Unsupervised Multiple Granularities Attention-Attribute Learning for Person Re-identification

Rui Yang, Guoqiang Xiao, Song Wu

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

Laboratory of Digital Media and Communications of Southwest University
Chongqing, China

Content

1 Research problem

Person re-identification (Re-id) is a key technology in many video surveillance applications, such as person association, multi-target tracking and behavior analysis.

Given an image of a target pedestrian, the aim of person re-identification is to match stated person across non-overlapping camera views.

Application scenarios

2 Present situation

Research status:

- Person re-identification is a key problem in computer vision
- Person re-identification has become a difficult problem in image retrieval
- Person re-identification is an important technical way of social public safety

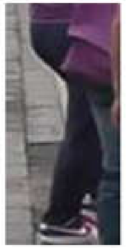
Existing problems:

- Camera field of view non overlapping
- There are big differences in shooting angle and pedestrian posture
- The traditional feature descriptor is unstable
- The amount of data explodes, but there is little labeled data

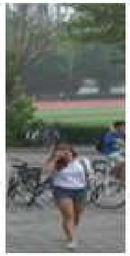
Angle transformation

Attitude transformation

Illumination transformation



Incomplete picture



Small proportion of person

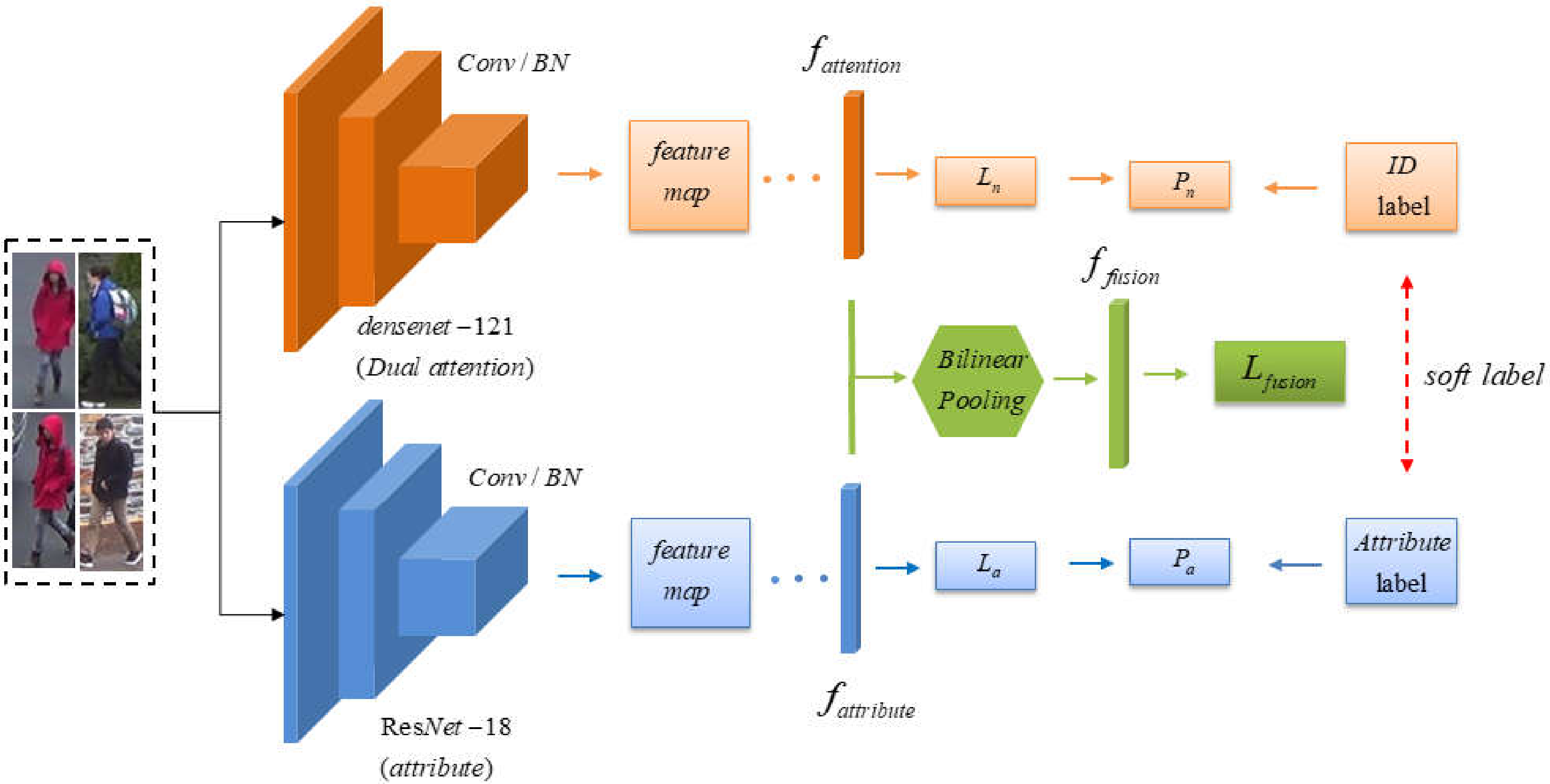
Occlusion

Solution

In recent years, there are two main ideas in Person re-identification

1. Extract a visual feature with both discriminant and robustness to describe person
2. Design an appropriate metric function to maximize the correct matching rate

