International Journal of Web Services Research
Volume 20 • Issue 1

Dual-Branch Network Fused With Two-Level Attention Mechanism for Clothes-Changing Person Re-Identification

Yong Lu, Minzu University of China, China*
Ming Zhe Jin, Minzu University of China, China

ABSTRACT

Clothes-changing person re-identification is a hot topic in the current academic circles. Most of the current methods assume that the clothes of a person will not change in a short period of time, but they are not applicable when people change clothes. Based on this situation, this paper proposes a dual-branch network for clothes-changing person re-identification that integrates a two-level attention mechanism and captures and aggregates fine-grained person semantic information in channels and spaces through a two-level attention mechanism and suppresses the sensitivity of the network to clothing features by training the clothing classification branch. The method does not use auxiliary means such as human skeletons, and the complexity of the model is greatly reduced compared with most methods. This paper conducts experiments on the popular clothes-changing person re-identification dataset PRCC and a very large-scale cross-spatial-temporal dataset (LaST). The experimental results show that the method in this paper is more advanced than the existing methods.

KEYWORDS

Channels, Clothes Features, Clothes-Changing Person Re-Identification, Complexity, Dual-Branch Network, Fine-Grained Person Semantic Information, Spaces, Two-Level Attention Mechanism

INTRODUCTION

Person re-identification technology, a key technology within intelligent surveillance systems, is regarded as an image retrieval problem. Person re-identification technology is a necessary technology for intelligent surveillance systems in public places for instances like locating criminals. It can also be applied to intelligent security, epidemiological investigations, and intelligent transportation. Through all-weather monitoring, the technology can prevent the occurrence of crimes like theft and robbery, locate lost persons, and assist intelligent transportation systems in completing the automatic dispatching of people, vehicles, and roads.

When monitoring large amounts of data, traditional manual processing methods are inefficient and costly. The person re-identification technology can improve such problems by quickly locating

DOI: 10.4018/IJWSR.322021 *Corresponding Author

and tracking the target. This saves labor costs, improves the accuracy of detection, and has a high application value in intelligent monitoring systems.

Person re-identification aims is to search for a targeted person via surveillance videos at different locations and times. Due to factors like the limitations of technology, most of the current research on person re-identification assume that the target's clothes are unchanged (Huang et al., 2018; Jin et al., 2022; Li et al., 2018). Thus, it uses the color, texture, and other features of the clothes as discriminant conditions. However, the problem of changing clothes is unavoidable when re-identifying a person over an extended time. There is also the problem of changing clothes in some short-term scenarios. For example, suspects usually change clothes to avoid identification and tracking. The original method will no longer be applicable in the clothes-changing scenario because people may be wrongly matched if wearing similar clothes. To address the issue, this article studies problems related to clothes-changing person re-identification.

To avoid the interference of clothes, some clothes-changing re-identification methods attach modal inputs along with the input image (Chao et al., 2019; Chen et al., 2021; Qian et al., 2020; Shu et al., 2021; Yang et al., 2019). These include three-dimensional (3D) shapes, bones, and contour (Chao et al., 2019; Chen et al., 2021; Qian et al., 2020). However, these methods often require additional models to capture multimodal information. This, in turn, increases the complexity of the model. In fact, original images contain rich clothing-independent information, which is largely underutilized.

This article aims to better mine information unrelated to clothes in the image. Thus, it adds a two-level attention module to the model, acting on the features extracted by the backbone network in space and channel, respectively. Then, it obtains a multi-scale fine-grained attention map. The module can more effectively capture the semantic information of persons in the channel and space, as well as eliminate the influence of irrelevant background as it focuses on features related to an individual. In view of the influence of the clothes feature, this article sets up a clothes classification branch. It also suppresses the sensitivity of the model to clothes features by training this branch. Experiments on popular datasets show that the proposed method is competitive (Shu et al., 2021; Yang et al., 2019).

The contributions of this article can be summarized as follows. First, the article uses IBN-Net as the backbone network. It uses the instance normalization (IN) layer to eliminate individual contrast and extracts features, such as person texture and outline, through the batch normalization (BN) layer. It uses a two-level attention module to capture and aggregate more fine-grained features in space and channels. Experiments show that this kind of two-level attention module enables the network to learn complementary global and local features. This is more suitable for clothes-changing person re-identification. Second, to be more suitable for the clothes-changing person re-identification, this article adds a clothes classification branch to suppress the sensitivity of the model to clothes features. Third, this article conducts extensive experiments on commonly used data sets and large-scale cross-temporal dataset. All have achieved strong results.

