ORIGINAL RESEARCH PAPER

Pedestrian re-identification based on attribute mining and reasoning

Chao Li | Xiaoyu Yang | Kangning Yin | Yifan Chang | Zhiguo Wang | Guangqiang Yin

Center for Public Security Technology, University of Electronic Science and Technology of China, Gaoxin West District, Chengdu, China

Correspondence
Guangqiang Yin, Center for Public Security Technology, University of Electronic Science and Technology of China, No. 2006, Xiyuan Avenue, Gaoxin West District, Chengdu, China.
Email: yingq@uestc.edu.cn

Chao Li and Xiaoyu Yang contributed equally to this article.

Abstract
The high-level semantic information extracted from the pedestrian attribute feature is an important element for pedestrian recognition. Pedestrian attribute recognition plays an important role in both intelligent video surveillance and pedestrian re-identification promoting the convenience of searching and performance of model. This paper tries finding a practical method to improve the performance of the pedestrian re-identification by combining pedestrian attributes and identities. The multi-task learning method combines pedestrian recognition and attribute information in a direct way that considers the correlation between pedestrian attributes and identities but ignores the principle and degree of such correlation. To solve this problem, a new pedestrian recognition framework based on attribute mining and reasoning is proposed in this paper. To enhance the expression ability of attribute features, it designs spatial channel attention module (SCAM) based on attention mechanism to extract features from every attribute. SCAM can not only locate the attributes on the feature map, but also effectively mine channel features with a higher degree of association with attributes. In addition, both spatial attention model and channel attention model are integrated by multiple groups of parallel branches, which further improve the network performance. Finally, using the semantic reasoning and information transmission function of graph convolutional network, the relationship between attribute features and pedestrian features can be mined. Besides, pedestrian features with stronger expression ability can also be obtained. Experiment work is conducted in two databases, DukeMTMC-reID and Market-1501, which are commonly used in pedestrian recognition tasks. On the Market-1501 dataset, the final effect of the algorithm model CMC-1 can reach 94.74%, and mAP can reach 87.02%; on the DukeMTMC-reID dataset, CMC-1 can reach 87.03%, and mAP can reach 77.11%. The results show that our method is at the top of the existing pedestrian recognition methods.

1 | INTRODUCTION

At present, most existing pedestrian recognition methods are based on convolutional neural network [1–4], which usually obtains the expression features of a whole picture. However, some significant details may be ignored in this way. Pedestrian attributes express the high-dimensional semantic information of pedestrians, which provide detailed information for pedestrian recognition tasks, such as hair length, sleeve length and etc. Character attributes are highly robust to illumination, pose changes and camera changes. Different attributes of pedestrians are important supplementary information for pedestrian recognition. As shown on the right side of Figure 1, the appearance and posture of the two pedestrians in the picture are very similar. If no other information is introduced, it is easy to identify the two people as the same person, but if gender information is introduced, the two pedestrians can be clearly distinguished. Therefore, it is critical to integrate...
attribute characteristics into pedestrian identity characteristics. We use the multi-task learning method to introduce attribute information into a pedestrian recognition task directly and simply.

As details of pedestrians, attribute information, which contains detailed information of pedestrians, is just related to some parts and channels of an image. For example, the length and the colour of the upper body are only related to the characteristics of the corresponding area of the upper body while the judgement of gender should be considered from the entire image; In addition, the attribute of sleeve length is related to the shape of the edge. Therefore, it is very important to know how to locate the correct region effectively from spatial channels, which can improve the expression ability of attribute features and the effect of pedestrian recognition. Then, we design a pedestrian recognition algorithm based on attribute mining to promote the expression ability of attribute features.

After a deeper observation on data, the degrees of correlation among attributes are different and some attributes can be obtained by mutual inference from each other. As shown in Figure 2, we can infer that there is a high probability of a male on the left and a female on the right based on hair length, clothing type and appearance. This example means that there are more semantic relations between attributes, which can be explored deeply. However, in reality scene, some attributes are difficult to judge directly due to the influence of perspective and occlusion. In this case, some attributes can be inferred by the relationship between them and other attributes. Then, we design the attribute reasoning module based on the information transmission ability of graph convolutional to explore the semantic relationship among multiple attributes, attributes and overall features.

The integration of pedestrian information into a pedestrian re-identification framework can effectively suppress the negative effect of pedestrian misalignment caused by cross camera. The pedestrian attributes are still different among similar pedestrians and are less affected by the misalignment caused by cross camera. Then, we design a new pedestrian re-identification framework based on attribute mining and reasoning. In this framework, the combination of attributes and pedestrian re-identification is based on multi-task learning[6]; to improve the expression ability of attribute features, we propose two attribute mining structures of spatial channel attention module for spatial and channel respectively [7]. These modules can locate different attributes in space and channel based on attention mechanism. In addition, we use parallel spatial attention and channel attention modules to integrate multiple attention structures, which further improve the expression ability of attribute features. Besides, we propose an inference function in the framework by using graph convolutional network [5] which can build the relationship between attributes and all features. The final designed network can accurately locate the attribute-related regions and channels, infer the relationship between attributes and the entire characteristics, so as to obtain more robust and expressive features.