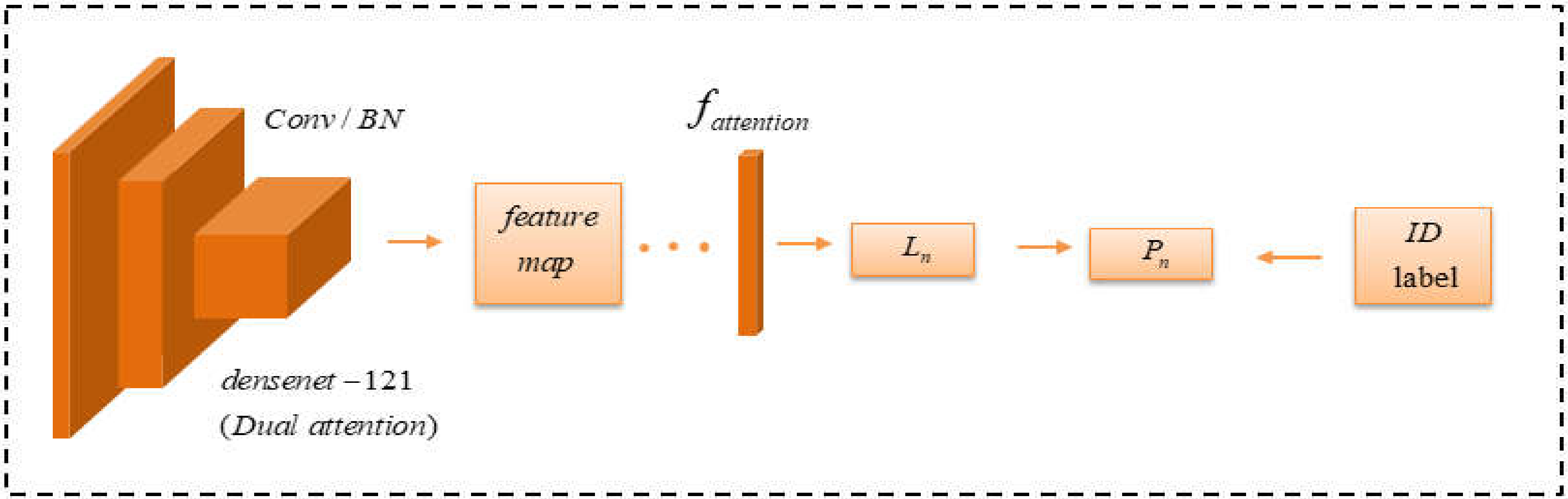
3 Multiple Granularities Attention-Attribute Learning



Attention branch

Attention loss：

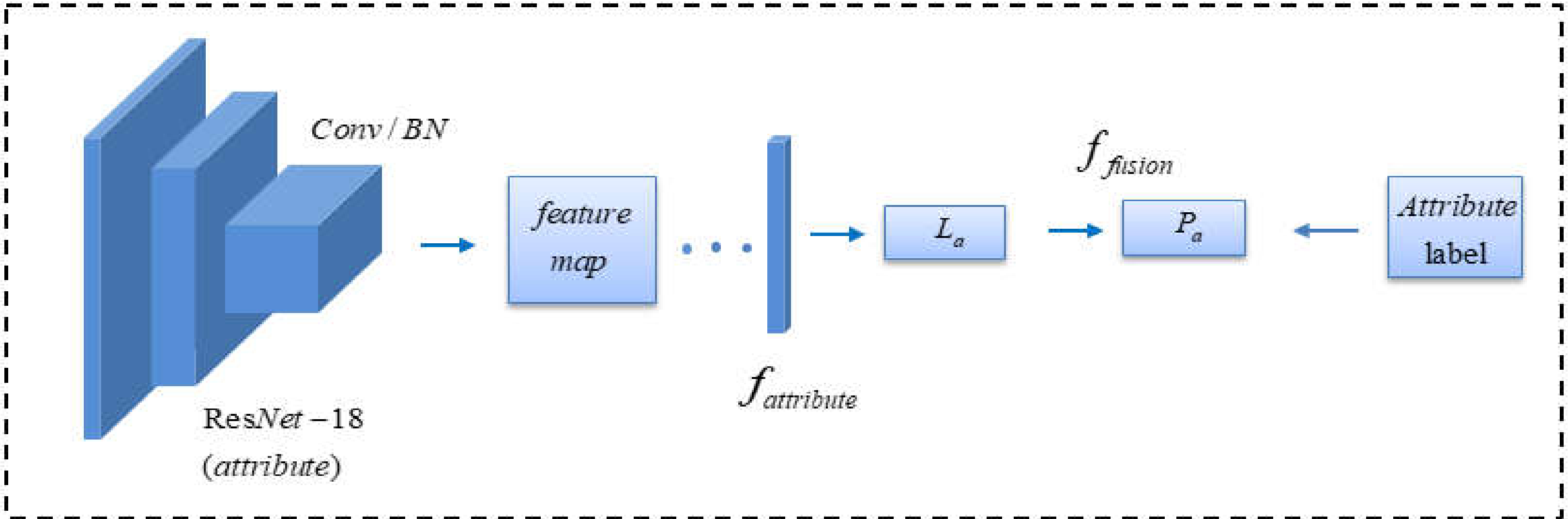
$$L_n = (x_i, y_i^j, \theta) = -\frac{1}{N_s}\sum_{i=1}^{N_s} p(x_i, y_i^j)\log p(x_i)$$



