Multimedia Systems (2022) 28:1069–1081
https://doi.org/10.1007/s00530-022-00893-y

REGULAR PAPER

Pedestrian attribute recognition based on attribute correlation

Ruijie Zhao[1] · Congyan Lang[1] · Zun Li[1] · Liqian Liang[1] · Lili Wei[1] · Songhe Feng[1] · Tao Wang[1]

Received: 8 September 2021 / Accepted: 12 January 2022 / Published online: 14 February 2022

Abstract

Pedestrian attribute recognition is widely used in pedestrian tracking and pedestrian re-identification. This task confronts two fundamental challenges. One comes from its multi-label nature; the other one comes from the characteristics of data samples, such as the class imbalance and the partial occlusion. In this work, we propose a Cross Attribute and Feature Network (CAFN) that fully exploits the correlations between any pair of attributes for the pedestrian attribute recognition to tackle these challenges. Concretely, CAFN contains two modules: Cross-attribute Attention Module (C2AM) and Cross-feature Attention Module (CFAM). C2AM enables the network to automatically learn the relation matrix during the training process which can quantify the correlations between any pair of attributes in the attribute set, and CFAM is introduced to fuse different attribute features to generate more accurate and robust attribute features. Extensive experiments demonstrate that the proposed CAFN performs favorably compared with state-of-the-art approaches.

Keywords Pedestrian attribute recognition · Attention mechanism · Multi-label classification · Attribute correlation · Visual feature correlation

1 Introduction

Pedestrian attribute recognition, which aims to predict a label set of semantic pedestrian attributes for each given image, has been widely used in various computer vision tasks, such as pedestrian detection, pedestrian retrieval [1] and pedestrian re-identification [2, 3].

Previous works in this field can be roughly categorized into two groups: studies which focus on feature learning and studies that value the attribute correlations. The first group of methods has shifted the research focus from global features to local features and attention-based features. Early studies [4, 5] utilize standard convolutional neural network to extract fully connected features as global features, which is shown in Fig. 1a. Concerning this kind of global features possess less discriminative capacity, there emerge a series of approaches that put emphasis on local features on equally partitioned or learned human parts. For example, Zhu et al. [6] divided the image into 15 overlapping patches and adopted corresponding local parts for the classification of specific attribute classification. Works [7–9] introduced clustering method to further refine the obtained part information. Studies [10, 11] applied poselet [12] to obtain the key points of the pedestrian and transform these key points into part information. Yang et al. [13] proposed a specific key point localization network and an adaptive generator of bounding box for each part. Li et al. [14] added human-centric and scene-level contexts to bounding box regions; Liu et al. [15] took advantage of EdgeBoxes to create regions proposals and obtain local features from the input. Nowadays, attention-based methods have become popular. Some works [16, 17] capture attention from multiple layers and Sarfraz

Communicated by Z. Liu.

✉ Congyan Lang
cylang@bjtu.edu.cn

Ruijie Zhao
19120454@bjtu.edu.cn

Zun Li
lznus2018@gmail.com

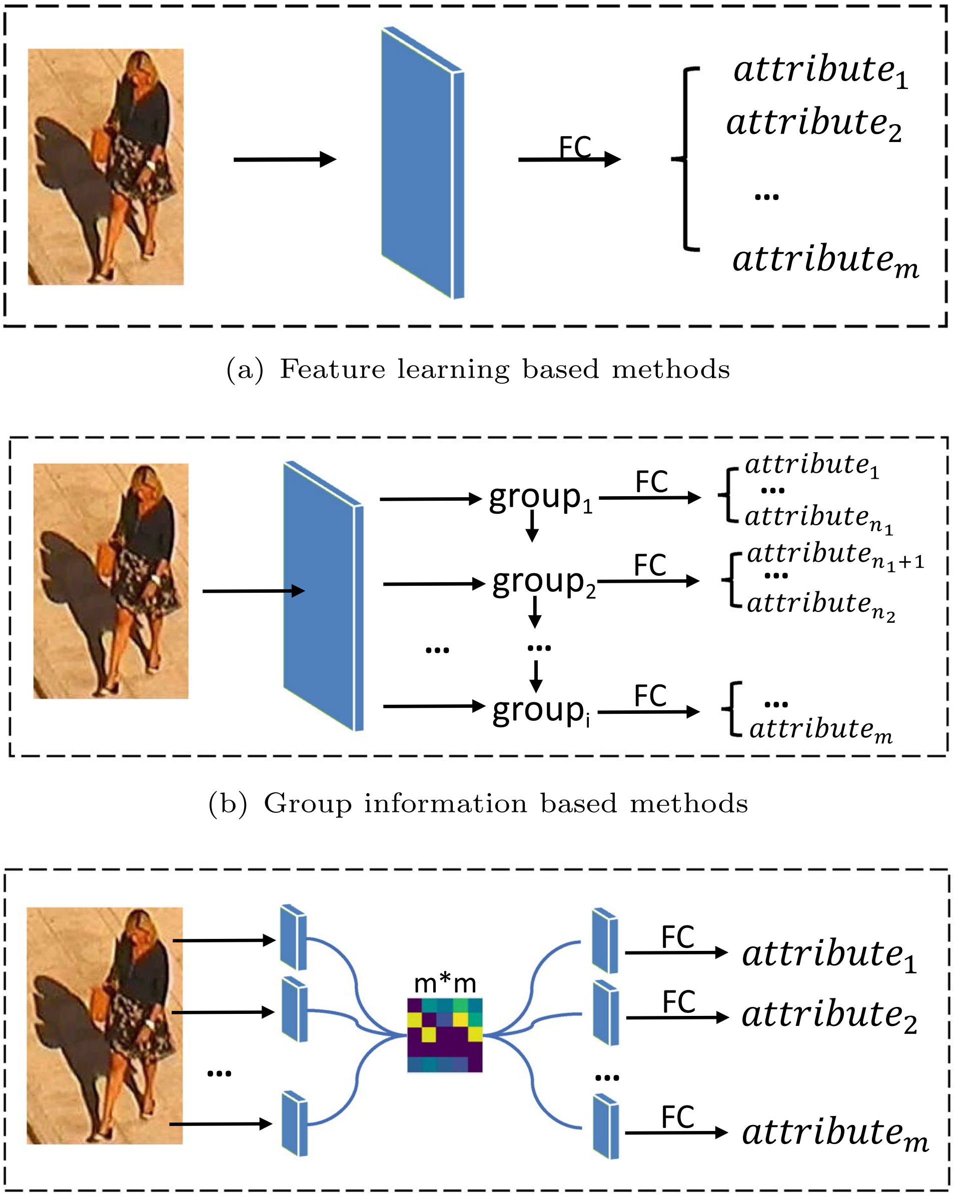
Liqian Liang
lqliang@bjtu.edu.cn

Lili Wei
20112014@bjtu.edu.cn

Songhe Feng
shfeng@bjtu.edu.cn

Tao Wang
twang@bjtu.edu.cn

[1] Beijing Jiaotong University, Beijing 100044, China



(a) Feature learning based methods

(b) Group information based methods

(c) Ours

Fig. 1 Illustration of different methods

et al. [18] captured attention from visual cues. Guo et al. [19] stressed the importance of refining attention heat map for each attribute. However, all the methods mentioned above pay less attention to the correlations between attributes, which plays a crucial role in alleviating the common sample imbalance and partial occlusion issues in this area. Here the sample imbalance means that there exist several attributes with only a small amount of training samples. And partial occlusion means that several pedestrian attributes are partially or completely blocked by other objects in the image. It is difficult for the network to learn distinguishable features to describe and recognize such attributes. However, with the help of the correlation information between attributes, the network can identify these attributes more accurately.

