Computer Science and Information Systems 20(2):793–812 https://doi.org/10.2298/CSIS220815016F

Pedestrian Attribute Recognition Based on Dual Self-attention Mechanism

Zhongkui Fan[1] and Ye-peng Guan[1, 2]

[1] School of Communication and Information Engineering, Shanghai University,
200444 Shanghai, China
{fanzkui, ypguan}@shu.edu.cn
[2] Key Laboratory of Advanced Displays and System Application, Ministry of Education,
200444 shanghai, China

Abstract. Recognizing pedestrian attributes has recently obtained increasing attention due to its great potential in person re-identification, recommendation system, and other applications. Existing methods have achieved good results, but these methods do not fully utilize region information and the correlation between attributes. This paper aims at proposing a robust pedestrian attribute recognition framework. Specifically, we first propose an end-to-end framework for attribute recognition. Secondly, spatial and semantic self-attention mechanism is used for key points localization and bounding boxes generation. Finally, a hierarchical recognition strategy is proposed, the whole region is used for the global attribute recognition, and the relevant regions are used for the local attribute recognition. Experimental results on two pedestrian attribute datasets PETA and RAP show that the mean recognition accuracy reaches 84.63% and 82.70%. The heatmap analysis shows that our method can effectively improve the spatial and the semantic correlation between attributes. Compared with existing methods, it can achieve better recognition effect.

Keywords: pedestrian attribute recognition; spatial self-attention; semantic self-attention; deep learning.

1. Introduction

Visual recognition of pedestrian attributes has recently drawn a large amount of research attention due to its great potential applications in computer vision. Pedestrian attributes are defined as semantic mid-level descriptions of people, such as gender, age, hairstyle, body fat or thin, clothing style, and accessories. The research has achieved much success in many fields such as image retrieval [1,2], object recognition [3,4], and person re-identification [5]. It has also shown great application prospects in smart video surveillance and video-based business intelligence. At present, pedestrian attributes recognition is a great challenge, the main difficulties are: (1) poor image quality, low resolution, occlusion, motion blur, etc.; (2) uncontrollable interference, such as illumination and camera viewing angle, which aggravate the difficulties of feature recognition. (3) some pedestrian attribute recognition tasks require local fine-grained information, such as “glasses”.

There are some pioneering works on attribute recognition in multiple application scenarios. Layne et al. [6] first using Support Vector Model(SVM) to recognize attributes (e.g. “gender”, “backpack”) to assist pedestrian re-identification. To solve the attribute recognition problem in mixed scenarios, Zhu et al. [7] use boosting algorithm to recognize attributes. Deng et al. [8] construct the pedestrian attribute database (PETA) and utilize SVM and Markov Random Field to recognize attributes. However, these methods [9,10,11] use handcraft features, which cannot recognize the attributes effectively. Due to the outstanding performance of CNN in computer vision tasks, some researchers have tried to use it for pedestrian attribute recognition. [12] proposed CNN-based methods DeepSAR and DeepMAR for pedestrian attribute recognition to achieve better results. [13] proposed segment the pedestrian into multiple parts, and then recognize the attributes of each part. However, these simple methods cannot accurately segment pedestrians and require pre-processing of images. In addition, the correlation among attributes is ignored, which is very important to attributes recognition. For example, longhair feature has a higher probability for women than men, so the hair length could help to recognize the gender. Siddiquie et al. Zhang et al. [14] propose a pose aligned neural networks to recognize pedestrian attributes (e.g. age, gender) on images in multiple scenarios. The method poselets[15] decompose the image into a set of parts, which provide a robust distributed representation of a person, attributes can be inferred without explicitly localizing body parts. Although Pedestrian attributes recognition has been studied for years, it is still a quite challenge in real application environments. Pedestrian attribute recognition includes both local fine-grained and global feature recognition. In low-resolution images, some local attributes (such as glasses) occupy a small area, and the attributes need to be identified through fine-grained features, some abstract attributes (such as gender) need to be judged by overall features. At the same time, some attributes are spatial correlation, and some attributes are semantical correlation (such as gender attributes and skirt attributes), it can be used to improve the recognition effect between attributes.

To overcome the shortcomings of current pedestrian attribute recognition, this paper designs a dual self-attention convolution neural network (CNN) is proposed. In the work, the spatial and semantic self-attention are used for locating key regions, and the key regions are paid more attention to pedestrian attributes recognized. Secondly, using the relationship between attributes to improve the recognition accuracy. The proposed methods obtain state-of-the-art results on datasets PETA and RAP.

In this paper, there are three contributions.

(1) An end-to-end pedestrian attribute recognition framework is proposed. feature extraction, key points location, and pedestrian attributes recognition form an end-to-end network.

(2) The dual attention mechanism is proposed. Spatial self-attention is used for finding out pedestrian attributes from regions, and semantic self-attention is used to obtain the constraints among attributes.

(3) A hierarchical recognition strategy is proposed. We classify the pedestrian attributes as global (such as gender and age) and local (such as hairstyle and has glass), the whole regions are used for the global attribute recognition, and the partial regions are used for the local attribute recognition.

The organization of the rest of this paper is as follows. In Section II, we introduce the research and development of pedestrian attribute recognition. In Section III, the pedestrian attribute recognition framework is introduced in particular. In this section, we divide the methods into several parts and discuss detailed process in each part respectively. The experiments are carried on the benchmark dataset and the experimental results are analysed in Section IV. In Section VI, we give a conclusion.

2. Related Work

Pedestrian attribute recognition is important to research in computer vision. Given a person's image, pedestrian attribute recognition aims to predict a group of attributes to describe the characteristics of the person from a pre-defined attribute list. In the past decade, researchers have done a lot of research on pedestrian attribute recognition, and many methods have achieved good results, which can be divided into two groups: hand-crafted features and deep learning-based methods.

2.1. Pedestrian Attribute Recognition with Hand-Crafted Features

Pedestrian attributes were first used for human recognition by [16], the person image is divided into multiple regions, and each region is associated with a classifier based on Haar-like features. Then, the attribute information is used to retrieve surveillance video streams. The approach proposed by [10] extracts a 2784 dimensional low-level color and texture feature vector for each image and trains an SVM for each attribute. The attributes are further used as a mid-level representation to help person re-identification. [17] introduced the pedestrian attribute database APiS. Their method determines the upper and lower body regions according to the average image and extracts color and gradient histogram features (HSV, MB-LBP, HOG) in these two regions, then, an Adaboost classifier is trained to recognize attributes. The drawback of these approaches is that accuracy is low and the correlation between attributes is not considered. To overcome this limitation, [18] proposed an interaction model, based on Adaboost approach, learning an attribute interaction regressor. The final prediction is a weighted combination of the independent score and the interaction score. [19] constructed the pedestrian attribute dataset "PETA", which uses a Markov Random Field (MRF) to predict the relation between attributes, attributes are recognized by exploiting the context of neighbouring images on the MRF-based graph. [20] uses a multilabel Multi-layer perceptron to classify all attributes at the same time.