RELATED WORK

Person re-identification was first recognized as an independent task in the field of computer vision at the 2016 CVPR conference. The task of clothes-changing (long-term) person re-identification was first proposed in 2019. At present, many researchers have devoted themselves to the research of person re-identification algorithm technology without changing clothes. They have achieved remarkable results. However, there are few studies on clothes-changing person re-identification. The results obtained are not ideal.

Non-Clothes-Changing Person Re-Identification

Person re-identification has developed rapidly in recent years. The research on person re-identification is divided into traditional methods and deep learning methods. Traditional methods include two steps of feature extraction and similarity measurement (Liao et al., 2015; Weinberger & Saul, 2008). The

purpose of feature extraction is to extract discriminative and robust feature representation like color, HOG, and SIFT. The purpose of similarity measurement is to design the measurement function. Thus, the smaller the intra-class distance, the larger the inter-class distance.

Weinberger et al. (2008) proposed to improve the nearest neighbor classification by learning the Mahalanobis distance metric. Liao et al. (2015) proposed a method of subspace and metric learning to make a stable representation of the change of viewpoint and obtain an effective feature representation. However, traditional methods have limited learning ability and are difficult to adapt to large data volume tasks.

With the development of deep learning, current work is based on neural networks to learn discriminative features (Gong et al., 2021; He et al., 2021; Li et al., 2018; Xu et al., 2018). For example, Gong et al. (2021) proposed a strategy to eliminate bias with bias. It balances the weight between color features and color-independent features in the neural network by discarding part of the color information in the training data. It, thereby, overcomes the influence of color segmentation. He et al. (2021) used transformer for person re-identification and, for the first time, proposed to use jigsaw patch module (JPM) to generate robust features with improved discriminative ability and more diverse coverage. Their work introduced a side information embeddings (SIE) that incorporate non-visual cues by inserting learnable embeddings to alleviate feature bias to camera viewpoint changes. Li et al. (2018) built a harmonious attention CNN (HA-CNN) model for joint learning of soft pixel attention and difficult area attention. It also optimized feature representation, which can solve the misalignment of the input image. Xu et al. (2018) added pose-guided part attention (PPA) and attention-aware feature composition (AFC) to the network. They aimed to learn the attention of rigid and non-rigid parts using pose information. Then, global features and partial features are used as the final feature embedding. Still, these works focused on short-term re-identification based on color appearance features. These do not perform well in the case of persons changing clothes.

Clothes-Changing Person Re-Identification

There are few studies on clothes-changing person re-identification that distinguish persons based on clothes-independent features like body shape, face, or 3D shapes. Wan et al. (2020) focused on detecting facial information and extracting facial features to improve the accuracy of the model. Chen et al. (2021) extracted texture-insensitive 3D shape embeddings directly from 2D images by adding 3D body reconstruction as an aid. However, most of these methods use auxiliary means, such as human skeleton. Additional masks, pose, or contour estimation increase the computational cost of these methods. At that same time, it increases the complexity of the model.

This article proposes an unassisted network used on RGB images, which learns the fine-grained features of the model through the attention mechanism. It also uses the clothes classification branch to suppress the clothes features extracted by the network. The method is competitive with existing methods.

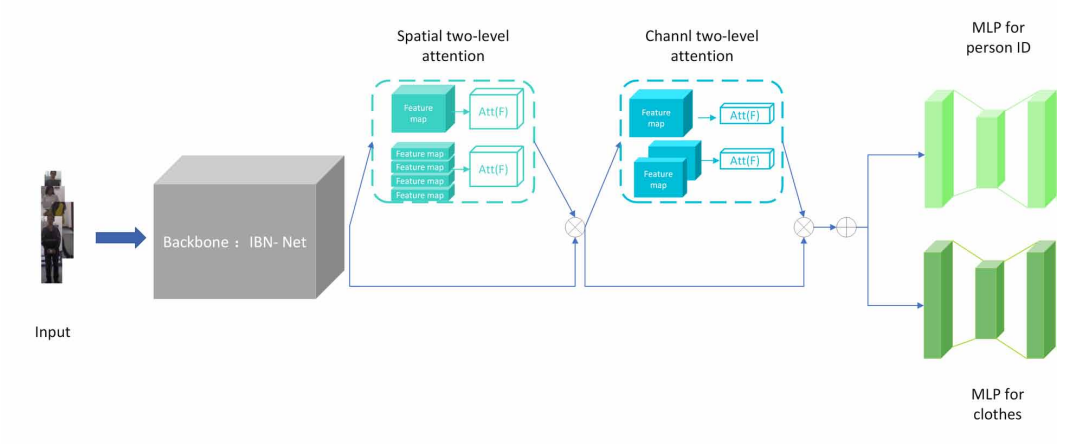
METHOD

Framework

Figure 1 illustrates the framework of the method. The backbone network uses IBN-Net with the last two layers removed. It also adds a spatial two-level attention module and a channel two-level attention module. The connection method of the two uses a concatenated structure. Finally, it uses multi-layer perceptron (MLP) to classify the extracted person features. A clothes classification branch is added outside the person classification branch to make the model more suitable for the scene of changing clothes.