To better leverage the attribute correlations, current studies normally concentrate on two aspects: multi-stage prediction and sequential prediction. Multi-stage based methods [20] often involve multiple recognition and prediction stages, where the classifiers need to be trained separately. In contrast, sequence-based methods [21–24] employed time series model to capture the interdependence and correlation among attributes, among which a typical group information related to attributes is introduced in [22, 24] to give a boost to intra-group feature sharing and inter-group feature competition, as shown in Fig. 1b. However, the group information requires much manual annotation and the model can be easily affected by the prediction order of attributes. Moreover, these methods only consider the correlations between a partial set of attributes instead of the whole set of attributes.

Motivated by the aforementioned observations, we consider investigating the attribute correlations more sufficiently by modeling the relationships between any pair of attributes in the attribute set. Towards this end, we propose a Cross Attribute and Feature Network (CAFN), which extracts multi-scale features and learns the attribute correlations via two newly proposed modules: Cross-attribute Attention Module (C2AM) and Cross-feature Attention Module (CFAM). The former one aims to deeply mine the correlations among attributes by automatically learning a relation matrix of attributes, as shown in Fig. 1c. In addition, the latter one is introduced on top of C2AM to fuse multiple matrices of attribute features, aiming to obtain a feature matrix with higher generalization capacity. Through these two modules, the abstract relationship information among attributes can be fully explored and exploited, which improves the stability and effectiveness of the network. The experiments on three common data sets (PETA [25], RAP [26], PA-100K [16]) have verified the effectiveness of our method. Our contributions are summarized as follows.

- We propose an effective network CAFN to sufficiently explore the attribute correlations for resolving the problem of pedestrian attribute recognition.
- We further propose two modules: Cross-attribute Attention Module and Cross-feature Attention Module to automatically learn the relation matrix and fuse multiple matrices of attribute features, respectively.
- Extensive experiments on three common data sets demonstrate the effectiveness of the proposed CAFN.

We introduce the related work of pedestrian attribute recognition in Sect. 2, our proposed method and the experiments are introduced in Sects. 3 and 4, respectively. In Sect. 5, we summarize the content of the paper.

2 Related work

In this section, we present a brief review of approaches about pedestrian attribute recognition in Sect. 2.1, and then introduce studies of feature pyramid architecture and attention mechanisms that are closely related to this work in Sects. 2.2 and 2.3, respectively.

2.1 Pedestrian attribute recognition

In practical applications, it is necessary to identify a series of attributes for the task of pedestrian attribute recognition. It is an intuitive idea to let the network to learn and identify each attribute independently, but it will bring redundancy and inefficiency to the network. Moreover, the learning of one attribute may relies on or constrain the others. Process single attribute independently is prone to be influenced by the multi-label nature of the pedestrian attribute recognition task. Therefore, most research methods integrate recognition tasks of different attributes into one model. These approaches could be categorized into two groups: methods without considering attribute correlation and methods considering attribute correlation.

2.1.1 Methods without considering attribute correlation

Some methods pay more attention to how to extract more accurate features and do not explore the correlation information between attributes in the process of recognition. Li et al. [4] and Sudowe et al. [5] focus on the global feature and proposes multi-task learning algorithms. These networks are relatively simple in which all attribute features share network parameters, resulting in a poor effect. Some works [6–11, 13–15] put forward different strategies to extract or generate local features. Zhu et al. [6] divide the image into 15 overlapping patches to extract local features and the spatial location information of attributes is used to identify each patch; Joo et al. [7] extracts features from selected windows in the image, and then K-means is conducted to cluster these features, to train the attribute detector; Some works [10, 11] apply poselet [12] to obtain the key points of the pedestrian and transform these key points into part information to obtain local features. Some works [15–19] add attention mechanism to improve utilization of visual feature information. Some works [15, 19] put forward the application of CAM (class activation map) network [27] which is used as guidance to assess the importance of local features to different attributes and locate attributes from the global feature. Liu et al. [16] proposes a multi-directional attention mechanism module (MDA) to fuse the multi-layer features in the attention area according to the semantics of different layers. Each layer in MDA generates an attention graph, which not only maps back to the original layer features but also applies to adjacent layers, so as to mine multi-scale attention features. The view predictor is introduced in method [18] to estimate the weight of the view, to fuse the features of different views. In method [17], it is suggested that the attention graph should be learned by weak supervision, and the classification performance can be improved through guiding the network to pay more attention on the spatial part containing relevant information. In method [28], the human posture key points are generated and then be used as auxiliary information to obtain the regional position information of specific attributes. However, these models only add attention to the network structure or focus on specific attributes, without considering the dependence of attributes.

2.1.2 Methods considering attribute correlation

Other methods take into account the attribute correlation information to assist attribute recognition. Bourdev et al. [20] trains SVM classifiers at poselet [12] level, person level and context level. When they trained classifiers for each attribute at the context level, they applied the scores of all person-level attribute classifiers to make use of the correlations among attributes. However, these classifiers need to be trained separately. Wang et al. [21] manually divide the given image into several horizontal regions, extracts attribute features with LSTM which could be able to capture higher order correlations between attributes. Based on [21], Zhao et al. [22] divided the whole attribute list into several groups, and LSTM was used to simulate spatial and semantic correlation in attribute groups. Some methods [21, 22] are based on sequence prediction, which is easily affected by manual partition and attribute order. The network [23] is capable of predicting multiple attributes simultaneously, which applied Transformer [29] as an attention module to model the correlation between attributes and align the long attribute sequences. Zhao et al. [24] proposed two models: circular convolution model (RC) and cyclic attention model (RA). These two models focused on the correlation between different attribute groups and the spatial correlation within groups, respectively. The MTA-Net [30] is constructed based on the LSTM who employs the information of the next time step, and mines deeper relations between images and attributes. However, dividing attributes into different groups only considers the rough correlation between partial attributes and does not accurately quantify the correlation between all attribute pairs. An attribute relationship attention module is designed in [31] to capture the latent relations among different attributes. However, this method combines two learning tasks: coarse attribute localization and fine attribute recognition, which is too complex.

2.2 Feature pyramid architecture

To make better use of the correlation between attributes, it is necessary to obtain an accurate feature description of each attribute. However, in a deep convolution network, lower layers can capture small-scale attribute information and the deep network has a larger receptive field which can capture more abstract information. Owing to different and complementary concerns of different layers, it is meaningful to add feature fusion operation in the network. The feature pyramid structure integrates the different features of the low layers and high layers. First, multiple features of different scales are extracted from the bottom up, and then the feature is upsampled from the top down. The features of the same scale are fused in the horizontal direction. This idea has been reflected in many previous works, such as [32, 33].