2.2. Pedestrian Attribute Recognition with Deep Learning

In recent years, deep learning has achieved success in automatic feature extraction using multi-layer nonlinear transformation, methods based on Convolutional Neural Network (CNN) models have been proposed for pedestrian recognition.[12] fine-tuned the CaffeNet trained on ImageNet to perform single and multiple attribute recognition. Sudowe et al. [21] proposed ACN(Attributes Convolutional Net), which uses an Alexnet network to extract features and set a classifier for each attribute to realize attribute recognition. Abdulnabi[22] proposed Multi-Task Convolutional Neural Network (MTCNN) , which shares feature pools each task corresponds to an attribute recognition. The above methods only extract the overall features of pedestrians, and pedestrian attribute recognition can be regarded as a fine-grained multi-label classification task. Highlighting local attribute features can greatly improve the recognition effect. To enhance the significance of local features, the existing methods extract local features by segmenting pedestrian images. For example, [18] proposed to divide the pedestrian images into 15 overlapping parts where each part connects to several CNN pipelines with several convolution and pooling layers. Some methods extract the local features of trunk position and use the correlation between local features for attribute recognition. In method [23], the pedestrian image is horizontally cut into multiple regions, and the features of each region are encoded and decoded, to make the local features more prominent. Another way is to use the attention mechanism to improve the accuracy of attribute recognition by assigning feature weights. Typical methods include the Spatial Regularization Network (SRN)[24], which applies the attention mechanism to the pedestrian attribute recognition for the first time, the Hydraplus-Net [25] proposes a multi-directional attention mechanism, and JVRKD(Join Visual-semantic Reasoning and Knowledge Distillation) [26], which combines local feature methods and attention mechanism methods.

2.3. Pedestrian Attribute Recognition with self-attention

In convolutional networks for image recognition, the layers of the network perform two functions. The first is feature aggregation, which the convolution operation performs by combining features from all locations tapped by the kernel. The second function is feature transformation, which is performed by successive linear mappings and nonlinear scalar functions: these successive mappings and nonlinear operations shatter the feature space and give rise to complex piecewise mappings. In computer vision tasks, the correlation between features can effectively improve accuracy. In order to obtain features outside the local area, it is necessary to increase the receptive field of deep neurons through the stacking of convolutional layers, which will increase the complexity of the convolutional neural network. Inspired by the idea of a No-local mean de-noising filtering algorithm [27], Wang et al. [28] proposed the No-Local network for video classification. In the literature [29], combining Transformer and No-Local, a self-attention mechanism was proposed to solve the problem of non-local feature dependence. In the task of image classification and pedestrian re-recognition, the literature [30, 31] used the self-attention mechanism to achieve better results. However, the method in the [27- 30] only uses the spatial correlation of features and does not use the correlation between channels. Literature [31] proves that semantic information has a strong correlation with channel features. [32-34] combine spatial attention and channel attention for image classification.

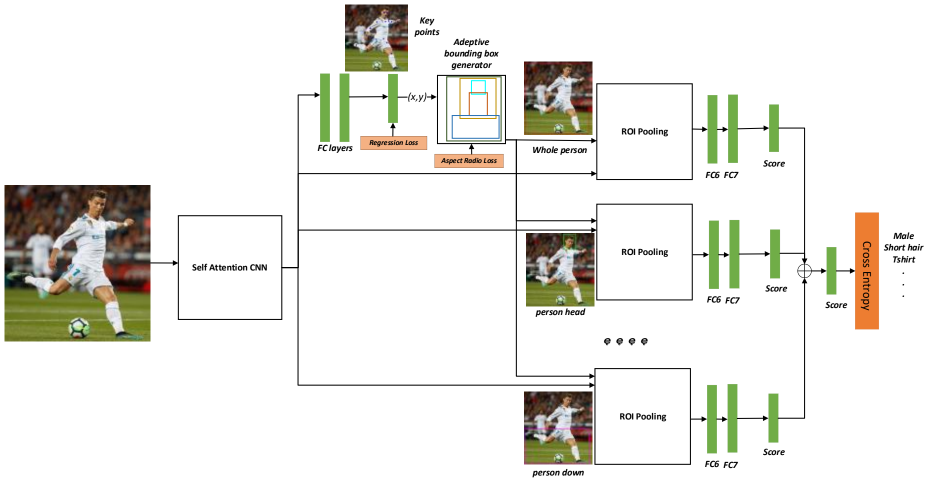


Fig.1. The pipeline of dual self-attention pedestrian attributes recognition. It consists of initial convolutional feature exaction layers, a key point localization network, an adaptive bounding box generator for each part, and the final attribute classification network for each part.

3. Methods

This paper proposes an end-to-end framework for pedestrian attributes recognition. In the method, high-quality, unique, and distinguishable features are obtained through spatial and semantic self-attention mechanisms, which are used for the pedestrian key points (Such as left eye, right eye, nose, left shoulder, right shoulder, etc. 17 points) are estimated, from the key points, the bounding boxes are generated adaptively. Instead of applying the transform to the whole image, we apply a spatial transform for each part and use the bilinear sampler to get the image features for subsequent attribute recognition. The whole network is learned end-to-end in a multi-task setting, attribute recognition as the main task and person key point prediction as to the auxiliary one. the framework is exemplified in Figure 1.

3.1. Dual self-Attention CNN

Good features are the premise for pedestrian attribute recognition. In order to obtain high-quality features, this paper proposes the dual attention mechanism of spatial self-attention and semantic self-attention. The spatial self-attention mechanism is used to salience the features that have correlation in space, and the semantic self-attention mechanism is used to salience the features that have correlations in channels. The network structure is shown in Figure 2.

The dual attention network is divided into four stages. From the top to the bottom of each subnet, the resolution of each level is gradually reduced by a multiple of 1/2, and the number of channels is increased by a factor of 2. The first stage consists of a high-resolution subnet, which contains 4 linearly transformed E-ecaneck residual modules.

The second, third, and fourth stages are composed of an improved E-ecablock module and a fusion module (Non-local block) of the spatial attention mechanism. The goal is to extract and merge features more extensively and deeper. The E-ecaneck module and E-ecablock module are constructed by channel attention as the basic modules to maximize the acquisition of useful channel information, Then integrate the Non-local block module in the spatial attention mechanism based on the channel attention mechanism to realize the effective extraction and fusion of spatial information. Finally, through the up-sampling operation, the features are output to realize the task of detecting key points of the human body and further realize the estimation of the human body posture.