This article uses the two-level attention module to further capture the semantic information of persons, suppress the influence of background information, and extract more fine-grained features inherent in persons. The details will be discussed in the two-level attention module. In addition to the

Figure 1.
Framework of the method in this paper



person classification branch, this article adds a clothes classification branch. The parameters of the adversarial loss are obtained by classifying the clothes. In the process of minimizing the adversarial loss, the model is forced to learn non-clothing features.

Backbone

Most of the current popular network structures use Resnet-50 as the backbone network. The BN in the network can make the overall sample features have different points for each sample. This is very effective for distinguishing different sample types. However, in the clothes-changing person re-identification, the appearance of similar samples change greatly. The use of the BN layer has a negative impact on the recognition accuracy of the network. Moreover, due to changes in the viewing angles of cameras, the captured images of persons can have undesirable factors like blur, occlusion, posture, color, style, and brightness. This is particularly true in the scene of changing clothes, where a person's appearance changes more often.

To solve these problems, this article uses IBN-Net (Pan et al., 2018) as the backbone. It combines the IN layer and BN layer in the network to complement the limitations of the two. The IN layer features images that do not change due to changes in appearance (i.e., style, color, and brightness). The BN layer features include texture and outline.

Two-Level Attention Module

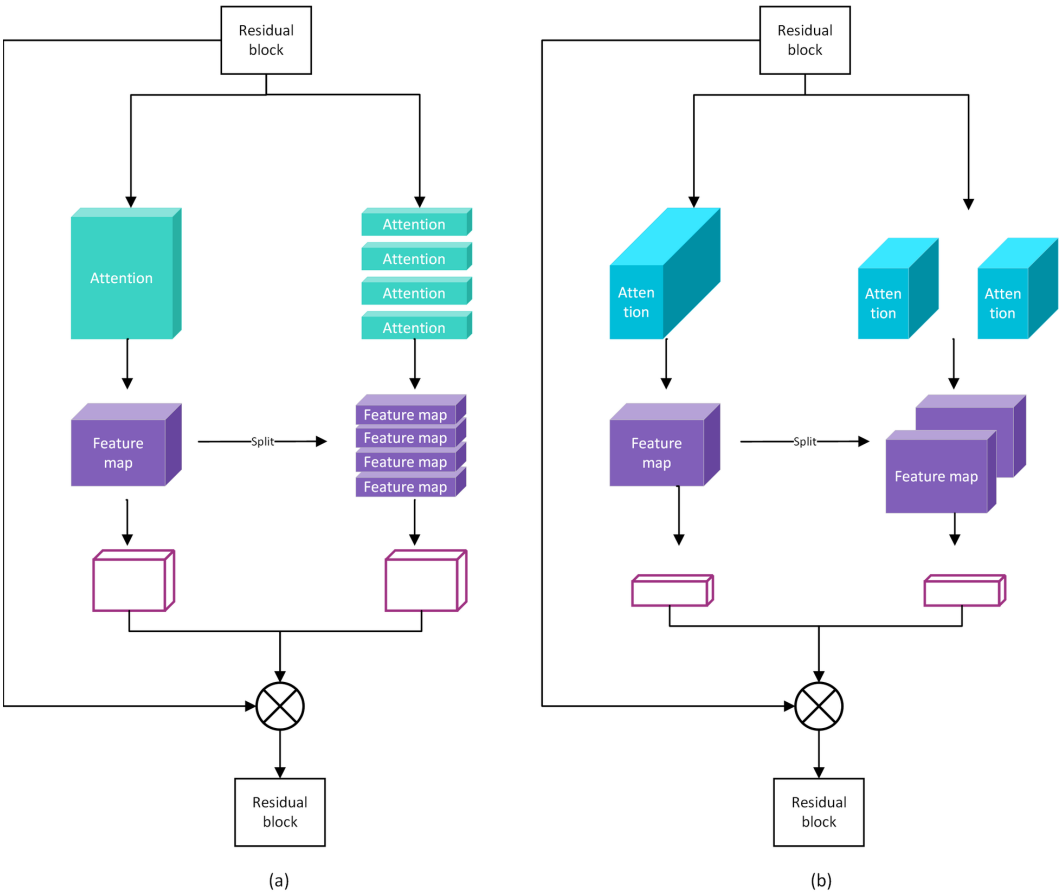
The focus of clothes-changing person re-identification is to extract features related to people. However, background information is often captured along with the person's image. This, in turn, interferes with the re-identification task. In clothes-changing person re-identification, the clothes feature is no longer used to distinguish persons; thus, there are fewer features that can be used.