The main research contents and innovations are summarized as below:
1. Combine attributes with pedestrian re-identification task; propose a new network framework with the ability of attribute mining and reasoning.
2. A spatial channel attention module(SCAM) based on attention mechanism is designed. Based on the two dimensions of space and channel, the specific features of each attribute are located and derived from the shared feature map, so as to improve the expression ability of attribute features.
3. The graph convolutional network is used to model the semantic reasoning between attributes; and the relationship between attributes and the entire characteristics to obtain more robust and expressive features. Finally improve the effect of pedestrian recognition.
4. The effectiveness and advances of the design model are verified by experiments and the effect of different designs on...
the experimental results is analyzed through the ablation experiment.

The construction of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 describes a new pedestrian re-identification framework based on attribute mining and reasoning. The experimental results and conclusion will be presented in Section 4 and Section 5 respectively.

2 RELATED WORK

2.1 Multi-task learning

Traditional semantic segmentation and instance segmentation are processed independently, that is, training the separate neural network for each task and the model space between each task is independent of each other. In reality, however, many problems are multi-modal. Multi-task learning (MTL) [8] is a kind of transfer-learning algorithm. Before the appearance of deep learning, it mainly attempts to model the common information among different tasks and hopes to produce better generalization performance. The researchers assumed that the task parameters should be close to each other, such as some distance metrics or sharing the same prior probability etc. When all tasks are correlated, these hypotheses will produce a better training effect. However, if information sharing occurs between uncorrelated tasks, the training effect may be reduced.

After the appearance of deep learning, MTL designs a network that can learn and share from multiple tasks. If related tasks can share complementary information, they have potential to improve performance, so as to achieve the interaction and promotion of multi-task supervisory information. Multi-tasking can accelerate the speed of reasoning of each task by avoiding repeated computation of features in the shared layer. Parameter sharing can be divided into soft sharing and hard sharing.

Cross-stitch networks [9] introduce parameter soft sharing into the deep MTL framework, which connects two independent networks by the way of parameter soft sharing. The cross-stitch module determines how to use other networks to learn knowledge from these task-related networks and linearly combines them with the output of the previous layers. MTAN [10] uses an attention mechanism to share a common feature library among specific task networks. NDDR-CNN [11] also brings dimension reduction technology into feature fusion layer. One problem of parameter soft sharing method is scalability because the size of MTL network increases linearly with the number of tasks.

Ubernet [12] is the first parameter hard sharing model, which jointly processes a large number of low, medium and high-level visual tasks. This model adopts multi-head design across different network layers and scales. Multilinear relationship network [13] allows models to learn relationships between tasks by adding matrix priors to the fully connected layer. Stochastic filter groups [14] redesign convolution kernels at each layer to support the behaviour of shared or specific task.

Lud et al. [15] proposed a fully adaptive feature sharing (FAFS) by combining multi-task training with face attributes. Starting from the thin network, it uses the index of automatic grouping of similar tasks to dynamically widen the network greedily. However, such methods cannot achieve the goal of global optimization, which assigns an accurate task allocation to each branch. It does not allow models to learn more complex interactions among tasks. Similar to FAFS, vandenhende et al. determined task grouping based on the score of task relevance obtained by pre-calculating. In contrast to FAFS, they measured task relevance based on feature similarity scores rather than sample difficulty. The main goal of Taskonomy [16] is to explore the correlation between different visual tasks, so as to avoid repeated learning of related tasks. Meanwhile, it provides a multi-task learning framework which is easy to expand and generalize. Inspired by Taskonomy, Branched multi-task architecture search (BMTAS) optimizes network topology directly without relying on the task correlation score. With the development of multi-task learning, it has been combined with pedestrian recognition in recent 2 years.

2.2 Attributes for person re-ID

Traditional methods use attributes to support pedestrian re-identification and enhance the low-level features. Laye et al. [17] [18] proposed that the low-dimensional feature descriptor and SVM could be used as attribute detectors to introduce attributes into a metric learning method, while Su et al. [19] added the middle-level attribute feature into pedestrian description. Su et al. proposed that low-level features and camera correlations learned from attributes can be used for re-identification. Khamis et al. [20] combined attribute attribute subspace by learning discrimination projection to effectively use the interaction between attributes and appearance for pedestrian matching.