2.3 Attention mechanism

The attention mechanism enables the model to focus on a small part of useful information from a large amount of input information while ignoring other information. At present, the mainstream attention mechanisms can be divided into the following three types: channel attention, spatial attention and self attention. Channel-attention based methods [34–36] could automatically obtain the importance of different channels in the feature, so as to strengthen the important features and suppress the non important features. Spatial-attention based methods [37] aims to improve the feature expression of key regions. These methods generate a mask and assign weight for each position, so as to enhance the specific target regions of interest and weaken the irrelevant background regions. The self-attention based methods [29, 38] reduce the dependence on external information and have a better ability of capturing the internal correlation of features. The self-attention mechanism can be viewed as a mapping operation from a query to a series of key-value pairs. In general, the input data could be set as a series of key-value data pairs. Giving a query of an element in the target data, the weight coefficient of each key can be obtained by calculating the similarity or correlation between the query and each key. After normalization by the softmax operation, the weight and corresponding value are weighted and summed to get the final attention value.

3 Methodology

In this section, we first present the overall architecture of the proposed network in Sect. 3.1. Then we give a detailed introduction to the Cross-attribute Attention Module and Cross-feature Attention Module in Sects. 3.2 and 3.3, respectively. Finally, we describe the loss function in Sect. 3.4.

3.1 Overall architecture

Figure 2 shows the overall pipeline of the proposed Cross Attribute and Feature Network (CAFN). Motivated by FPN [33], we construct CAFN with a three-layer feature pyramid structure. Specifically, we adopt inception-v2 as the backbone network and build feature pyramid architecture with three different levels: the inception_3b, inception_4d, and inception_5b. The feature pyramid architecture mainly contains three horizontal connection operations and two top-down connection operations. For horizontal connections, we use a 1×1 convolution kernel to reduce the channel dimension uniformity to 256. For top-down connections, we perform an up-sampling operation. Given the input pedestrian image as I, we denote the top-down features in the three layer as $\phi_j(I), j \in \{1, 2, 3\}$, respectively. The input images

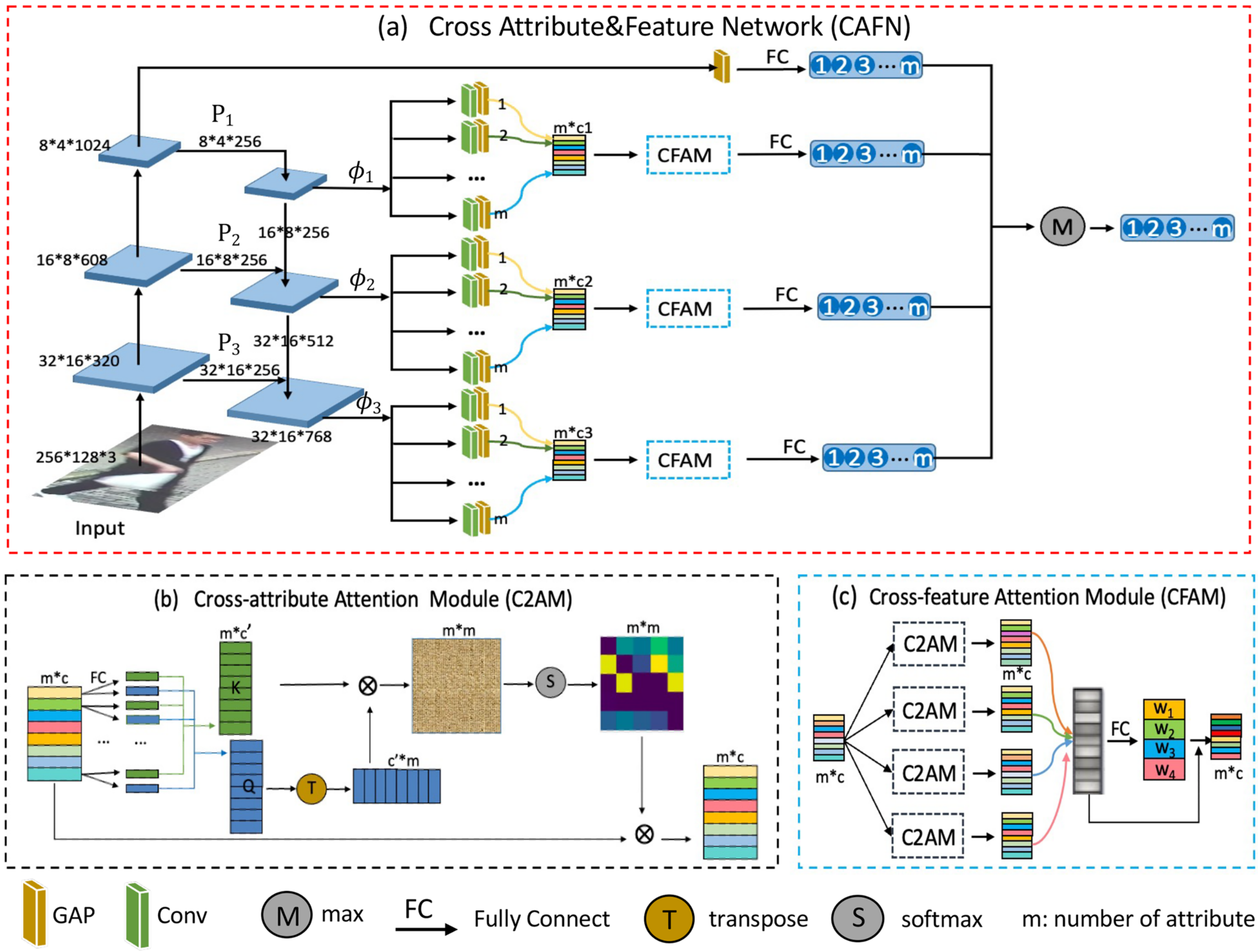


Fig. 2 Overview of the proposed Cross Attribute and Feature Network (CAFN). **a** is the overall pipeline of CAFN. **b** and **c** are the illustration of Cross-attribute Attention Module (C2AM) and Cross-feature Attention Module (CFAM), respectively

are resized to 256 × 128 and sent into the multilayer convolution network, and the feature maps of $32 \times 16, 16 \times 8$ and 8×4 are obtained successively. The number of channels is 320, 608 and 1024 respectively. In the process of building the feature pyramid, to reduce the network parameters, 256 channels are integrated and extracted from the three feature maps to obtain $\phi_j(I), j \in \{1, 2, 3\}$. Where $\phi_1(I)$ is obtained directly from $P_1(I)$, and the size is $8 \times 4 \times 256$. After an upsampling operation, $\phi_1(I)$ is concatenated with $P_2(I)$ and $\phi_2(I)$ is obtained with size of $16 \times 8 \times 512$. Similarly, after an upsampling operation, $\phi_2(I)$ is concatenated with $P_3(I)$ and $\phi_3(I)$ is obtained with size of $32 \times 16 \times 768$.