The attention mechanism has been proved to be effective for extracting key information in images. However, in the clothes-changing person identification, it is necessary to capture more fine-grained person features under limited conditions due to the failure of clothes features. Capturing more fine-grained person features is the focus of solving the problem of clothes-changing person re-identification. To address this issue, this article uses a two-level attention module to guide the network to extract features from images at different scales. In addition, it utilizes the captured coarse-grained and fine-grained features for complementary learning.

The traditional attention mechanism is often learned on the global scale of the image (Hu et al., 2018; Woo et al., 2018). The attention map obtained by the traditional method is effective for the extraction of key features in the image. However, in the clothes-changing person re-identification, the clothes feature is no longer used as the discriminant condition. Currently, the judgment of the person ID requires more fine-grained features. The extraction of traditional attention is often coarse-grained, which is no longer applicable to the clothes-changing person re-identification.

The two-level attention module used in this article does not abandon the coarse-grained features. Instead, it integrates the coarse-grained and fine-grained features to jointly guide the network to learn (see Figure 2). Figure 2(a) shows the spatial two-level attention module. Figure 2(b) shows the channel two-level attention module. The features extracted by the backbone network $F \in \mathrm{R}^{C\times H\times W}$ are divided into n parts; n features are obtained after division. For spatial attention, the divided feature is $F_i \in \mathrm{R}^{C\times \frac{H}{n}\times W}$. For the channel attention, the divided feature is $F_j \in \mathrm{R}^{\frac{C}{n}\times H\times W}$. The attention operation is directly performed on the features $Att\left(F\right)$ of the original scale. A coarse-grained attention map is obtained, A. For the divided feature map, taking spatial features F_i as an example, a set of sub-attention modules is used to capture the key information of each feature. Then, the

Figure 2.
Spatial two-level attention module (a) and channel two-level attention module (b)



attention map A_i of each sub-feature is obtained. Concat all sub-attention maps to obtain a global-scale attention map A' of the same size as the attention map A. This process can be formulated as:

$$A' = Concat\left(A_i\right)_{i=1}^{n} \tag{1}$$

Finally, the coarse-grained and fine-grained feature maps obtained from different scale feature maps are fused and applied to the feature map. This process can be formulated as:

$$F' = \sigma\left(A * A'\right) * F \tag{2}$$

where F' is the feature map after the attention module, * denotes the element-wise product, σ is the sigmoid function.

In the specific implementation, this article uses the attention mechanism proposed in Woo et al. (2018). The channel attention mechanism is designed to group and aggregate those semantically similar channels, allowing the network to pay more attention to person features and reduce the influence of background. The input feature map is passed through two parallel max pooling layers and average pooling layers. It changes the feature map from $C \times H \times W$ to the size of $C \times 1 \times 1$. Then, it passes through an MLP with shared parameters. Finally, it passes through the activation function to get two results after activation.

The two output results are added element by element. Then, the output result of channel attention is obtained through an activation function. The output result is multiplied by the original image to change back to the size of the input feature. The calculation method of channel attention is:

$$\begin{aligned} M_c\left(F\right) &= \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^{c}\right)\right) + W_1\left(W_0\left(F_{max}^{c}\right)\right)\right) \end{aligned} \tag{3}$$

where F denotes the feature map and σ denotes the sigmoid function.

The spatial attention mechanism is used to capture and aggregate those semantically related pixels in the spatial domain (i.e., determine person pixel locations in space based on channel attention). The input features are obtained through max pooling and average pooling to obtain two $1 \times H \times W$ feature maps. Then, the two feature maps are spliced through the concat operation, becoming a single-channel feature map through convolution. It passes an activation function to obtain the feature map of spatial attention. Finally, it multiplies the output result by the original image to change back to the input feature size. The calculation method of spatial attention is:

$$\begin{aligned} M_s\left(F\right) &= \sigma\left(f\left(\left[AvgPool\left(F\right); MaxPool\left(F\right)\right]\right)\right) \\ &= \sigma\left(f\left(\left[F_{avg}^{s}; F_{max}^{s}\right]\right)\right) \end{aligned} \tag{4}$$

where f denotes the convolution operation. F and σ have the same definition as Equation (3).