Recently, with the development of deep learning technology, Franco et al. [21] proposed a deep learning framework from coarse to fine. However, this paper ignored the correlation among attributes. Su et al. [22] firstly trained the network on dataset with attribute labels and, then, use triplet loss function to fine-tune the network. Besides, there is also a combination work of predictable attribute labels and independent dataset for the final adjustment. Schumann et al. [23] pre-trained the network on independent dataset with labelled attributes then fine-tune on another dataset with individual identities. Wang et al. [24] proposed an unsupervised re-identification method which can share knowledge from source domain based on attributes learned from tagged source dataset. Shi et al. [25] used pedestrian attributes to conduct attribute mining analysis and reasoning analysis, which achieved good experimental results, but the authors did not make too many improvements in feature networks, loss functions, attribute labels, attribute classifiers, etc. There is further room for improvement.
METHOD

3.1 Multi-task learning

Multi-task learning combines two or more tasks for learning, to achieve the interaction and promotion of the supervision information and get better results. As the detailed information of pedestrian, attribute information is very appropriate to be combined with pedestrian re-identification by multi-task learning.

We design a network based on multi-task learning (MTNet) and the structure is shown as Figure 3.

(1) Stride = 1: The original 50-layer residual backbone network (ResNet50[26]) will decrease the input size to 1/32 of the original size. In this task, it can only obtain the output with the size of 8×4, while the original size is 256×128. The feature map with very small size will lose a lot of image information, which poses a negative influence on pedestrian feature extraction. Therefore, we remove the down-sampling operation in the last residual unit of stage 4, but keep the down sampling operation of the first three stages, so that the size of the feature map obtained in each stage is 64×32, 32×16, 16×8 and 16×8, respectively.

(2) Double pooling: To combine attribute recognition with pedestrian re-identification, most network-based methods[17][18, 19], after global pooling, allocate a branch to a fully connected layer in addition to allocating a fully connected layer to every attribute. This method is simple but ignores the difference between attribute feature space and identification attributes space. Compared to the requirement of re-identification that feature should contain the information of whole image, the feature used in attribute identification should contain more location information. As shown in Figure 4, the middle two heat maps are the response of network for attribute training only to the input images. The right two images are the response of network for identities training only to input image and the red-coloured part represents the attention area from network. It can be concluded that the attribute-trained network focuses more on the location of pedestrian components, such as backpacks, jackets, shoes, hair etc. The identity-trained network pays more attention to pedestrian structure, and even pays attention to the surrounding environment of the pedestrian. This is because people in the same environment are more likely to have the same identity. So one of the two images on the right focuses more on the pedestrian structure, and the other focuses more on the surrounding environment. Therefore, it will pose a negative effect on the final result when we use the same attribute to finish re-identification and attribute identification simultaneously.

We use dual-pooling method to overcome above problems. Firstly, the feature map obtained from backbone network passes through the global average pooling and global max pooling respectively and obtains two different 2048-dimensional feature vectors. For each different attribute, the feature vector obtained from global max pooling is sent to a fully connected layer to predict the probability of the attribute. Besides, the feature vector obtained by global average pooling is sent to the pedestrian identity classifier (fully connected layer) to predict the pedestrian identity. Such design of network can classify attribute feature space and identity feature space very well, which can lead to a promotion effect on these two tasks.
Meanwhile, the network can increase attribute significance by combining features supervised by attribute can effectively promote the performance of pedestrian re-identification.

### 3.2 Attribute mining

The attributes of pedestrians are diverse, such as age and gender, belonging to the overall attributes of pedestrians, while the attributes of hair and clothes belong to the local attributes of pedestrians. Although all information is detailed, the overall attributes should consider the whole feature area of pedestrians in the judgement and make a comprehensive judgement on the whole body of pedestrians. The network model shares the same basic network. There are different correlation degrees between the learned high-dimensional channels and the different attributes. Alternatively, the negative correlation could be found between some channels and features. In response to the above-mentioned attribute diversity problem, the pedestrian recognition algorithm based on attribute mining is proposed by using attention mechanism. Moreover, the spatial attention model and the channel attention model are designed respectively by employing attention mechanism, and these two attention models are combined to present a spatial channel attention module (SCAM), which locates attributes from the two dimensions of space and channels to implement attribute mining. The diversity of pedestrian attributes is also reflected in the different contents of concerns. The attributes such as the colour of the upper garment and the colour of the lower garment pay more attention to the colour features of a certain area of the image, while the attributes such as the type of the upper garment and the type of the lower garment pay more attention to the contour features of a certain area of the image.

We design a pedestrian recognition network structure (AMNet) based on attribute mining, as shown in Figure 5.

1. **Spatial attention**
   - The shared feature map obtained from basic network first passes through the convolution layer with pad = 1, stride = 1 and convolution kernel size of 3×3 to reduce dimensions and the number of output channels is c’, which makes the normalization operation[27] on the dimension-reduced feature and activation operation by ReLu function[28]. A 1×1 convolution layer is posed after ReLU layer which is used to produce the attention value of Softmax on the corresponding position of the feature map. As shown in Figure 6.