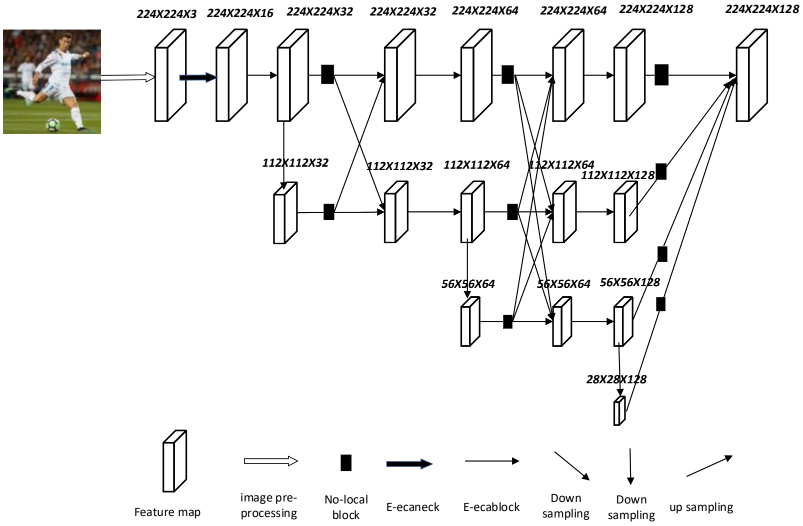


Fig 2. self-attention CNN. Features are extracted from the spatial self-attention network and semantic self-attention network.

Spatial Self-attention non-Local block

In deep learning networks, the dependency between features at different locations plays an important role in computer vision task. No-local has the following characteristics, Firstly, which calculates the features of a certain location as the weighted sum of the features of other locations; Secondly, which directly captures dependencies by calculating the interaction between any two locations, regardless of their distance; Third, which unfixed input size and can be embedded in many deep learning network structures. The non-local block structure is shown in Figure 3.

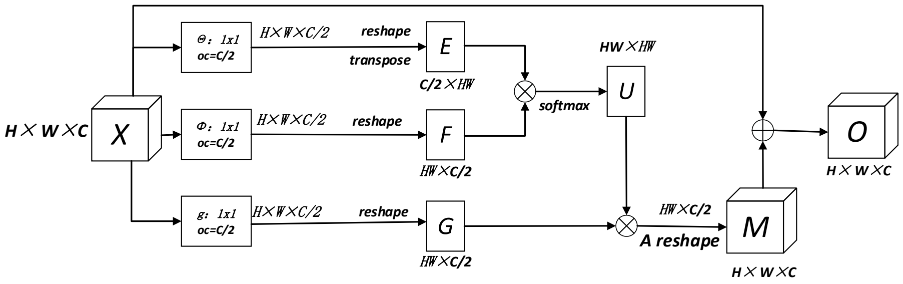


Fig.3. Non-local block

The Non-local block module is the core module of the spatial domain attention model (Non-local Neural Networks). As a non-local spatial attention method, it is not only limited to the local area but operates in the global area. The calculation formula is as follows the equation (1) shows:

$$y_i = \frac{1}{c(x)} \sum\nolimits_{\forall j} f(x_i, x_j) g(x_j) \qquad (1)$$

In equation (1), *i* and *j* are some spatial bits, *c(x)* is the normalization factor, and *x* and *y* are input and output data with the same size respectively. x_i is a vector, $f(x_i,x_j)$ is to calculate the correlation function between x_i and all x_j. $g(x_j)$ calculate the characteristic value of the input signal at position *j*. After formula (1), formula (2) can be obtained, where Z_i is the final output, which is the linear conversion matrix realized by the *1×1* convolution operation.

$$z_i = w_z y_i + x_i \qquad (2)$$

In this paper, the spatial attention module (non-local block) module is used to improve the fusion of the second, third and fourth stages, and it is added to the fusion of each resolution characterization based on integrating the channel attention, as shown in Figure 4. Shown. The non-local block module is operated in the global region, which can expand the receptive field. Therefore, it can effectively extract more beneficial information by different resolutions used for multi-scale fusion to get a better effect.

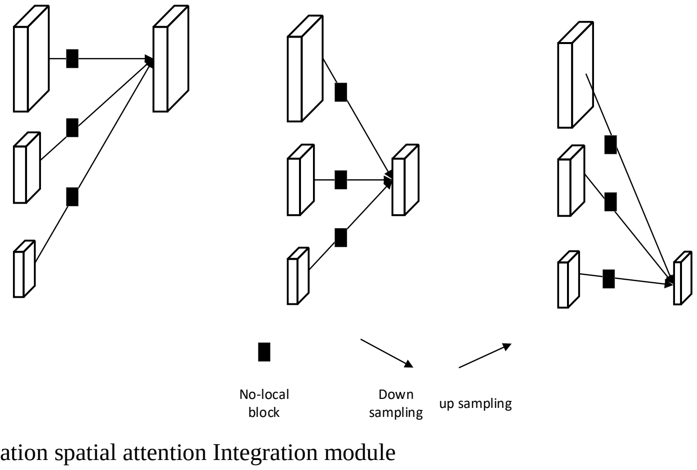


Fig.4. Integration spatial attention Integration module

Efficient Channel Attention network

The ECA module is a method of capturing local cross-channel information interaction. It conducts cross-channel information interaction without reducing the channel dimension and aims to ensure computational performance and model complexity. The ECA module can realize information interaction between channels through one-dimensional convolution with a convolution kernel size *K*, as shown in Figure 5.

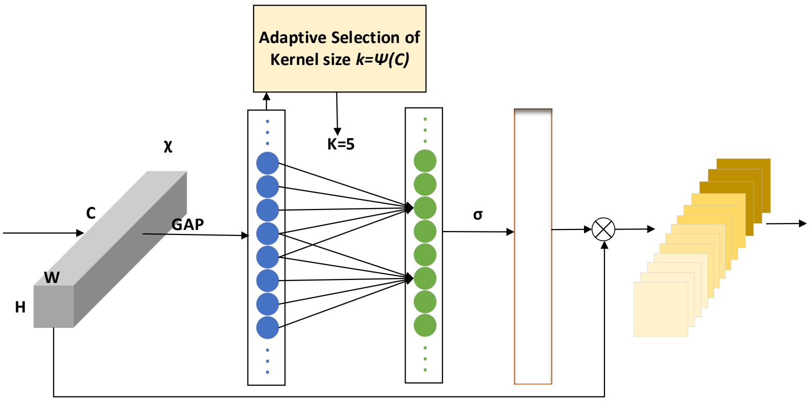


Fig.5. ECA (Efficient Channel Attention network) structure diagram

The output of a convolution block is $\chi\epsilon R_{H\times W\times C}$, *W*, *H*, *C*, represent width, height, and number of channels, respectively, and GAP represents global average pooling. The ECA module passes a one-dimensional convolution with a convolution kernel size of *k*. To realize the information exchange between channels, as shown in formula (1) Where σ is a *Sigmoid* function, *C1D* represents one-dimensional convolution, and *k* represents *k* parameter information, which is the output signal. This method of capturing cross-channel information interaction ensures performance results and model efficiency.

$$\omega = \sigma\left(C1D_K(y)\right) \tag{3}$$

Inspiring from the method of bottleneck and basic block in ResNet network, this paper integrates ECABlock into the bottleneck (bottleneck) and residual (basic block) to get an improved E-ecablock and E-ecaneck, which replace the bottleneck module and basic block module of the original network, to realize cross-channel information interaction without reducing the channel dimension.