Loss Function

In the clothes classification branch, this article uses the ASL classification loss (Ridnik et al., 2021). It can balance the weight between positive and negative samples and solve the problem of labeling errors that may occur in clothes labeling. ASL loss can be formulated as:

$$ASL = \begin{cases} (1-p)^{\gamma_+} \log(p), positive \\ (p)^{\gamma_-} \log(1-p), negative \end{cases} \tag{5}$$

where γ denotes the focusing parameter (setting γ_+ to 0 and γ_- to 4 in this article). p denotes the probability of each category output by the model, which can be formulated as:

$$p = \frac{e^{(f(x)_k)}}{\sum_{i=1}^{N} e^{(f(x)_i)}}, k \in S_{clo} \tag{6}$$

where N denotes the number of categories of clothes, $f(x)$ denotes the feature vector obtained through the clothes classification branch, k denotes the *k*-th clothes categories, and S_{clo} denotes the set of clothes. The purpose of training the clothes classification branch is to obtain the adversarial loss.

In person classification, this article jointly uses the label smooth loss (Szegedy et al., 2016) commonly used in representation learning, the triplet loss (Zhai et al., 2019) commonly used in metric learning, and the adversarial loss used to suppress the sensitivity of the model to clothes features.

The label smooth loss can be formulated as:

$$L_{id} = -\sum_{i=1}^{N} q_i \log(p_i) \tag{7}$$

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon, i = y \\ \frac{\varepsilon}{N}, \text{otherwise} \end{cases} \tag{8}$$

where N denotes the number of persons in the training set, p_i denotes the predicted probability of the output person identity, and y denotes the real signature information of the person identity. Equation (8) denotes the label smoothing operation. ε denotes a hyperparameter with a small value ($\varepsilon = 0.1$ in this article).

The triplet loss can be formulated as:

$$L = max(d(a,p) - d(a,n) + margin, 0) \tag{9}$$

where a denotes an anchor example. p denotes a sample of the same category as a and n denotes a sample of a different category. The margin denotes a hyperparameter greater than 0. In this article,

the margin is set to 0.3. The final optimization goal is to decreae the distance between a and p and increase the distance between a and n.

To make the model pay more attention to non-clothing features (based on the above losses), this article adds a clothes adversarial loss. In the person classification branch, the model has increased the distance of person features with different ids. It will also increase the distance between the same person features that have changed clothes. The purpose of adding adversarial loss is to make people with the same identity but the same persons changing clothes closer.

This article further divides the clothes classification. Based on the original clothes classification, the clothes belonging to the same person in the dataset are classified as positive samples. The others are negative samples. The purpose is to make the person classification branch unable to distinguish samples with the same id but wearing different clothes. Thus, the same person wearing different clothes can be classified into one category. The adversarial loss can be formulated as:

$$L_A = \begin{cases} \dfrac{1}{N_{S_{clo}^+}} \sum_{i=1}^{N_{S_{clo}^+}} \left(1-p^+\right)^{\gamma_+} \log\left(p^+\right), positive \\ \dfrac{\varepsilon}{N_{S_{clo}^-}} \sum_{i=1}^{N_{S_{clo}^-}} \left(p^-\right)^{\gamma_-} \log\left(1-p^-\right), negative \end{cases} \tag{10}$$

where $N_{S_{clo}^+}$ denotes the number of positive samples and $N_{S_{clo}^-}$ denotes the number of negative samples. For example, if a person has 10 sets of clothes, these 10 sets of clothes are positive samples. The rest of the clothes are negative samples. ε denotes a hyperparameter with a small value. In this article, $\varepsilon = 0.1$.

p^+ denotes the probability of a positive sample, which can be formulated as:

$$p^+ = \frac{e^{\left(f(x)_k\right)}}{e^{\left(f(x)_k\right)} + \sum_{i=1}^{N_{S_{clo}^-}} e^{\left(f(x)_i\right)}}, k \in S_{clo}^+ \tag{11}$$

p^- denotes the probability of a negative sample, which can be formulated as:

$$p^- = \frac{e^{\left(f(x)_k\right)}}{\sum_{i=1}^{N} e^{\left(f(x)_i\right)}}, k \in S_{clo}^- \tag{12}$$

where N denotes the total clothes type.