2. **Channel attention**
   - The shared feature images obtained from basic network first pass through the global average pooling layer[29], which can change the dimension of feature map from h×w×c to 1×1×c. The transformed feature passes through the convolution layer with pad = 0, stride = 1, convolution kernel size of 1×1 and c’ channels. The dimension-reduced feature passes through the BatchNorm layer to do normalization operation and ReLU layer to do activation operation. Then, the feature passes through the convolution layer with pad = 0, stride = 1, convolution kernel size of 1×1 and c channels, which can revert the dimension to the original size and obtain the attention value. Finally, feature passes through the sigmoid function and obtains the map of channel attention, as shown in Figure 7.

3. **Spatial channel attention module**
   - Since there could be some bias in attention model obtained by semi-supervised learning, we tend to use model integration approach to solve this problem. To be more specific, we train multiple attention model simultaneously and integrate them as a final model which is called Spatial Channel Attention Module (SCAM), as shown in Figure 8.
An attention map is obtained by integrating the results of multiple attention branches which are trained separately, the specific strategy is shown below.

$$A_E = \frac{1}{m} \sum_{i=1}^{m} A_i$$  \hspace{1cm} (1)

$A_E$ represents the result of attention after integration; $A_1, A_2 ... A_m$ represent the sequence of attention result before integration; $m$ represents the number of branches.

### 3.3 Attribute reasoning

According to the observation on dataset, there exists a kind of semantic reasoning relationship among pedestrian attributes and between attributes and pedestrians’ identities. It can be concluded that the person is female based on the long hair, skirt-dressing and other attributes without any facade information of this person. Besides, it is easy to make a general judgement based on the attribute information of height, the level of strength, the colour of the hair etc. Moreover, people are good at describing a special pedestrian based on the type of clothes, appearance information, gender, age and other attributes. Therefore, we use graph convolutional network to model the semantic reasoning relationship among attributes; between attributes and pedestrians’ identities. In this way, we propose a pedestrian re-identification algorithm based on attribute reasoning, which combines the attribute identification with pedestrian re-identification more reasonably.

We define graph convolutional operation cited from Kipf et al.[5] and Li et al.[30]:

$$O = GVW_g = ((I - A_g)V)VW_g.$$ \hspace{1cm} (2)

$A_g$ represents the $M \times N$ matrix, which is used for information diffusion, which design three transformations to show the three different connections among nodes: Simply connected, Fully connected and Adaptively connected. $W_g$ represents the parameter of state update equation, $I$ is used to decrease the difficulty of calculation optimization. Meanwhile, make a Laplacian Smoothing before convolution operation.

The global features of 2048 dimension were obtained by global average pooling of the shared feature map, and 2048 dimensions features of each attribute were obtained by spatial channel attention module and full connection layer ascending dimension, so we can get the characteristic node matrix $V$.

$$V = \begin{bmatrix} v_{global} \\ v_{atr1} \\ \vdots \\ v_{atrm} \end{bmatrix}.$$ \hspace{1cm} (3)

(1) Simply connected

$$A_g = \begin{bmatrix} 0 & 1 & \cdots & 1 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$ \hspace{1cm} (4)

1 indicates that there exists an edge between nodes and 0 indicates that there is no edge between nodes. The values of elements in the first row and column of the matrix are 1 and others are 0. Such connection will lead to less information transfer among nodes. Information needs all nodes to make indirect
transfer. So the attribute reasoning ability is relatively weak in this way. As shown in Figure 9.

(2) Fully connected

\[
\mathbf{A}_{\text{fg}} = \begin{bmatrix}
0 & 1 & \cdots & 1 & 1 \\
1 & 0 & \cdots & 1 & 1 \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
1 & 1 & \cdots & 0 & 1 \\
1 & 1 & \cdots & 1 & 0
\end{bmatrix}.
\]

(5)

Except for the diagonal elements, value of all elements equal to 1. Such connection method will not only leads an effective attribute information transfer to all nodes and all nodes can transfer information to each other directly. The advantage lies in the direct reasoning among attributes rather than transfer by all nodes; the disadvantage is that irrelevant nodes can also transfer information, thus causing error interference, as shown in Figure 10.

(3) Adaptively connected

\[
\mathbf{A}_{\text{ag}} = \begin{bmatrix}
0 & \gamma_{1,0} & \cdots & \gamma_{1,m-1} & \gamma_{0,m} \\
\gamma_{1,0} & 0 & \cdots & \gamma_{1,m-1} & \gamma_{1,m} \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
\gamma_{m-1,0} & \gamma_{m-1,1} & \cdots & 0 & \gamma_{m-1,m} \\
\gamma_{m,0} & \gamma_{m,1} & \cdots & \gamma_{m,m-1} & 0
\end{bmatrix}.
\]

(6)

Adaptive connection is obtained by learning based on connection relationship among nodes. The dotted line indicates that the connection among nodes needs to be learned. When the adaptive connection is adopted, all other elements are assigned with random values by back-propagation learning except the elements on the diagonal line, which are assigned 0 and do not participate in learning. The value closer to 1 means a stronger connection between two nodes and the value closer to 0 means the weaker connection between two nodes. As shown in Figure 11.