E-ecablock module and E-ecaneck module

The E-ecablock module proposed in this paper contains two ECABlock modules with the size of the 3×3 convolution kernel and one residual connection, as shown in Figure 6. The E-ecaneck module includes two ECABlock modules with a size of 1×1 convolution kernel, one ECABlock module with a size of 3×3 convolution kernel, and a residual connection, as shown in Figure 7.

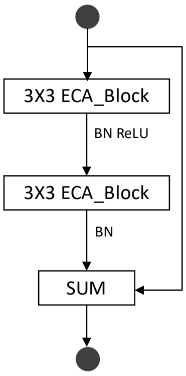


Fig.6. E-ecablock

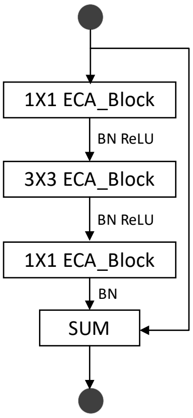


Fig.7. E-ecaneck

Inspired by the literature [35], this article adds the channel attention module before the convolution to get better results in feature extraction. This article adds two 3×3 convolutions in the E-ecablock module. Using ECA_Block, before each convolution, to obtain meaningful features by assigning attention weights to different channels of the input feature map. Finally, the output result after the convolution and the feature of the initial input is summed through the residual connection, which can obtain a better channel feature. In order to increase the efficiency of channel information extraction, the same method is adopted, adding the ECA_Block module before each layer of convolution in the E-ecaneck module. The channel dimension of the feature map is reduced and added through the ECA_Block module, multiple convolution operations are performed for feature extraction, which can obtain more useful feature information.

3.2. Key Point Estimation and Adaptive Part Generation

Key Point Estimation As shown in Figure 1, there are three fully-connected layers after the convolutional features, with output dimensions 2048, 2048, and *2N,* respectively.

Here *N* is the number of key points. After each fc layer,the ReLU activation function and drop out layer (ratio 0.5) is used. We use L2 distance loss for key point estimation, $\sum_i^n \|\widehat{p_\iota} - p_i\|_2^2$, where $\widehat{p_\iota}, p_i \in R^2$ are the normalized ground truth and estimation for key point *i*.

Some attributes are clearly associated with certain object parts. We encode such prior knowledge by specifying a subset of key points p_t for each part *t*. For example, in Figure 8, for body we have P_{body} = {body, head, torso, upbody, lowerbody }. The initial part bounding box $b_t = [w_t, h_t, x_t, y_t]$ is defined as an enlarged bounding box of key points in P_t,

$$w_t = s \cdot (\max_x(p_t) - \min_x(p_t)) \qquad h_t = s \cdot \left(\max_y(p_t) - \min_y(p_t)\right) \quad (4)$$

$$x_t = \frac{1}{2}(\max_x(p_t) + \min_x(p_t) - w_t) \qquad y_t = \frac{1}{2}\left(\max_y(p_t) + \min_y(p_t) - h_t\right) \quad (5)$$

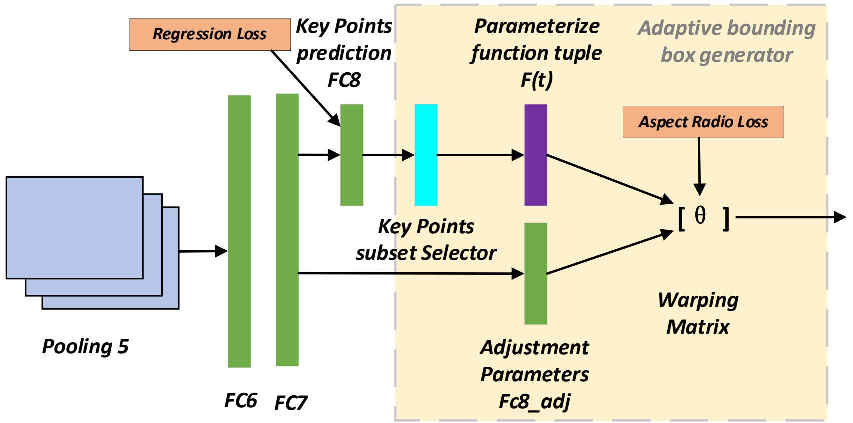

Regression Loss
Key Points prediction FC8
Parameterize function tuple F(t)
Adaptive bounding box generator
Aspect Radio Loss
[θ]
Key Points subset Selector
Warping Matrix
Pooling 5
Adjustment Parameters Fc8_adj
FC6
FC7

Fig.8. Adaptive bounding box generator

Here, $\max_x(p_t)$is the maximum *x* coordinate in key points of P_t. Other notations are similar. *w/h* is the initial box's width/height. *x/y* is initial box's upper left corner. *s* is a constant scalar larger than *1.0*. It is set to *1.5* in our experiment. The initial bounding box is then adaptively adjusted by free parameters $\Delta = [\Delta_w, \Delta_h, \Delta_x, \Delta_y]$. The final bounding box is defined as $[w_t(1 + \Delta_w), h_t(1 + \Delta_h), x_t + \Delta_x, y_t + \Delta_y]$. To learn the adjustment parameters, we add one more fully connected layer (fc8_adj) to the previous layer (fc7), with *4* output values. This is depicted in Figure 8. The final bounding box could have too distorted dimensions due to the free adjustment parameters. In order to alleviate this issue, we introduce a bounding box aspect ratio loss as L_r^t.

$$L_r^t = \begin{cases} \frac{1}{2}\{[\alpha[h_t(1+\Delta_h)]^2 - w_t(1+\Delta_w)]^2\}_+ & if\ h_t(1+\Delta_h) > w_t(1+\Delta_w) \\ \frac{1}{2}\{[\alpha[w_t(1+\Delta_w)]^2 - h_t(1+\Delta_h)]^2\}_+ & if\ w_t(1+\Delta_w) > h_t(1+\Delta_h) \end{cases} \quad (6)$$

where α is a ratio threshold (set to 0.6 in the experiment). The loss is 0 when the value in bracket $\{\}_+$ is less than 0.