For the learning of correlation between attributes, we perform convolution and pooling operations on features $\phi_1(I)$, $\phi_2(I)$ and $\phi_3(I)$ to extract the features of each attribute, getting the initial rough feature matrix F. And then, use the Cross-attribute Attention Module (C2AM) to learn the relation matrix between attributes. According to the weight information in the relation matrix, each attribute feature in the feature matrix F will be fused with all other attribute features using different weights, so as to get the feature matrix $\tilde{F}$ which contain more rich semantics. In the Cross-feature Attention Module (CFAM), we convert the input feature matrix F to different $\tilde{F}_p$ by constructing h different C2AM modules, and perform weighted fusion on these $\tilde{F}_p$ ($p \in \{1, 2, \ldots, h\}$, h denotes the number of C2AM contained in CFAM). The output of CFAM is one feature matrix $\hat{F}$, which will be sent to fully connect layers to obtain the recognition results of attributes.

As illustrated in Fig. 2, four individual prediction vectors are obtained from one global branch and three pyramid layers. During inference, four prediction vectors are aggregated and the final prediction output value of each attribute uses the maximum value of the corresponding result in the four vectors.

3.2 Cross-attribute attention module

To model the relationships between any pair of attributes in the attribute set, we propose a Cross-attribute Attention Module(C2AM), as illustrated in Fig. 2b. We input each feature $\phi_j(I)$, $(j \in \{1, 2, 3\})$ into m different convolutional layers and global average pooling layers, respectively, and obtain features f_i, $(i \in \{1, 2, \dots, m\})$, which is corresponding to m different attributes. These f_i construct the attribute feature matrix F with size of $m \times c$ (c is the feature dimension of each attribute). To mine the attribute correlation information based on such features, we utilize self-attention mechanism to learn a relation matrix between attributes, which is denoted as R , with a shape of $m \times m$ (m is the number of attributes). Specifically, each f_i is input into different fully connect layers to obtain the corresponding vector k_i and vector q_i, respectively. In addition, the dimension of vector k_i and vector q_i is reduced to c'(c'=c/8). After concatenation of all the k_i and q_i, the matrix K and matrix Q are obtained. The matrix Q is transposed and multiplied by the matrix K, and the relationship matrix R of $m \times m$ size is obtained after softmax operation, formulated as

$$R = s(K \times Q^\top),\\ R \in \mathbb{R}^{m\times m}, K \in \mathbb{R}^{m\times c'}, Q \in \mathbb{R}^{m\times c'}. \tag{1}$$

where s refers to the softmax operation for each row vector of the matrix. Specifically, the ith row of matrix K is learned from the ith attribute feature in matrix F through the fully connect operation. As a result, the ith row of matrix K contains the information of the ith attribute. Therefore, is the matrix Q. Therefore, in the matrix of $m \times m$ size obtained by multiplying the K and $Q^\top$, the value of row i and column j is obtained by multiplying the ith attribute feature in K and the jth attribute feature in Q. After normalizing the matrix of $m \times m$ size with softmax operation, matrix R is obtained which can represent the relationship between attributes. Moreover, the value of the element in row i and column j of the relationship matrix R represents the correlation between the ith attribute and the jth attribute. Multiplied R with the attribute feature matrix F, the output feature $\tilde{F}$ of the module is obtained: formulated as:

$$\tilde{F} = R \times F,\\ \tilde{F} \in \mathbb{R}^{m\times c}, R \in \mathbb{R}^{m\times m}, F \in \mathbb{R}^{m\times c}. \tag{2}$$

In the process of this operation, for each f_i $(i \in \{1, 2, \dots, m\})$ in F, we consider the relationship between f_i and all other f_v, $(v \in \{1, 2, \dots, m\}$ and $v \neq i)$, and fuse them according to the corresponding value in the relationship matrix R, respectively. After this operation, we get the attribute feature matrix $\tilde{F}$ which is further refined.

3.3 Cross-feature attention module

To generate more accurate and robust attribute features, we propose a Cross-feature attention module(CFAM), as illustrated in Fig. 2c. For each feature $\phi_j(I)$, $(j \in \{1, 2, 3\})$, we can get an initial rough feature matrix F. In CFAM, we construct multiple C2AM modules. By setting the number of C2AM as h, we will get h different $\tilde{F}_p$, $(p \in \{1, 2, \dots, h\})$ after inputting F to each C2AM. Although the F input to multiple C2AM modules is the same, the parameters of fully connect layer in each C2AM module are different. Therefore, the obtained matrix K and Q are different in different C2AM, and different relationship matrix R will be obtained. Finally, different $\tilde{F}_p$ are obtained.

In Fig. 2c, we set h as 4. In addition, CFAM can learn different weights for different $\tilde{F}_p$ and fuse them more efficiently. After concatenation of each input feature $\tilde{F}_p$, we input them to two fully connect layers and set the output dimension of the last fully connect layer as h, so as to map the feature information to weight information. In addition, the last fully connect layer will output h weight parameters w_p, $(p \in \{1, 2, \dots, h\})$. Through the continuous learning and updating of the parameters in the fully connect layers, the network can directionally learn h parameters to represent the importance of h attribute features $\tilde{F}_p$ according to the fusion information of h attribute features. The w_p corresponding to each $\tilde{F}_p$ represents the importance of $\tilde{F}_p$ to the feature matrix $\hat{F}$. We multiply the weight w_p by the corresponding $\tilde{F}_p$ and sum to get the output features of Cross-feature Attention Module, which are denoted as $\hat{F}$. Each row of feature matrix $\hat{F}$ represents a feature that belongs to an attribute after feature fusion which is formulated as

$$\hat{F} = \sum_{p=1}^{h} w_p \times \tilde{F}_p,\\ \hat{F} \in \mathbb{R}^{m\times n}, w_p \in \mathbb{R}, \tilde{F}_p \in \mathbb{R}^{m\times n}. \tag{3}$$

The CFAM makes use of the interior information of features and makes it easier for the network to learn a more reasonable combination of features. Moreover, CFAM contains multiple C2AM, which can alleviate the deviation caused by a single relation matrix.

3.4 Loss function

For the four individual prediction vectors obtained in the network, each of them is compared to the ground-truth labels and is involved in the calculation of loss function to update the gradient of each branch in the network more reasonably. Given the input pedestrian image as I, and the corresponding label is represented as $y = [y_1, y_2, \dots, y_m]^T$, where m and y_i

denote the number of pedestrian attributes and the label of the ith attribute, respectively. In particular, when I has the ith attribute, $y_i = 1$, otherwise $y_i = 0, (i \in \{1, 2, \dots, m\})$. We denote the predictions generated from the jth branch on ith attribute as $\tilde{y}_i^j$, $j \in \{1, 2, 3, 4\}$. Then the final output of the network $\tilde{y} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y_m}]^T$, where $\tilde{y}_i = \max\,(\tilde{y}_i^1, \tilde{y}_i^2, \tilde{y}_i^3, \tilde{y}_i^4)$. This paper uses the weighted binary cross entropy loss function in [4], which is formulated as

$$L_j(\tilde{y}_j, y) = -\frac{1}{m}\sum_{i=1}^{m} \gamma^i (y^i \log(\sigma(\tilde{y}_j{}^i)) \\ + (1 - y^i) \log(1 - \sigma(\tilde{y}_j{}^i))), \tag{4}$$

where $\gamma^i = e^{-a_i}$, and a_i is the loss weight value of the ith attribute; a_i is the prior class distribution of the ith attribute; σ is the sigmoid activation function; j is the loss function of jth branch, $j \in \{1, 2, 3, 4\}$. The final loss function consists of the sum of the loss functions of backbone and three pyramid layers:

$$L = \sum_{j=1}^{4} L_j(\tilde{y}_j, y). \tag{5}$$

4 Experiments

4.1 Settings

4.1.1 Data set

To verify the effectiveness of the proposed model, we conduct experiments on three public data sets, PETA, RAP and PA-100K, respectively.

