 1: (continued).

PCB[3]	41.8	38.7	23.5	10.0
IANet[4]	46.3	45.9	25.0	12.6
ISP[5]	36.6	-	27.8	11.9
GI-ReID[6]	-	-	23.7	10.4
Ours	**52**	**51.7**	**28.9**	**13.7**

5. Conclusion

In this article, we suggest using perspective information for CC-ReID tasks. A framework containing a perspective label generation module and a perspective-based approach to counter learning loss is proposed. The framework improves robustness to clothing changes by suppressing the model's sensitivity to perspective information. Our experimental results on two public CC-ReID datasets show that the algorithm has good performance.

References

[1] *Gu X, Chang H, Ma B, et al. Clothes-changing person re-identification with rgb modality only[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1060-1069.*

[2] *Li, Wei, Xiatian Zhu, and Shaogang Gong. "Harmonious attention network for person re-identification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.*

[3] *Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 480-496.*

[4] *Hou R, Ma B, Chang H, et al. Interaction-and-aggregation network for person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9317-9326.*

[5] *Zhu K, Guo H, Liu Z, et al. Identity-guided human semantic parsing for person re-identification[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer International Publishing, 2020: 346-363.*

[6] *Jin X, He T, Zheng K, et al. Cloth-changing person re-identification from a single image with gait prediction and regularization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 14278-14287.*