1.1 Hierarchical Pedestrian Attribute Recognition

There are some differences between various pedestrian attributes recognition, that is, some attributes characteristics scattered in various parts of the human body, such as gender, fat or thin, and other attributes characteristics are concentrated in a certain part of the human body, such as hairstyle. According to the characteristics scattered in the whole body or gathered in a local, we divided the pedestrian attributes into global and local. In real deployment, this can be implemented according to the pre-configuration.

Our goal is to recognize a set of human attributes $\{a \in A\}$ for all the people in an image I in unconstrained scenes. Assume that a target person's bounding box in I as b, divide human body b into a set of parts s $\{s \in S\}$ according to human body key points. Probability is used to determine whether an attribute exists, this is, estimated the probability of existence of attribute a on the target person given by measurements V. The attributes are recognized from both human body b and parts S. The measurements V is defined as $V = \{b, S, I\}$. Formula 8 is used to evaluate for each attribute a the conditional probability from the given measurements V :

$$Score(a; b, S) = \omega_{a,b}^{T} \cdot \emptyset(b; I) + \max_{s \in S} \omega_{a,s}^{T} \cdot \emptyset(s; I) \tag{8}$$

where φ(b; I) is the extracted fc7 features from region b in image I, while wa,· are the scoring weights of attribute a for different regions.

The scoring terms for person bounding box b and parts $\{s \in S\}$ form the basis of our model and are shown on the upper two paths in Fig. 8. Their sum can be regarded as a pose-normalized deep representation at the score level. Such score fusion is found to be more effective than feature fusion in our task because the latter would generate a very large feature vector from the many parts and overfits easily. In our CNN, the scoring weights and feature vectors are jointly learned for all attributes $a \in A$.

Note for the part set S, we select the most informative part s for each attribute by a max score operation and only add the maximum to the final attribute score. This is because human attribute signals often reside in different body parts, so not all parts should be responsible for recognizing one particular attribute. For example, the head part can hardly be used to infer the "long pants" attribute. Through the max pooling of part scores, we are now able to capture those distributed attribute signals from the rich part collection.

4. Experiments

In this paper, the model training is divided into two stages: first, key point detection model is trained with COCO2017 dataset, and then attribute recognition model is trained with PETA and RAP dataset.

1.2 Data Set and Evaluation Indicators

COCO2017 dataset has 17 key points of human posture, it contains 200000 images, 250000 images with 17 key points. 57000 images are used for training, 5000 images are used for validate, 20000 images are used for test. The 17 key points are nose, right eye, left eye, right ear, left ear, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, left ankle.

Table 1. Thirty-five attributes of PETA dataset

No	attribute	No	attribute
1	accessory Muffler	19	carryingMessengerBag
2	personalLarger60	20	personalLess45
3	accessoryHat	21	lowerBodyJeans
4	personalMale	22	carryingBackpack
5	hairLong	23	footwearShoes
6	footwearLeatherShoes	24	upperBodyTshirt
7	upperBodyPlaid	25	lowerBodyShortSkirt
8	personalLess60	26	footwearSneaker
9	personalLess30	27	carryingNothing
10	upperBodyShortSleeve	28	upperBodyJacket
11	accessoryNothing	29	carryingOther
12	carryingPlasticBags	30	lowerBodyShorts
13	upperBodyFormal	31	accessorySunglasses
14	upperBodyOther	32	upperBodyThinStripes
15	upperBodyCasual	33	footwearSandals
16	lowerBodyFormal	34	upperBodyLogo
17	lowerBodyCasual	35	upperBodyVNeck
18	lowerBodyTrousers		

Table 2. Fifty-one attributes of RAP dataset

No	attribute	No	attribute
1	Female	27	lb-Dress
2	Clerk	28	attach-Backpack
3	ub-Vest	29	attach-SingleShoulderBag
4	lb-TightTrousers	30	attach-Other
5	lb-Jeans	31	attach-HandBag
6	Customer	32	ub-Jacket
7	shoes-Boots	33	ub-ShortSleeve
8	action-Calling	34	Age31-45
9	lb-LongTrousers	35	Age17-30
10	lb-Skirt	36	action-Pulling
11	hs-LongHair	37	ub-TShirt
12	ub-TShirt	38	shoes-Sport
13	shoes-Leather	39	ub-Tight
14	hs-Glasses	40	attach-PlasticBag
15	action-Pusing	41	action-Gathering
16	hs-Hat	42	attach-PaperBag
17	hs-BaldHead	43	action-Holding
18	hs-BlackHair	44	ub-Sweater
19	AgeLess16	45	hs-Muffler
20	ub-Colton	46	BodyFat
21	ub-SuitUp	47	action-CarrybyArm
22	action-CarrybyHand	48	shoes-Casual
23	attach-HandTrunk	49	BodyThin
24	lb-ShortSkirt	50	BodyNormal
25	shoes-Cloth	51	action-Talking
26	attach-Box		

The PETA dataset consists of 19,000 people images collected from 10 small-scale pedestrian datasets. The entire data set is randomly divided into three non-overlapping parts: 9500 for training, 1900 for verification, and 7600 for testing. Due to the imbalance of attribute samples, attributes with a sample ratio of more than 5% in the 35 attribute labels are generally selected for evaluation. The 35 attributes are shown in Table 1.

The RAP data set contains 41,585 images from 26 indoor monitoring cameras, and each image has 69 binary attributes and 3 multi-category attributes. According to the official agreement, the entire data set is divided into 33,268 training images and 8,317 test images. The recognition performance of 51 binary attributes is evaluated. The 51 attributes are shown in Table 2.

The evaluation indicators for quantitative comparison use the general label-based mean Accuracy (mA) indicator and the Example-based accuracy rate (Accuracy, Acc) indicator and precision rate (Precision, Prec) index, recall rate (Recall, Rec) index and F1 value index.

4.1. Comparative Experiment

This article compares two sets of experiments: Experiment 1 compares the benchmark network Base-CNN and the various modules proposed in this article on the two data set test sets; Experiment 2 compares the model in this article with some current pedestrian attribute recognition models Comparison of quantitative evaluation index results.