The **PETA data set** [25] contains 8705 pedestrians with a total of 19,000 images (resolution span range from 17 × 39 to 169 × 365). Each pedestrian is labeled with 61 binary and four multi-class attributes. However, some attributes will not be used according to the established protocol. We only use 35 attributes with a positive label ratio higher than 5%. In addition, using the same method as [18] to divide the PETA data set, the number of images in the training, verification, and test sets are 9500, 1900, and 7600, respectively.

The **RAP data set** [26] is collected from the real indoor environment. A total of 26 cameras were used to collect images of the surveillance scene, including 41,585 samples, with a resolution ranging from 36 × 92 to 344 × 554. Specifically, there are 33,268 training images and 8317 test images. Each image sample contains 72 fine-grained attributes (69 binary attributes and 3 multi-class attributes). However, we only make use of 51 attributes whose positive label ratio is higher than 1%.

The **PA-100K data set** [16] is composed of pictures captured from 598 real outdoor surveillance cameras. There are 100,000 samples in total, and the resolution of each sample image is between 50 × 100 and 758 × 454. The PA-100K data set is by far the largest pedestrian attribute recognition data set. The whole data set is randomly divided into the training set, validation set, and test set at the ratio of 8:1:1. Each image in the data set is labeled with 26 attributes.

4.1.2 Implementation details

In the entire process of the experiment, we use the adam [39] optimizer with initial learning rate of 1×10^{-4}, momentum of 0.9, and weight decay of 5×10^{-4}. The batch size is set to 8 and the initial learning rate equals 0.0001 which decays by 0.1 every 10 epochs. The training samples are augmented by random horizontal flipping. In both training and testing phases, input images are resized to 256 × 128. We use Inception-v2 as the backbone, whose parameters are initialized with the corresponding model pre-trained on ImageNet [40].

4.1.3 Evaluation metrics

For the evaluation of PETA, RAP, and PA-100K, we rely on two types of indicators: label-based and example-based metrics.

For the label-based evaluation, we adopt the mean accuracy (mA). For each attribute, we calculate the accuracy of all samples, regardless of positive or negative. Then we calculate the average value of all attributes to get mA. Therefore, mA is not affected by the class imbalance, so the error caused by less and more frequent tag values is punished equally strongly. The evaluation criterion mA can be calculated by the following formula:

$$mA = \frac{1}{2N}\sum_{i=1}^{m}\left(\frac{\mathrm{TP}_i}{P_i} + \frac{\mathrm{TN}_i}{N_i}\right), \tag{6}$$

where m is the number of attributes and N is the number of samples. TP_i and TN_i are the numbers of positive and negative examples that are correctly predicted for the ith attribute, respectively. P_i and N_i are the numbers of positive and negative examples of the ith attribute, respectively.

To explain the consistent attribute predictions in each person image, we further use four widely used metrics, including: accuracy, precision, recall rate and F1 value, which are defined as follows:

$$\text{accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\,Y_i \cap f(x_i)\,|}{|\,Y_i \cup f(x_i)\,|}, \tag{7}$$

$$\text{precision} = \frac{1}{2N}\sum_{i=1}^{N}\frac{|\,Y_i \cap f(x_i)\,|}{|\,f(x_i)\,|}, \tag{8}$$

$$\text{recall} = \frac{1}{2N}\sum_{i=1}^{N}\frac{|\,Y_i \cap f(x_i)\,|}{|\,Y_i\,|}, \tag{9}$$

$$F1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \tag{10}$$

where Y_i means the ground truth positive labels of the ith example, and $f(x_i)$ returns the predicted positive labels for ith example. In addition, | . | is the set cardinality.

4.2 Comparison with state-of-the-arts

We make a comparison of the performance of our proposed network against several state-of-the-art networks, such as GAM [41], PGDM [11], GRL [22], RCRA [24], MT-CAS [42] , DTM [28], MTMS [31] and so on. The experimental results on PETA data set and RAP data set are shown in Table 1, and the result of PA-100K is shown in Table 2. As for evaluation criteria, mA, accuracy, precision, recall, and F1 are all listed. Besides, we add the number of parameters (#P) and complexity (GFLOPS) for the RAP data set. By observing the data in these tables, we can be informed of the advantages of our model clearly. On the PETA data set, both accuracy and F1 score reach the optimal level, which are 78.81% and 86.57%, respectively. Although LG-net [15] is the best in precision and F1 value and DTM [28] is better in recall value on RAP data set, our model is superior to these two models in the number of model parameters and computational complexity. MTMS [31] has the best result on mA, but the results on accuracy, precision and F1 is far worse than our model. Moreover, on the PA-100K data set, our model has the best performance in mA and recall and outperforms the sub-optimal model MT-CAS [42] by 2.56% and 3.48%, respectively. These results show the necessity and effectiveness of mining attribute correlation information.

Table 2 Quantitative comparisons on PA-100K data set

Methods	PA-100K data set				
	mA	Accu	Prec	Recall	F1
DeepMar [4]	72.7	70.39	82.24	80.42	81.32
PGDM [11]	74.95	73.08	84.36	82.24	83.29
LG-Net [15]	76.96	75.55	86.99	83.17	85.04
HP-net [16]	74.21	72.19	82.97	82.09	82.53
VeSPA [18]	76.32	73	84.99	81.49	83.2
MT-CAS [42]	77.2	**78.09**	**88.46**	84.86	**86.62**
Ours	**79.76**	77.12	86.18	**88.34**	86.14

Best results are in bold

Table 1 Quantitative comparisons against previous methods on PETA and RAP data sets

Methods	Data set											
	PETA					RAP						
	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	GFLOPS	#params
ACN [5]	81.15	73.66	84.06	81.26	82.64	69.66	62.61	80.12	72.26	75.98	–	–
DeepMar [4]	82.89	75.07	83.68	83.14	83.41	73.79	62.06	74.92	76.21	75.56	0.72	58.5M
GAM [41]	–	–	–	–	–	79.73	**83.97**	76.96	78.72	77.83	–	–
PGDM [11]	82.97	78.08	86.86	84.68	85.76	74.31	64.57	78.86	75.9	77.35	1	87.2M
LG-Net [15]	–	–	–	–	–	78.68	68	**80.36**	79.82	**80.09**	>4	>20M
JRL [21]	85.67	–	86.03	85.34	85.42	77.81	-	78.11	78.98	78.58	–	–
GRL [22]	**86.7**	–	84.34	**88.82**	86.51	81.2	–	77.7	80.9	79.29	>10	>50M
RA [24]	86.11	–	86.03	88.51	86.56	81.16	-	79.45	79.23	79.34	–	–
HP-net [16]	81.77	76.13	84.92	83.24	84.07	76.12	65.39	77.33	78.79	78.05	–	–
VeSPA [18]	83.45	77.73	86.18	84.81	85.49	77.7	67.35	79.51	79.67	79.59	>3	17.0M
DIAA [17]	84.59	78.56	86.79	86.12	86.46	–	–	–	–	–	–	–
MT-CAS [42]	83.17	78.78	**87.49**	85.35	86.41	–	–	–	–	–	–	–
MTA-Net [30]	84.62	78.80	85.67	86.42	86.04	77.62	67.17	79.72	78.44	79.07	-	-
DTM+AWK [28]	85.79	78.63	85.65	87.17	86.11	82.04	67.42	75.87	**84.16**	79.80	4.09	23.7
MTMS [31]	86.23	77.21	84.52	87.22	85.85	**82.45**	49.10	55.00	80.44	65.33	-	-
Ours	85.97	**78.81**	85.68	88.08	**86.57**	77.9	66.87	78.65	80.38	79.37	2.31	24.7M

Best results are in bold

4.3 Ablation studies

To analyze the effectiveness of each key component of the CAFN network and the influence of other factors, we conduct ablation experiments on the PETA data set.