We designed the Pedestrian Re-identification network (ARNet) based on attribute reasoning, as shown in Figure 12.
3.4 Loss function

Regarding the loss of attribute classification, cross entropy loss is used as a loss function. For M attributes, each attribute contains 2 types: 0 and 1. The loss function of attribute m can be expressed as follows.

$$L_{ATT}^{(m)} = -y^{(m)} \log(\hat{y}^{(m)}) + (1 - y^{(m)}) \log(1 - \hat{y}^{(m)}).$$ (7)

Accordingly, for all attributes, the loss function of attribute classification can be obtained as follows:

$$L_{ATT} = \frac{1}{M} \sum_{i=1}^{M} L_{ATT}^{(m)}. (8)$$

Regarding the loss of pedestrian identity classification, softmax loss is used as the loss function, where y is the tag of one’s identity.

$$L_{ID} = -\sum_{k=0}^{N} 1\{y = k\} \log(\hat{p}_k).$$ (9)

For the loss of pedestrian identity characteristics, triplet loss is used as the loss function. The purpose of this is to shorten the distance for the negative sample $x_n$ between one of sample $x_a$ and its positive sample $x_p$.

$$L_{tri} = \max(||x_a - x_p|| - ||x_a - x_n|| + \text{margin}, 0).$$ (10)

The loss of pedestrian identity classification and attribute classification is combined by weighted, forming the loss of multi-task learning.

$$L = L_{ATT} + L_{ID} + \alpha L_{ATT},$$ (11)

$\alpha$ is used to control the effect of attributes in the process of multi-task learning and the hyperparameter of pedestrian re-identification algorithm based on multi-task learning.

| Entry name | Configuration description |
|------------|--------------------------|
| CPU        | Intel® Xeon E5-2678V3 CPU @2.5GHz |
| GPU        | NVIDIA TITAN RTX 24G |
| Memory     | RECC4 2666MHz 32Gx4 |
| Operating System | Ubuntu 18.04 X86_64 |
| Experimental Framework | PyTorch 1.6.0 cuda 10.2 |

4 | EXPERIMENTS

Pytorch is used as the experimental framework to implement the proposed pedestrian re-identification, using Titan RTX graphics card for training and testing. With the excellent performance of the graphics card in data processing and matrix calculation, it can effectively improve the efficiency of the experiment and save time and cost. Table 1 lists the main hardware and software environment of this experiment.

4.1 Dataset and evaluation

4.1.1 Dataset

In order to test the effectiveness of the designed network, we conducted experiments on two public pedestrian recognition datasets with pedestrian attribute data. The two datasets are described as follows:

(1) Market-1501[31]: This dataset contains 32668 pedestrian images taken by six cameras from different perspectives, with 1501 pedestrian identities. The dataset is divided into a training set and a test set. The training set contains 751 pedestrian identities with a total of 12,936 images, while the test set contains 750 pedestrian identities with a total of 19,732 images. During the test, the test set is divided into a query set with 3368 images and a gallery set with 16,364 images. Each pedestrian is labelled with 27 types of attributes, including gender, hair length, cuff length, bottom length, bottom type, whether to wear a hat, whether to carry a handbag, whether to carry a backpack, whether to have a shoulder bag, age, 8 colours of upper outer garment and 9 kinds of trousers. Except for 4 types of age, the others are only marked with 2 types.

(2) DukeMTMC-reID[32, 33]: This dataset is a subset of the DukeMTMC cross camera-tracking dataset. It contains 36,411 images of 1812 pedestrian identities captured by eight cameras, of which 1404 pedestrian identities appear in more than two cameras, and 408 pedestrian identities only appear in one camera as pedestrian identities in the interference test set. 1404 pedestrian identities are divided into training set with 702 identities and test set with 702 identities. Each pedestrian’s identity is labelled with 23 types of attributes, including gender, shoe type, hat, handbag, backpack, shoe colour, coat length, 8 colours of upper outer garment and 7 colours of trousers. Each attribute only has two types.
4.1.2 | Evaluation

There are two evaluation indexes since this paper involves two identification tasks. For attribute recognition tasks, we calculate the classification accuracy for each attribute; for pedestrian re-identification task, we select two official evaluation indicators, which are CMC and mAP, respectively. CMC index firstly calculates the distance between the question set and query set to get a distance ranking table from small to large, then calculates the rank-k accuracy rate by calculating whether the images in the top k query set contain the identity of the query. We only use rank-1 results as a representative of CMC indicators. CMC can be expressed as follows:

\[ CMC - k = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{r(q_i) \leq k\}. \quad (12) \]

CMC-k cannot reflect the all query ability of pedestrian re-identification system. So mAP index should be introduced. Indicator mAP is the average accuracy (AP) of all images in the query set, where AP calculates the first n (n = 1, 2, 3, ...) of query results. It can also be concluded that the CMC index reflects the retrieval accuracy while map reflects the recall rate.

\[ AP = \frac{1}{R_n} \sum_{j=1}^{n} \frac{\mathbb{1}\{c(q_j)\}}{j} R_j, \quad (13) \]

\[ mAP = \frac{1}{m} \sum_{i=1}^{m} AP_i. \quad (14) \]

4.1.3 | Experimental setting

(1) Data pre-processing: First, the size of input data is decreased to 256x128 and each batch is set to 64. Input data will be enhanced by horizontal mirror transformation, filling, random clipping and random erasing. After that, the data are de-averaged and normalized.