In this article, three kinds of losses are used jointly. The total loss function can be formulated as:

$$L_{\text{total}} = L_{id} + \beta L_{triplet} + L_A \tag{13}$$

where β denotes a hyperparameter. It is set to 0.1 in this article.

EXPERIMENTS

Datasets

This article verifies the effectiveness of the proposed algorithm on the popular clothes-changing person re-identification datasets PRCC and LaST.

PRCC is a public dataset for person re-identification, utilized in 2020. It consisted of 221 identities and three camera views (Yang et al., 2019). The images in the PRCC dataset includes a variety of changes, such as clothes, lighting, occlusion, pose, and perspective. There are 50 images per person in each camera's view. There are approximately 152 images per person in the dataset (a total of 33,698 images).

LaST is a large-scale spatiotemporal person re-identification dataset for person re-identification released in 2021. This includes 10,862 person identities and more than 228,156 images (Shu et al., 2021). Compared with existing data sets, LaST is more challenging and diverse. It includes children to people over the age of 70, differing time dimensions (from day to night), and persons in different cities and different countries. Seventy-six percent have changed clothes.

Evaluation Index

This article evaluates the Re-ID accuracy of the method on two datasets through the cumulative matching characteristic (CMC) and mean average precision (mAP). CMC uses *Rank-n* to represent the accuracy of ReID. The calculation method aims to calculate the accuracy of the identity query in the first *n* results. The final *Rank-n* value is the average of the *Rank-n* of all tested data. The average precision mAP score sums and averages the average precision across tasks, which reflects the overall Re-ID accuracy.

Experimental Details

The operating system of the experimental platform is Ubuntu16.04. It uses two NVIDIA 2080TI GPUs, each with 12GB of video memory. The entire network is built in the PyTorch framework. The batch size is set to 32. Each batch contains eight people. Each person contains four images.

The input image is resized to 384×192. Random horizontal flip, random crop, and random wipe are used for data augmentation. This article adds a spatial two-level attention module and a channel two-level attention module after the second residual block and the fourth residual block of the model. It sets the two as a concatenated structure. In this article, the number of segmentation n of the spatial attention is set to 4. The segmentation of the channel attention is 2. The model was trained by Adam optimizer proposed by Kingma et al (2014) for 60 epochs. The learning rate was initially set to 3.5e-4 and divided by 10 every 20 epochs.

Experimental Results

This article conducts experiments on PRCC and LaST datasets. It is then compared with popular methods. The comparison results are shown in Table 1 and Table 2.

As shown in the Table 1, regarding the PRCC dataset, the method is superior to RCSANet (Huang et al., 2021), Liget MBN (Herzog et al., 2021), 3DSL (Chen et al., 2021), and other methods in terms of Rank-1 and mAP indicators. The method does not use the assistance of other modal data. The model complexity is lower.

The method in this article is competitive on the LaST dataset, surpassing HOReID on the Rank-1 indicator (Wang et al., 2020). In the two datasets, the method has achieved excellent performance. This verifies the effectiveness of the method. Thus, it can be applied to the scene of re-identification of people changing clothes.

This article also conducted a series of experiments on the clothes-changing data of the PRCC dataset to verify the influence of the IBN-Net as the backbone network, the two-level attention module, and dual-branch structure on the performance of the algorithm in the clothes-changing scene. The results are shown in Table 3.

Table 1.
Comparison with other popular methods on the PRCC dataset

Methods	Time	PRCC			
		Clothes-Changing		Same-Clothes	
		Rank-1	mAP	Rank-1	mAP
HACNN (Li et al., 2018)	2018	21.81	-	82.45	-
GI-ReID (Jin et al., 2022)	2022	33.3	-	80.0	-
ISP (Zhu et al., 2020)	2020	36.6	-	92.8	-
RGA-SC (Zhang et al., 2020)	2020	42.3	-	98.4	-
RCSANet (Huang et al., 2021)	2021	50.2	48.6	100.0	97.2
LighetMBN (Herzog et al., 2021)	2021	50.5	51.2	100.0	98.9
3DSL (Chen et al., 2021)	2021	51.3	-	-	-
Ours	2022	52.0	52.5	100.0	99.9

Table 2.
Comparison with other popular methods on the LaST dataset

Methods	Time	LaST		
		Rank-1	Rank-5	mAP
PCB (Sun et al., 2018)	2018	50.6	68.0	15.2
ABD-Net (Chen et al., 2019)	2019	48.5	67.6	16.1
QAConv (Liao et al., 2020)	2020	64.6	82.4	22.4
HOReID (Wang et al., 2020)	2020	68.3	82.3	25.5
Ours	2022	68.9	81.4	24.1

Notes: LaST is a new and huge dataset, some existing popular methods are not suitable for this dataset or the experimental effect is not ideal, so in Table 2 authors reselect some baselines with better effects for comparison.