4.3.1 Effectiveness of cross-attribute attention module

We first remove the CFAM module in the third pyramid layer and add a C2AM module to make the structure of the three layers consistent. Then we replace the original C2AM with two alternatives. One is merely removing the self-attention mechanism in C2AM, which means the attribute feature matrix F will be sent to the fully connected layer instead of the attribute feature matrix $\tilde{F}$. The other one is directly removing C2AM, which means adding corresponding fully connected layers directly on each feature $\phi_i(I), i \in \{1, 2, 3\}$ to make recognition. We refer to them as $C2AM-$ and $C2AM--$. The results are shown in the first three rows of Table 3. For $C2AM--$, the mA and F1 score are 84.97% and 85.82%, respectively. Compared with $C2AM--$, the mA and $F1$ score of $C2AM-$ are increased by 0.05% and 0.13%, which indicates that convolution and global average pooling operations are important and can further refine the attribute features. Compared with $C2AM-$, the mA and $F1$ score of $C2AM$ are increased by 0.31% and 0.14%, which reveals that the self-attention mechanism can improve the model discrimination ability by learning attribute correlation. In addition, the increased mA and $F1$ scores show the success of C2AM. The corresponding result of attribute-wise mA for $C2AM-$ and $C2AM$ is shown in Fig. 3. Compared with $C2AM-$, $C2AM$ achieves improvement in many attributes.

Table 3 Ablation study for C2AM and CFAM on PETA data set

Method	mA	F1
$C2AM--$	84.97	85.82
$C2AM-$	85.02	85.95
$C2AM$	85.33	86.09
$CFAM^1$	85.33	86.09
$CFAM^2$	85.52	86.32
$CFAM^4$	85.97	86.51
$CFAM^8$	85.77	86.41
Full Module($CFAM^4$)	85.97	86.51

4.3.2 Effectiveness of cross-feature attention module

To verify the effectiveness of the CFAM and explore the optimal structure of the CFAM, we set up four groups of experiments. Except for the different number of C2AM

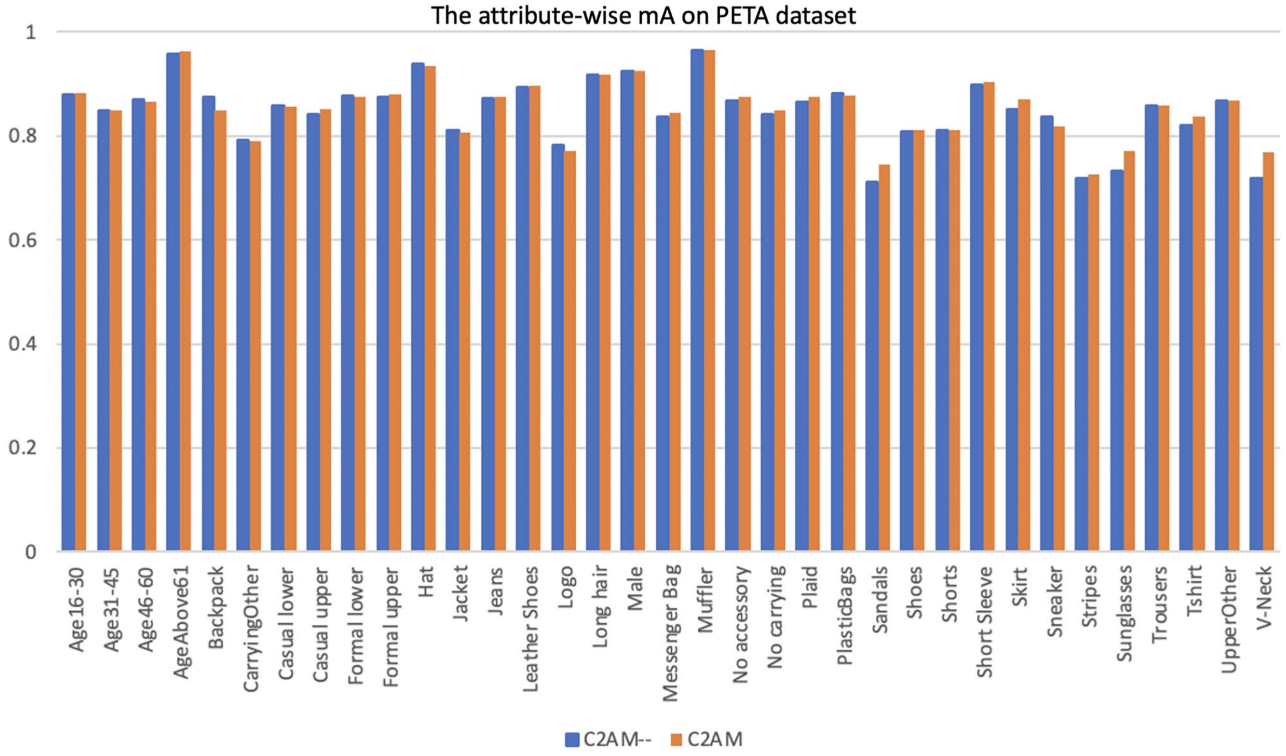


Fig. 3 Attributewise mA of $C2AM-$ and $C2AM--$

modules included in CFAM, the other conditions of these four groups of experiments are the same. We set the number of C2AM modules to 1, 2, 4, 8, respectively, and refer to them as $CFAM^1$, $CFAM^2$, $CFAM^4$ and $CFAM^8$, respectively. It should be noted that setting the number of C2AM modules to 1 is equivalent to removing the CFAM module, which means $CFAM^1$ is equal to *C2AM*. Results in Table 3 reveal that the CFAM module is instrumental in further improving the ability of the model to distinguish attributes. In addition, when the number of C2AM modules is 4 in CFAM, the experimental result is the best. Compared with $CFAM^1$, the mA and F1 score of $CFAM^4$ are increased by 0.64% and 0.32%, respectively. However, when we add the number of CAM modules from 4 to 8, the mA and F1 score are reduced by 0.2% and 0.1%, respectively. The reduction of performance suggests that when C2AM exceeds a certain number, it may cause redundancy of the model and bring negative effects. The corresponding result of attributewise mA for $CFAM^1$ and $CFAM^4$ is shown in Fig. 4. Compared with $CFAM^1$, $CFAM^4$ achieves improvement on a number of attributes.