(2) Network: To compare with other advanced methods more conveniently, we choose ResNet50 which has been pre-trained on ImageNet as the basic network. The convolution step size of the last convolution layer of ResNet50 is changed from 2 to 1. After passing through the basic network, a 16x8 size feature map will be obtained.

(3) Training setting: Stochastic Gradient Descent (SGD) is used to train the network. During training, the weight decay is set to 0.00035 and the initial learning rate is set to 0.00035. The total number of iterations of the model is 200. When the number of iterations reaches 40, 70, 140, 170, the learning rate becomes 1/10 of the original.

4.2 | The results and analysis of hyperparametric experiment

4.2.1 | Parameter \( \alpha \)

Parameter \( \alpha \) is used to control the role of attributes in the process of multi-task learning. The larger value means that there is the greater influence of attribute on pedestrian re-identification and vice versa. For different values of parameters, comparative experiments have been finished on Market-1501 and DukeMTMC-reID. The results are shown in the Figure 13.

It can be found that with the increase of the value, the detail information can make positive effect on pedestrian re-identification. However, when the value is increasing, too large value will lead the poor result, which means that too much detailed information may make model ignore the overall information. According to the experimental results, we choose 0.3 as the parameter in the experiment.

FIGURE 12 Pedestrian re-identification network based on attribute reasoning (ARNet)
4.2.2 Branch experiment

In the spatial channel attention module, branch number $m$ as an important super parameter will have a significant impact on the experimental result. Specifically, if the number of branches is small, the invalid information is more likely to affect the integration effect. On the contrary, if the number of branches is too large, some weak but still valid information will be ignored, the training process will be too complex and time-consuming. In the process of the experiment, for different values, comparative experiments are conducted on Market-1501 and DukeMTMC-reID. The results are shown in the Figure 14.

It can be found that with the increase of the number of branches, the effect is improved obviously at the beginning and becomes worse after reaching the top, but the process of deterioration is slow. This shows that the increase in the number of branches can help the model focus on more important features and the less important information is gradually eliminated. According to the experimental results, we choose $m=4$ as the parameter set in the experiment.

4.2.3 Connection experiment

We design three kinds of connection methods: simply connected, fully connected and adaptively connected. Three connection methods are corresponding to different semantic relationships. The results are shown in the Table 2.

The experimental results show that a simple connection can only slightly improve the performance of the attribute-based method and the promotion effect of fully connection is better than that of simple connection. However, both of them cannot perform and adaptive connection. The information transfer in simple connections is less than that in fully connections. However, adaptively connections can adaptively select the correlation among attributes and overall features by learning. We finally chose the adaptive connection.
4.3 Pedestrian re-identification results and analysis

4.3.1 Ablation experiment

We conduct ablation experiments on pedestrian re-identification algorithm proposed in this paper. According to the experimental results on Market-1501 and DukeMTMC-reID, the influence of each module on the performance is analyzed. The results are shown in Table 3. Baseline represents the result of training pedestrian identity tags with resnet50 without using attribute tags.

From the data in Table 3, it can be concluded that baseline can only achieve CMC-1 = 88.84% and mAP = 71.59% on Market-1501 dataset. The pedestrian re-identification algorithm based on multi-task learning can achieve CMC-1 = 86.41%, mAP = 94.36%. The performance of the model can be improved by modifying the stride = 1 and double pooling. By reducing the stride = 1 and double pooling module, we verified that they could improve the performance of the model. It is proved that the attribute information can provide more details for the model and improve the performance of the model in the pedestrian recognition task. By using attention mechanism, the result reaches CMC-1 = 94.56%, mAP = 86.71%, which shows that the idea of using attention mechanism to mine attributes and improve pedestrian recognition performance is correct. When the graph convolutional network is used, another significant performance improvement is achieved, the result reaches CMC-1 = 94.74%, mAP = 87.02%. The conjecture that there is a reasoning relationship between the new root identity and the attributes and between the attributes is proved. It is correct to combine the attributes and pedestrian re-identification by using the reasoning method. The same is true for the DukeMTMC-reID dataset.