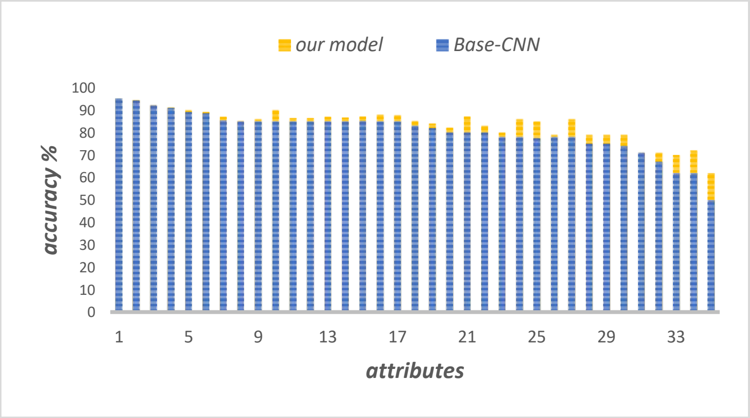
Experiment Related Settings

In this experiment, the COCO2017, RAP, and PETA image data sets will be cropped with the hip of the human body as centre, and adjusted to a fixed ratio of 4:3 for height and width, the size of the image being cropped to 256 × 192. Dataset are used for backbone network training for efficient and accurate human key point detection. In this experiment, the SGD optimizer is selected to optimize the model. The training epochs is set to 200, the training batch size is set to 20, and the learning rate is 0.001. Pedestrian attribute features and key points recognition share the feature map obtained by the backbone network. The feature map of each bounding box is pooled to obtain a 7x7x128 matrix through the corresponding backbone area features. The process of pedestrian attribute recognition training will fix the parameters of the backbone network, and only optimize the parameters of the full link layer. SGD optimization method is used in the training process, with an initial learning rate of 0.001. The whole system is trained and validate based on the pytorch deep learning framework.

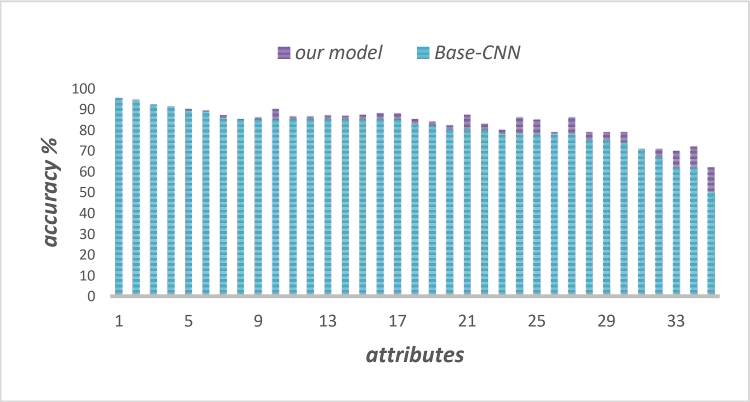
Method validity experiment

In this group of experiments, this paper adds non-local, E-ecablock, and E-ecaneck on the reference network Base-CNN respectively to compare and verify that each method recognizes attributes. The experimental results on the PETA and RAP data sets are shown in Table 3. The content in bold in the table indicates the best result under this index. It can be seen from Table 3 that after the non-local block, on the two data sets, the mA index increased by 1.57 percentage points and 1.94 percentage points respectively compared with the benchmark model; Adding the E-ecablock can also make the recognition effect of the model have a small gain (mA index gain ranges are

0.42 percentage points and 0.26 percentage points respectively); on the basis of adding E-ecaneck module proposed in this article, the value of each evaluation index can be improved in the two data sets, and the mA index has increased by 1.92 and 1.85 respectively. In general, compared with the benchmark model Base-CNN, the mA index on the PETA and RAP data sets has increased by 5.91 percentage points and 6.05 percentage points, respectively. In Figure 9, by comparing the recognition accuracy rates of the benchmark model Base-CNN and this model in the 35 attributes of the PETA dataset and the 51 attributes of the RAP dataset, it can be seen that the model in this paper has a recognition effect on most attributes. The improvement, especially attributes with a small number of samples in dataset, is more obvious. The Y-axis in Figure 4 represents the accuracy of recognition, and the X-axis represents the attribute number, which corresponds to the attribute numbers in Table 1 and Table 2, respectively.



(a) PETA



(b) RAP

Fig. 9. The recognition accuracy of each attribute on PETA and RAP of the benchmark model and our model

Table 3. Effectiveness comparison of different modules on PETA dataset

Model	mA	Prec	Rec	F1
Base-CNN	80.72	85.42	81.76	83.55
Base-CNN +non_local	82.29	83.41	82.49	82.95
Base-CNN+non_local+ E-ecablock	82.71	83.87	84.21	84.04
Base-CNN +non_local + E-ecablock + E-ecaneck	**86.63**	**86.79**	**85.73**	**86.26**

Table 4. Effectiveness comparison of different modules on RAP dataset

Model	mA	Prec	Rec	F1
Base-CNN	76.65	76.93	76.93	76.93
Base-CNN +non_local	78.59	78.16	78.85	78.5
Base-CNN+non_local+ E-ecablock	78.85	78.36	79.25	78.8
Base-CNN +non_local + E-ecablock + E-ecaneck	**82.70**	**81.85**	**83.82**	**82.82**

Comparison with other models

Talbe 5. Effect comparison of different models on PETA unit:%

model	mA	Prec	Rec	F1
DeepMAR	82.89	83.68	83.14	83.41
HPNet	81.77	84.82	83.24	84.02
PGDM	82.97	86.86	84.68	85.76
MPAR	83.57	86.19	85.07	85.63
IA^{2}Net	84.13	85.73	**86.07**	85.90
Our model	**86.63**	**87.79**	85.73	**86.75**

Talbe 6. Effect comparison of different models on RAP unit:%

model	mA	Prec	Rec	F1
DeepMAR	73.79	74.92	76.21	75.56
HPNet	76.12	77.33	78.79	78.05
PGDM	74.31	78.86	75.90	77.35
LGNet	78.68	80.36	79.82	80.09
MPAR	75.64	78.29	79.23	78.76
VSGR	77.91	82.05	80.64	81.34
RCRA	78.47	**82.67**	76.65	79.55
IA^{2}Net	77.44	79.01	77.45	78.22
Our model	**82.70**	81.85	**83.82**	**82.82**

Many models are based on PETA and RAP dataset for pedestrian attribute recognition, mainly including PGDM (Posed Guided Deep Model) [13], VSGR (Visual-Semantic Graph Reasoning net) [36], HPNet (Hydra Plus Net) [25], DeepMAR (Deep Multi-Attribute Recognization model) [12], LGNet (Location Guided Network) [37], MPAR (Multistage Pedestrian Attribute Recognition method) [38], RCRA (Recurrent Convolutional and Recurrent Attention model) [39] And IA^{2}Net (Image-Attribute reciprocally guided Attention Network) [40]. Among them, DeepMAR only extracts the global features of pedestrians. PGDM and LGNet extract the local features of different parts of pedestrians through the pedestrian local area network. RCRA and IA^{2}Net apply the attention mechanism to the network model. The bold font in the table 5 and 6 means the best result. From Table 5 and Table 6, we can see that our model has strong competitiveness in mean accuracy, precision and F1 in PETA and RAP dataset. The mean accuracy of attribute recognition on PETA and RAP datasets reached 86.63% and 82.70%, superior to the existing models.