In addition, three examples of different visual angles from the PETA data set are given in Fig. 5 for qualitative analysis. As we can see, the proposed *C2AM* and $CFAM^4$ can successfully recognize age, gender, clothing, footwear, and other attributes. In the first example, the pedestrian's clothing is unfavorable to the judgment of gender, but the attribute of long hair assists the identification of gender attribute. In the second example, the lower part of the pedestrian's clothing is partially occluded, but the upper part of the clothing attributes assists the correct recognition of the lower part of the clothing attributes. A failure case is also provided in the third example. Because of the correlation between short sleeves and shorts, *C2AM* mistakenly identified trousers as shorts. However, the wrong prediction is well corrected in $CFAM^4$.

4.3.3 Visualization of the relation matrix among attributes

The relation matrix among attributes is the key learning content of the network proposed in this paper. We proposed to use the correlation among attributes to assist the detection and recognition of each attribute. To obtain the correlation information, we let the network lean the relation matrix among attributes to quantify each pair of attributes in the attribute set. This section visualizes the relation matrix learned in the network after convergence, as shown in Fig. 6. The brighter the color, the greater the correlation. It can be seen that the relation matrix learns more abstract information, such as the obvious correlation between male and long hair in Fig. 6a. The network CAFN will learn multiple different relation matrices at the same time, and work together

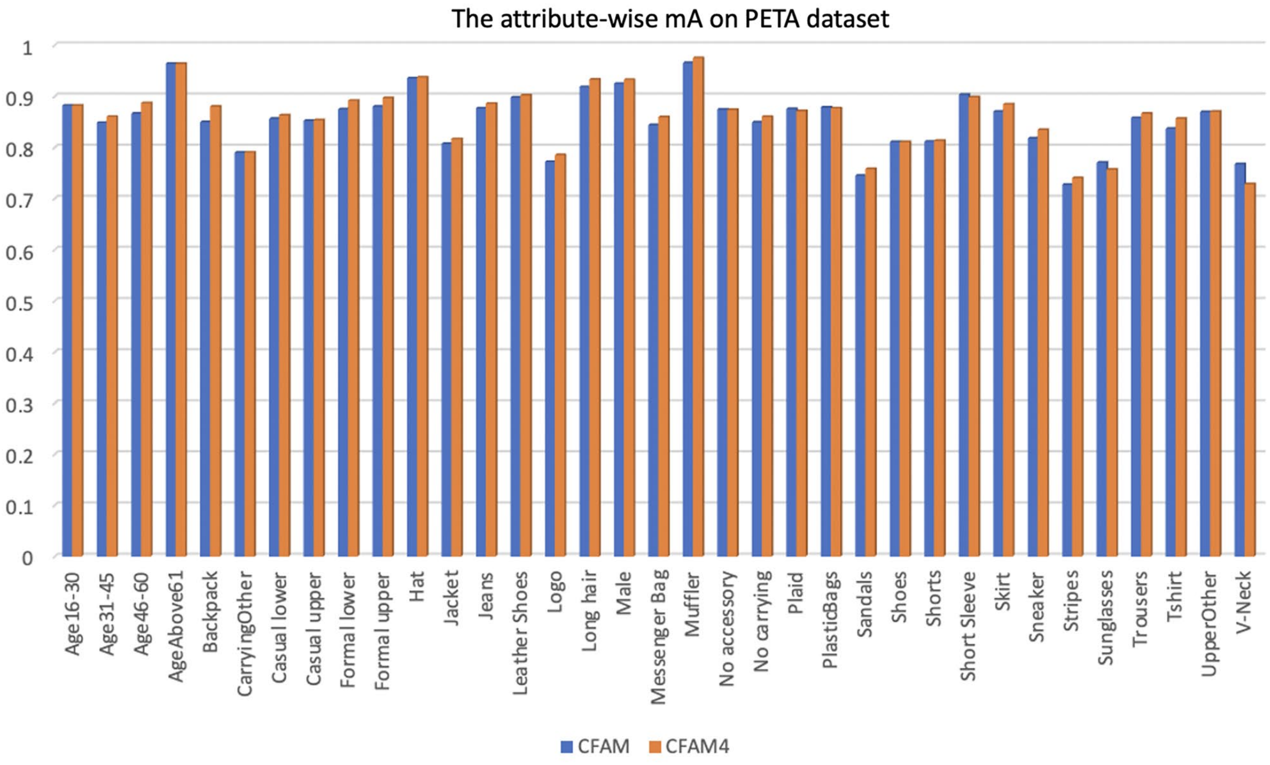
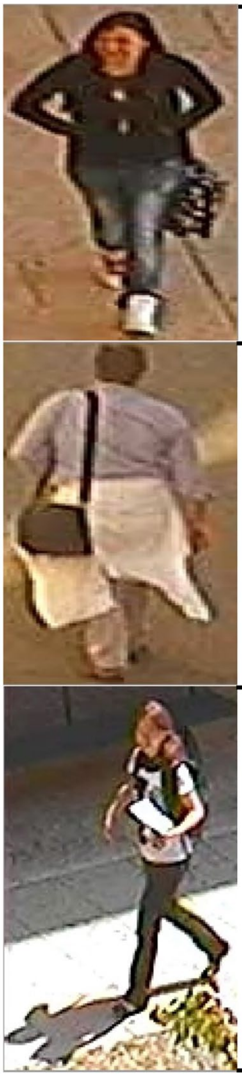


Fig. 4 Attributewise mA of $CFAM^1$ and $CFAM^4$

Fig. 5 Qualitative results from PETA data set of *C2AM* and *CFAM*^{4}. The top two indicate the effective examples. The last denotes the failure case in *C2AM* and the wrong prediction is well corrected in *CFAM*^{4}



Ground Truth	C2AM	CFAM⁴
Age16-30	✓	✓
Casual lower	✓	✓
Casual upper	✓	✓
Jacket	✓	✓
Jeans	✓	✓
Long hair	✓	✓
Messenger Bag	✓	✓
No accessory	✓	✓
Shoes	✓	✓
Age31-45	✓	✓
Formal lower	✓	✓
Formal upper	✓	✓
Male	✓	✓
Messenger Bag	✓	✓
No accessory	✓	✓
Shoes	✓	✓
Trousers	✓	✓
	CarryingOther	CarryingOther
Age16-30	✓	✓
Backpack	✓	✓
Casual lower	✓	✓
Casual upper	✓	✓
Long hair	✓	✓
Shoes	✓	✓
Short Sleeve	✓	✓
Trousers		✓
Tshirt	✓	✓
	CarryingOther	CarryingOther
	Shorts	