The experimental results in Table 3 show that using IBN-Net as the backbone network has excellent performance in the person re-identification task. The performance of the two-level attention module alone or the dual-branch pair algorithm has a good performance improvement. Incorporating both Rank-1 and mAP into the network has additional performance improvements. The IBN-Net backbone network is a powerful person feature extraction network; the two-level attention module can guide

Table 3.
Comparison results on the PRCC clothes-changing dataset

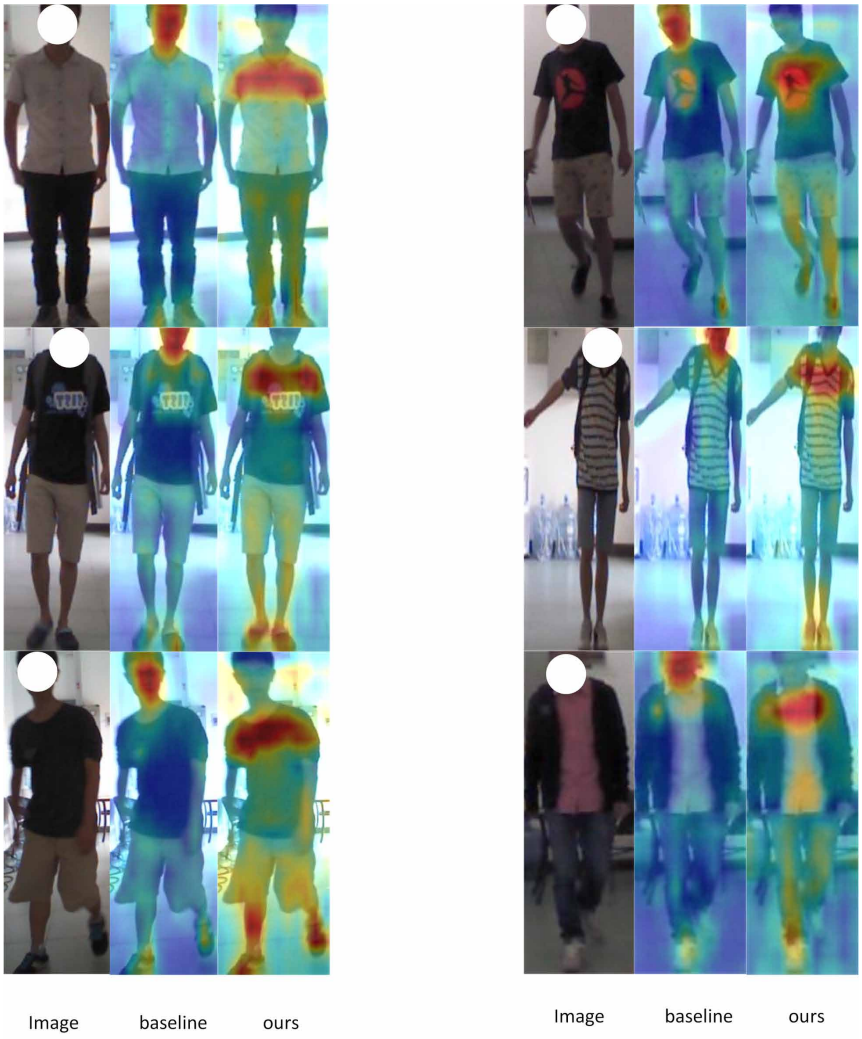
Methods	PRCC	
	Rank-1	mAP
ResNet-50	44.6	42.4
IBN-Net	45.4	43.8
IBN-Net+ Two-level Attention	46.5	47.7
IBN-Net+ Dual-branch	50.8	51.9
IBN-Net+ Two-level Attention + Dual-branch	52.0	52.5

the network to capture more fine-grained features of persons. Each module has different degrees of performance improvement on the PRCC dataset. This reflects the robustness of each module and algorithm. In addition, it can effectively improve the precision and accuracy of the clothes-changing person re-identification.

Visual Analysis

The heat map is used to visualize the attention of different models to the person image area to better show the effect of the method in the actual application process. As shown in Figure 3, the first column is the original person image, the second column is the visualization heat map using IBN-Net, and the third column is the method proposed in this article. Per Figure 3, as compared with the baseline, the model in this article pays more attention to the body area. It also has a wider range of attention. The feature maps tend to focus on shoulders, body shape, posture and body features unrelated to clothes. Therefore, the method in this article is more suitable for clothes-changing person re-identification.