Sample recognition analysis:

Figure 10 shows 8 images labelled *a-h*. The histogram is the prediction results of our model and StrongBaseline [41]. The horizontal axis of the histogram represents the predicted attribute, the vertical axis represents the probability of predicted attribute. The blue bar is the probability predicted by the our model, and the orange bar is the probability predicted by the StrongBaseline model. For the attributes of clear samples, such as samples *a*, *c,* and *e*, both methods have excellent and similar prediction results. For the attributes of blurry and occluded samples, such as the *g* sample is blurry and the 'ShoulderBag' attribute is occluded by body, the prediction probabilities of our model and StrongBaseline are only 0.0033 and 0.0027, which are shown zero on the histogram. In addition, Not only we found that the attribute with high resolution, such as 'Back', 'UpperLogo', 'Trousers', 'Age', ' Side', etc. have excellent prediction probability by our model, but also the 'Back' attribute of sample *d* and *f*, the 'Side' attribute of sample *g* and *h*, the prediction results of our model are much higher than StrongBaseline. Based on these, we infer that the spatial self-attention module in our model network plays an important role in capturing the global contextual information and improving the recognition accuracy of the model.

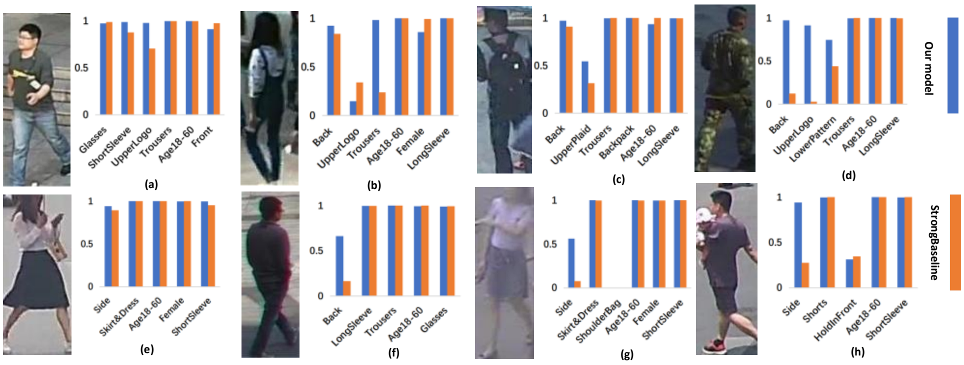


Fig.10. Comparison of pedestrian attribute recognition accuracy. The accuracy of our model is significantly higher than StrongBaseline on low quality image

Sample heatmap analysis:

To better understanding where the model focuses on, we visualize these localization results in Figure 11, which shows randomly selected samples on the RAP dataset. Based on our model: (1) The two groups *a* and *b* are 'Skirt' positive and 'BackPack' positive, respectively. Samples *a.1* and *b.1* are the areas that the main network Base-CNN focus on, *a.2* and *b.2* are the areas that are concerned after dual self-attention. We can observe that the attribute regions of interest are more complete after dual self-attention model. (2) The two groups *c* and *d* are ’LongSleeve’ negative and ’BackPack’ negative, respectively. *c.1* and *d.1* are the areas of interest for the main network Base-CNN, *c.2* and *d.2* are the regions of interest after dual self-attention model. We observe that the former focuses on invalid regions, while the latter does not pay attention to any regions. From Figure 12 the results can be seen, (1) Group *e* is ‘Skirt’ positive; *e.1* and *e.2* are the regions concerned by StrongBaseline and our model. We can observe that most of the regions concerned by StrongBaseline are invalid, while our model basically pays attention to all regions of the ‘Skirt’.(2) Group *f* is ’BackPack’ negative. Similarly, *f.1* and *f.2* are the regions that StrongBaseline and our model focus on respectively. As analysed above, StrongBaseline pays attention to an invalid region, while our model does not focus on any regions. According to these, we infer that our model will first predict an attention region for positive or negative attribute, and then strengthen the attention region for positive attribute or weaken the attention region for negative attribute by dual self-attention, such as groups *a* and *c*.

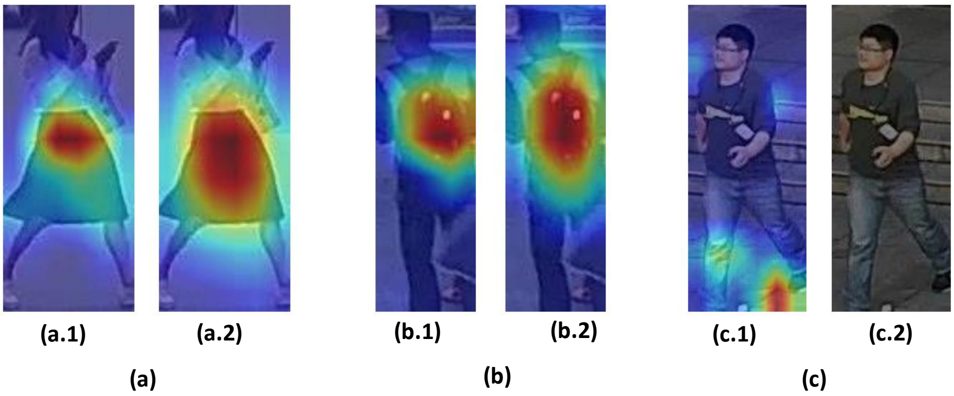


Fig.11. Heatmap of Base-CNN and our model on samples.

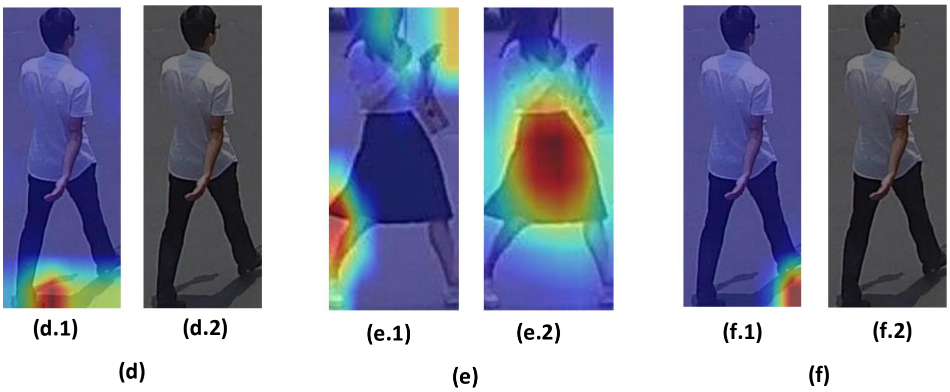


Fig.12. Heatmap of StrongBaseline and our model on samples.