4.3.2 Comparison with the existing method

We make comparison between our model and current methods which are based on Market-1501 dataset and DukeMTMC-reID dataset. For DukeMTMC-reID and Market-1501, resolution and background of images in DukeMTMC-reID are more variable. DukeMTMC-reID’s variability results from a wider camera view and complexes camera scene for sampling. Thus, DukeMTMC-reID is more challengeable than Market-1501 in person search task. There is a comparison between the ARNet proposed in this paper and other existing methods based on Market-1501 and DukeMTMC-reID datasets. The results shown in Table 4 include methods based on traditional manual features and methods based on deep learning.

As can be seen from Table 4, CMC-1 and mAP of our method in Market-1501 dataset are 97.74% and 86.20%, respectively. Compared with AANet, which also uses attribute features, our mAP and CMC-1 are improved by 3.61% and 0.81% respectively. Compared with other experimental results, we can find that the amplitude of mAP improvement is significantly higher than that of CMC-1, and even our mAP is higher than BFE, but CMC-1 is lower than BFE, which indicates that our network model can improve the recall rate more than the accuracy rate.

4.4 Attribute recognition results and analysis

Since the test set of pedestrian recognition is divided into gallery set and query set, and the query set contains more images, we use query set as the test set of attribute recognition. Besides, since each colour of clothes is taken as a label, we select the mean values of 27 and 23 accuracy sums. The results are shown in Tables 5 and 6.

### Table 5 Attribute recognition accuracy on Market-1501

| Attribute | young | teenager | adult | old | Backpack | Bag  | Handbag | Clothes | Down | Up  |
|-----------|-------|----------|-------|-----|----------|------|---------|---------|------|-----|
| ARNet(Ours) | 99.80 | 87.60 | 89.70 | 99.40 | 84.70 | 76.60 | 90.50 | 90.10 | 94.00 | 93.50 |
| Attribute | Hair | Hat | Gender | upblack | upwhite | upred | uppurple | upyellow | upgray | upblue |
| ARNet(Ours) | 87.60 | 97.10 | 92.20 | 90.70 | 89.70 | 90.30 | 97.20 | 90.90 | 86.70 | 91.60 |
| Attribute | upgreen | downblack | downwhite | downpink | downpurple | downyellow | downgray | downblue | downgreen | downbrown | Average |
| ARNet(Ours) | 92.80 | 86.40 | 94.60 | 97.30 | 99.90 | 83.00 | 80.80 | 97.30 | 93.10 | 91.50 |

### Table 6 Attribute recognition accuracy on DukeMTMC-reID

| Attribute | Backpack | Bag  | Handbag | Boots | Gender | Hat  | Shoes | Top | upblack | upwhite | upred | uppurple |
|-----------|----------|------|---------|-------|--------|------|-------|-----|---------|---------|-------|---------|
| ARNet(Ours) | 79.30 | 83.50 | 93.60 | 88.80 | 84.60 | 80.00 | 90.00 | 88.00 | 81.70 | 93.90 | 95.40 | 99.60 |
| Attribute | upgray | upblue | upgreen | upbrown | downblack | downwhite | downred | downgray | downblue | downgreen | downbrown | Average |
| ARNet(Ours) | 88.90 | 91.60 | 97.60 | 98.00 | 76.10 | 92.70 | 98.50 | 78.60 | 99.70 | 95.80 | 89.95 |
As can be seen from the results of Tables 5 and 6, we achieved relatively high recognition accuracy on Market-1501 and DukeMTMC-reID datasets. Most of the existing work adopts the method of single attribute multi-classification, while we adopt the method of single attribute binary classification. Therefore, we only compared the same dichotomous attributes with AANet. On the Market-1501 dataset, the dichotomous attributes of gender, hair and handbag showed a certain improvement, while the three attributes of backpack, bag and hat showed a certain decrease. In this framework, attribute features are used to assist the extraction of global features of pedestrians in pedestrian re-recognition. We do not use attribute to judge the difference of pedestrians, so no specific adjustment is made to the network details. In the aspect of attribute recognition, it has no outstanding contribution.

5 | CONCLUSION

In this paper, we start from pedestrian re-identification based on multi-task learning and design a new pedestrian re-identification framework based on attribute mining and reasoning. The final network can accurately locate regions and channels related to attributes. Besides, it can also infer the relationship between attributes and overall feature, so as to obtain more robust and expressive features. On the Market-1501 dataset, the final effect of the algorithm model CMC-1 can reach 94.74%, and mAP can reach 87.02%; on the DukeMTMC-reID dataset, CMC-1 reaches 94.74%, and mAP can reach 77.11%. Therefore, it is superior to the existing advanced re-identification method on multiple benchmark datasets.

ACKNOWLEDGEMENTS

The authors also would like to thank Center for Public Security Information and Equipment Integration Technology, UESTC for providing computation platform.