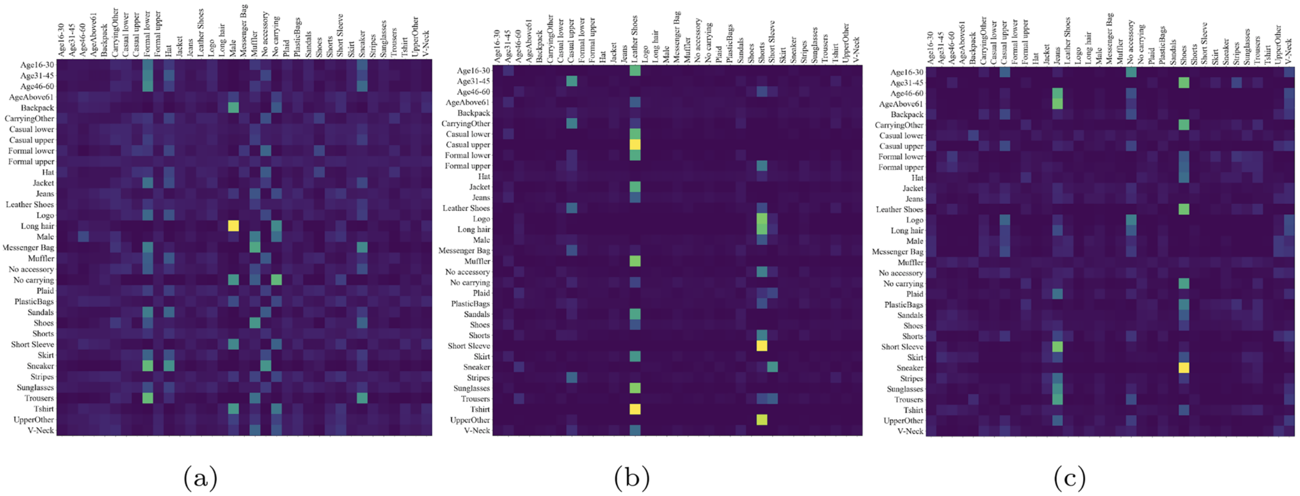


Fig. 6 Visualization of the relation matrix for three examples. The brighter the color, the greater the correlation

for the final attribute recognition. Another relation matrix in Fig. 6b highlights the correlation between short sleeves and shorts, while another relation matrix in Fig. 6c highlights the correlation between sneaker and shoes.

5 Conclusion

In this paper, considering how to exploit the correlations between any pair of attributes, we presented a novel architecture CAFN for pedestrian attribute recognition. It contains two essential modules: Cross-attribute Attention Module and Cross-feature Attention Module. As a result of the cooperation between the two modules, the performance of CAFN is promoted. We have carried out experiments on three public data sets (PETA, RAP, PA-100K) and achieved convincing results. The experimental results show that the network CAFN outperforms the most existing methods. Furthermore, extensive experiments verify the effectiveness of the two key modules in the network. In the future, it is meaningful to focus on how to explore and mine the correlation between images and attributes in a multi-modal perspective, which can further improve the model's ability to distinguish different attributes.

Acknowledgements This work was supported by the Fundamental Research Funds for the Central Universities (No. 2020JBM403), the National Natural Science Foundation of China (No. 62072027, No. 61872032), the Beijing Natural Science Foundation (Grants No. 4202057, No. 4202058, No. 4202060).

References

1. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3800–3808 (2017)
2. Tay, C.-P., Roy, S., Yap, K.-H.: Aanet: Attribute attention network for person re-identifications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7134–7143 (2019)
3. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., Yang, Y.: Improving person re-identification by attribute and identity learning. Pattern Recogn. **95**, 151–161 (2019)
4. Li, D., Chen, X., Huang, K.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 111–115. IEEE (2015)
5. Sudowe, P., Spitzer, H., Leibe, B.: Person attribute recognition with a jointly-trained holistic cnn model. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 87–95 (2015)
6. Zhu, J., Liao, S., Yi, D., Lei, Z., Li, S.Z.: Multi-label cnn based pedestrian attribute learning for soft biometrics. In: 2015 International Conference on Biometrics (ICB), pp. 535–540. IEEE (2015)
7. Joo, J., Wang, S., Zhu, S.-C.: Human attribute recognition by rich appearance dictionary. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 721–728 (2013)
8. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2470–2478 (2015)
9. Diba, A., Pazandeh, A.M., Pirsiavash, H., Van Gool, L.: Deepcamp: Deep convolutional action & attribute mid-level patterns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3557–3565 (2016)
10. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1637–1644 (2014)
11. Li, D., Chen, X., Zhang, Z., Huang, K.: Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018)
12. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1365–1372. IEEE (2009)
13. Yang, L., Zhu, L., Wei, Y., Liang, S., Tan, P.: Attribute recognition from adaptive parts. arXiv preprint arXiv:1607.01437 (2016)
14. Li, Y., Huang, C., Loy, C.C., Tang, X.: Human attribute recognition by deep hierarchical contexts. In: European Conference on Computer Vision, pp. 684–700. Springer (2016)
15. Liu, P., Liu, X., Yan, J., Shao, J.: Localization guided learning for pedestrian attribute recognition. arXiv preprint arXiv:1808.09102 (2018)
16. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 350–359 (2017)
17. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 680–697 (2018)
18. Sarfraz, M.S., Schumann, A., Wang, Y., Stiefelhagen, R.: Deep view-sensitive pedestrian attribute inference in an end-to-end model. arXiv preprint arXiv:1707.06089 (2017)
19. Guo, H., Fan, X., Wang, S.: Human attribute recognition by refining attention heat map. Pattern Recogn. Lett. **94**, 38–45 (2017)
20. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: 2011 International Conference on Computer Vision, pp. 1543–1550. IEEE (2011)
21. Wang, J., Zhu, X., Gong, S., Li, W.: Attribute recognition by joint recurrent learning of context and correlation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 531–540 (2017)
22. Zhao, X., Sang, L., Ding, G., Guo, Y., Jin, X.: Grouping attribute recognition for pedestrian with joint recurrent learning. In: IJCAI, pp. 3177–3183 (2018)
23. Liu, H., Wu, J., Jiang, J., Qi, M., Ren, B.: Sequence-based person attribute recognition with joint ctc-attention model. arXiv preprint arXiv:1811.08115 (2018)
24. Zhao, X., Sang, L., Ding, G., Han, J., Di, N., Yan, C.: Recurrent attention model for pedestrian attribute recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9275–9282 (2019)
25. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 789–792 (2014)
26. Li, D., Zhang, Z., Chen, X., Ling, H., Huang, K.: A richly annotated dataset for pedestrian attribute recognition. arXiv preprint arXiv:1603.07054 (2016)
27. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database (2014)

28. Zhang, J., Ren, P., Li, J.: Deep template matching for pedestrian attribute recognition with the auxiliary supervision of attribute-wise keypoints. arXiv preprint arXiv:2011.06798 (2020)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
30. Ji, Z., Hu, Z., He, E., Han, J., Pang, Y.: Pedestrian attribute recognition based on multiple time steps attention. Pattern Recogn. Lett. **138**, 170–176 (2020)
31. Gao, L., Huang, D., Guo, Y., Wang, Y.: Pedestrian attribute recognition via hierarchical multi-task learning and relationship attention. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1340–1348 (2019)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241. Springer (2015)
33. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
34. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
35. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519 (2019)
36. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: efficient channel attention for deep convolutional neural networks, 2020 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2020)
37. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
38. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
39. KingaD, A.: A methodforstochasticoptimization. Anon. InternationalConferenceon Learning Representations. SanDego: ICLR (2015)
40. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
41. Fabbri, M., Calderara, S., Cucchiara, R.: Generative adversarial models for people attribute recognition in surveillance. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
42. Zeng, H., Ai, H., Zhuang, Z., Chen, L.: Multi-task learning via co-attentive sharing for pedestrian attribute recognition. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.