5. Conclusions

According to the semantic correlation between pedestrian attributes and the spatial correlation between each attribute and body parts, this paper proposes to use spatial self-attention mechanism to extract spatial dependent features, and the channel self-attention mechanism to extract semantic relevance of features between channels. Use Non-Local network structure to highlight important features and weaken unimportant features Two attention modules, E-eaclock and E-ecanenck, are designed as the basic modules to form a feature extraction network, which can effectively improve the representation ability of spatial and semantic features. The key points of pedestrians and five bounding boxes of body part are extracted through feature extraction in the previous stage. The accuracy of attribute recognition is improved by making full use of attribute space and semantic constraints in attribute recognition. The experimental results show that the model in this paper can improve the recognition effect of attributes in both the PETA and RAP pedestrian attribute data sets. Through the analysis of the heat map of the experimental results, it can be seen that our model will first predict an attention region for positive or negative attribute, and then strengthen the attention region for positive attribute or weaken the attention region for negative attribute by dual self-attention. This method is important for pedestrian attribute recognition and can be applied to various practical applications.

Acknowledgment. This research work was supported in part by National Key R&D Program of China (2019YFC1520500, 2020YFC1523004).

References

1. Gordo A, Almazan J, Revaud J, et al. End-to-end learning of deep visual representations for image retrieval[J]. International Journal of Computer Vision, 2017, 124(2): 237-254.
2. Dubey S R. A decade survey of content based image retrieval using deep learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021.
3. Cffa B , Gg A , Jvdw A , et al. Saliency for fine-grained object recognition in domains with scarce training data[J]. Pattern Recognition, 2019, 94:62-73.
4. Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3367-3375).
5. Ding, S., Lin, L., Wang, G., & Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition, 48(10), 2993-3003.
6. Ryan Layne, Timothy M Hospedales, Shaogang Gong, and Q Mary. Person re-identification by attributes. In Bmvc, volume 2, page 8, 2012.
7. Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei,and Stan Z Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In ICB. IEEE, 2015.
8. Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In ACM MM, 2014..
9. Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In Proc. ACM Multimedia, 2014.
10. R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person reidentification by attributes. In Proc. BMVC, 2012.

11. J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In Proc. ICCV Workshops, 2013.
12. D. Li, X. Chen, and K. Huang, “Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios,” in Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on. IEEE, 2015, pp. 111–115.
13. D. Li, X. Chen, Z. Zhang, and K. Huang, “Pose guided deep model for pedestrian attribute recognition in surveillance scenarios,” in 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018, pp. 1–6.
14. N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In Proc. CVPR,2014.
15. L. Bourdev and J. Malik, “Poselets: Body part detectors trained using 3d human pose annotations,” in Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 1365–1372.
16. D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In Workshop on Applications of Computer Vision (WACV), pages 1–8, 2009.
17. J.Q. Zhu, S.C. Liao, Z. Lei, D. Yi, S.Z. Li. Pedestrian Attribute Classification in Surveillance: Database and Evaluation. IEEE ICCV Workshop, 2013.
18. J.Q. Zhu, S.c. Liao, D.Yi, Z. Lei, S.Z. Li. Multi-Label CNN Based Pedestrian Attribute Learning for Soft Biometrics. IAPR ICB, 2015.
19. Y.Deng, P.Luo, c.c. Loy, X.O. Tang. Pedestrian Attribute Recognition at Far Distance. ACM MM, 2014.
20. W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In AAAI, 2017.
21. P. Sudowe, H. Spitzer, B. Leibe. Person Attribute Recognition with a Jointly-trained Holistic CNN Model. IEEE ICCV Workshop, 2015.
22. A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, “Multi-task cnn model for attribute prediction,” IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 1949–1959, 2015.
23. Yao, C. , et al. "Hierarchical pedestrian attribute recognition based on adaptive region localization." IEEE International Conference on Multimedia & Expo Workshops IEEE, 2017.
24. Feng, Z. , et al. "Learning Spatial Regularization with Image-Level Supervisions for Multi-label Image Classification." IEEE (2017).
25. X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang,“Hydraplus-net: Attentive deep features for pedestrian analysis,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 350–359.
26. Li, Qiaozhe, et al. "Pedestrian Attribute Recognition by Joint Visual-semantic Reasoning and Knowledge Distillation." IJCAI. 2019.
27. Huang, Lingli. "Improved non-local means algorithm for image denoising." Journal of Computer and Communications 3.04 (2015): 23.
28. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7794-7803).
29. Zhang, Han, et al. "Self-attention generative adversarial networks." International conference on machine learning. PMLR, 2019.
30. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
31. Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
32. Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.
33. Chen, Tianlong, et al. "Abd-net: Attentive but diverse person re-identification." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

34. Tan, Zichang, et al. "Attention-based pedestrian attribute analysis." IEEE transactions on image processing 28.12 (2019): 6126-6140.
35. Yang, Z. , et al. "Gated Channel Transformation for Visual Recognition." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2020.
36. Q. L. X. Z. R. H. K. HUANG, “Visual-semantic graph reasoning for pedestrian attribute recognition,” in Association for the Advancement of Artificial Intelligence, AAAI, 2019.
37. P. Liu, X. Liu, J. Yan, and J. Shao, “Localization guided learning for pedestrian attribute recognition” 2018.
38. Zheng S F, Tang J, Luo B, et. al. Multistage pedestrian attribute recognition method based on improved loss function. Pattern Recognition and Artificial Intelligence.2018.31(12):1085-1095.
39. Zhao X, Sang L, Ding G, et al. Recurrent attention model for pedestrian attribute recognition. AAAI Press, 2019.
40. Ji Z, He E, Wang H et al. Image-attribute reciprocally guided attention network for pedestrian attribute recognition. Pattern Recognition Letters.2019, pp. 89-95.
41. J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang,“Rethinking of Pedestrian Attribute Recognition: Realistic Datasets with Efficient Method,“ arXiv preprint arXiv:2005.11909, 2020.

Zhongkui Fan receive the BS degree in Computer engineering from Huanghe science and technology College, and the MS degree in Computer Engineering from Jiangxi University of Science and Technology. He is currently working toward the PhD degree in the School of Communication and Information Engineering with Shanghai University. His research interests include computer vision and machine learning.

Ye-peng Guan was born in Xiaogan, Hubei Province, China, in 1967. He received the B.S. and M.S. degrees in physical geography from the Central South University, Changsha, China, in 1990, 2006, respectively, and the Ph.D. degree in geodetection and information technology from the Central South University, Changsha, China, in 2000. From 2001 to 2002, he did his first postdoctoral research at Southeast University in electronic science and technology. From 2003 to 2004, he did his second postdoctoral research at Zhejiang University in communication engineering, and he had been an Assistant Professor with the Department of Information and Electronics Engineering, Zhejiang University. Since 2007, he has been a Professor with School of Communication and Information Engineering, Shanghai University. He is the author of more than 120 articles, and more than 20 patents. His research interests include intelligent information perception, digital image processing, computer vision, and security surveillance and guard.

Received: August 15, 2022; Accepted: January 03, 2023.