FUNDING

None

ORCID

Chao Li https://orcid.org/0000-0002-2259-196X
Xiaoyu Yang https://orcid.org/0000-0001-7788-4324
Kangning Yin https://orcid.org/0000-0001-8652-0151
Yifan Chang https://orcid.org/0000-0002-2903-3418
Zhiguo Wang https://orcid.org/0000-0002-5652-5362
Guangqiang Yin https://orcid.org/0000-0003-2178-2147

REFERENCES

1. Li, W., et al.: DeepReID: Deep filter pairing neural network for person re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 152–159
2. Cheng, D., et al.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 1335–1344
3. Karanman, S., et al.: Leveraging local neighborhood topology for large scale person re-identification. Pattern Recognition 2014, pp. 3767–3778
4. Lin, J., et al.: Consistent-aware deep learning for person Re-identification in a Camera Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3396–3405
5. Kipf, T., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. Presented at ICLR 2017
6. Ou, X., et al.: Improving person re-identification by multi-task learning. Multitimed Tools Appl 78, 28257–28283 (2019)
7. Woo, S., et al.: CBAM: Convolutional Block Attention Module. In: Computer Vision - ECCV. Lecture Notes in Computer Science, vol 11211. Springer, Cham (2018)
8. Caruana, R.: Multitask Learning. Mach. Learn. 28, 41–75 (1997)
9. Misra, I., et al.: Cross-stitch networks for multi-task learning. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 3994–4003
10. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1871–1880
11. Gao, Y., et al.: NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3200–3209
12. Kokkinos, I.: UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5454–5463
13. Long, M., Wang, J., Yu, P.S.: Learning multiple tasks with nonlinear relationship networks. NIPS (2015)
14. Bragman, F., et al.: Stochastic filter groups for multi-task CNNs: Learning specialist and generalist convolution kernels. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1385–1394
15. Lu, Y., et al.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1131–1140
16. Zamir, A.R., et al.: Taskonomy: Disentangling task transfer learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 3712–3722
17. Layne, R., Hospedales, Gong, S.: Re-id: Hunting Attributes in the Wild. In: Proceedings of the British Machine Vision Conference. BMVA Press, UK (2014)
18. Layne, R., et al.: Person re-identification by attributes. In: Proceedings of the The British Machine Vision Conference, pp. 1–11. (2012)
19. Su, C., et al.: Multi-task learning with low rank attribute embedding for person re-identification. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 3739–3747
20. Khamis, S., et al.: Joint learning for attribute-consistent person re-identification. In: European Conference on Computer Vision, pp. 134–146. Lecture Notes in Computer Science, vol 8927. Springer, Cham (2014)
21. Franco, A., Oliveira, L.: Convolutional covariance features: Conception, integration and performance in person re-identification. Pattern Recognit. 61, 593–609 (2016)
22. Su, C., et al.: Multi-task learning with low rank attribute embedding for person re-identification. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 3739–3747
23. Schumann, A., Stieflahgen, R.: Person re-identification by deep learning attribute-complementary information. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, 2017, pp. 1435–1443
24. Wang, J., et al.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 2275–2284
25. Shi, Y., et al.: Person retrieval in surveillance videos via deep attribute mining and reasoning. In: IEEE Transactions on Multimedia (2020)
26. He, K., et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770–778 (2016)
27. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML’15: Proceedings of the 32nd International Conference on International Conference on Machine Learning 37, 448–456 (2015)
28. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. J. Mach. Learn. Res. 315–323 (2011)
29. Lin, M., Chen, Q., Yan, S.: Network In Network. Comput. Sci. (2013)
30. Li, Q., Han, Z., Wu, X.M.: Deeper insights into graph convolutional networks for semi-supervised learning. In: Thirty-Second AAAI Conference on Artificial Intelligence, pp. 3538–3545 (2018)
31. Zheng, L., et al.: Scalable person re-identification: A benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV) Santiago, 2015, pp. 1116–1124
32. Ristani, E., et al.: Performance measures and a data set for multi-target, multi-camera tracking. In: Computer Vision- ECV Workshops. Lecture Notes in Computer Science, vol 9914. Springer, Cham (2016)
33. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3774–3782
34. Sun, Y., et al.: SVDNet for pedestrian retrieval. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3820–3828
35. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. IEEE Trans. Circuits Syst. Video Technol. 29(10), 3037–3045 (2019)
36. Si, J.: Dual attention matching network for context-aware feature sequence based person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 5363–5372
37. Sun, Y., Zheng, L., Yang, Y.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11208. Springer, Cham (2018)
38. Sun, Y., Xu, Q., Li, Y.: Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. (2020)
39. Kalayeh, M.M., et al.: Human semantic parsing for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 1062–1071
40. Hou, R., et al.: Interaction-and-aggregation network for person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 9309–9318
41. Tay, C., Roy, S., Yap, K.: AANet: Attribute attention network for person re-identifications. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 7127–7136
42. Dai, Z., Chen, M., Gu, X.: Batch dropblock network for person re-identification and beyond. In: The IEEE International Conference on Computer Vision (ICCV) 2019, pp. 3691–3701