Figure 3.
PRCC data visualization results



CONCLUSION

This article proposes a two-level attention module and dual-branch structure for clothes-changing person re-identification. It uses the pre-trained IBN-Net as the backbone network, adding a series of two-level spatial and channel attention mechanisms to the network. Then, it uses the person and clothes classification dual branches to suppress the network's extraction of clothes features. The method does not use the auxiliary means of other modalities. In addition, the model is more lightweight. The network structure captures the inherent features of clothes-changing persons. The final network can better mine non-clothing features from images and can cope with clothes-changing person re-identification.

The superiority and robustness of this method are verified by experiments on two popular person Re-ID datasets and comparison with existing popular methods. The article also experimentally demonstrates the effectiveness of the proposed two-level attention module and dual-branch structure. The method proposed in this article can be applied to intelligent surveillance systems, especially for tracking criminals (who try to evade detection by changing clothes) and missing persons.

REFERENCES

Chao, H., He, Y., Zhang, J., & Feng, J. (2019, July). Gaitset: Regarding gait as a set for cross-view gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 8126–8133. doi:10.1609/aaai.v33i01.33018126

Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., & Zheng, W. S. (2021). Learning 3D shape feature for texture-insensitive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8146–8155). doi:10.1109/CVPR46437.2021.00805

Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., & Wang, Z. et al. (2019). Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8351–8361). IEEE.

Gong, Y., Huang, L., & Chen, L. (2021). Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method. In Computer Vision and Pattern Recognition (pp. 1–12). Academic Press.

He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15013–15022). IEEE.

Herzog, F., Ji, X., Teepe, T., Hörmann, S., Gilg, J., & Rigoll, G. (2021, September). Lightweight multi-branch network for person re-identification. In *2021 IEEE International Conference on Image Processing* (pp. 1129–1133). IEEE.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141). IEEE.

Huang, Y., Wu, Q., Xu, J., Zhong, Y., & Zhang, Z. (2021). Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 11895–11904). IEEE. doi:10.1109/ICCV48922.2021.01168

Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., ... Hua, X. S. (2022). Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14278–14287). IEEE. doi:10.1109/ICCV48922.2021.01168

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *Computer Science*. doi:10.1109/ICCV48922.2021.01168

Li, W., Zhu, X., & Gong, S. (2018). Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2285–2294). IEEE.

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2197–2206). doi:10.1109/CVPR.2015.7298832

Liao, S., & Shao, L. (2020, August). Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *European Conference on Computer Vision* (pp. 456–474). Springer. doi:10.1007/978-3-030-58621-8_27

Pan, X., Luo, P., Shi, J., & Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision* (pp. 464–479). doi:10.1007/978-3-030-01225-0_29

Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., & Xue, X. et al. (2020). Long-term cloth-changing person re-identification. *Proceedings of the Asian Conference on Computer Vision.*

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., & Zelnik-Manor, L. (2021). Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 82–91). IEEE.

Shu, X., Wang, X., Zang, X., Zhang, S., Chen, Y., Li, G., & Tian, Q. (2021). Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision* (pp. 480–496). doi:10.1007/978-3-030-01225-0_30

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826). doi:10.1109/CVPR.2016.308

Wan, F., Wu, Y., Qian, X., Chen, Y., & Fu, Y. (2020). When person re-identification meets changing clothes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 830–831). IEEE.

Wang, G. A., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., & Sun, J. et al. (2020). High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6449–6458). doi:10.1109/CVPR42600.2020.00648

Weinberger, K. Q., & Saul, L. K. (2008, July). Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1160–1167). doi:10.1145/1390156.1390302

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision* (pp. 3–19). Academic Press.

Xu, J., Zhao, R., Zhu, F., Wang, H., & Ouyang, W. (2018). Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2119–2128). IEEE.

Yang, Q., Wu, A., & Zheng, W. S. (2019). Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(6), 2029–2046. doi:10.1109/TPAMI.2019.2960509 PMID:31869783

Zhai, Y., Guo, X., Lu, Y., & Li, H. (2019). In defense of the classification loss for person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. doi:10.1109/CVPRW.2019.00194

Zhang, Z., Lan, C., Zeng, W., Jin, X., & Chen, Z. (2020). Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3186–3195). IEEE.

Zhu, K., Guo, H., Liu, Z., Tang, M., & Wang, J. (2020, August). Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision* (pp. 346–363). Springer. doi:10.1007/978-3-030-58580-8_21