Attribute branch

Attribute loss：

$$L_a = (x_i^k, \boldsymbol{a}_i^k, \theta) = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{k=1}^{M}(a_i^k \log(p(x_i^k)) + (1 - a_i^k)\log(1 - p(x_i^k))$$



Feature fusion:

Most feature fusion is based on feature connect. However, for person re-identification , this connect method will lose the corresponding relationship between different pedestrian features, which is not the real feature fusion;

Supervised algorithm:

Most of the existing person re-identification methods rely on a large number of labeled data to train the convolution neural network model, which is unrealistic

Bilinear Pooling Embedding:

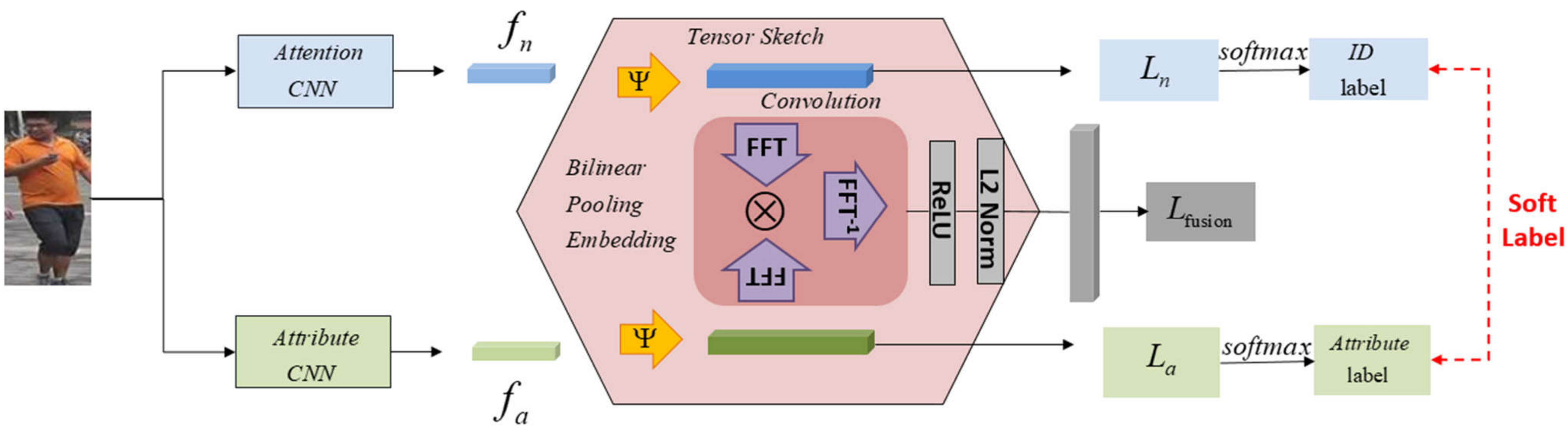
The outer product is used to further mine the corresponding relationship between features, and the "attribute correlation" principle is used to jointly optimize the network, so as to improve the scalability of the network model in practical application.

Unsupervised Bilinear Pooling Embedding

$$f_n = D(x_i)$$

$$f_a = R(x_i)$$

$$\rightarrow \quad b^* = \underset{b \in B}{\arg\max}\; p(b \mid f_n, f_a; \theta)$$



（1）Bilinear Embedding：

$$A(x_i) = \sum_{i=1}^{N_s} f_n f_a^{\;T}$$

$$f_n \otimes f_a = \begin{bmatrix} f_n^1 f_a^1 & f_n^1 f_a^2 & \cdots & f_n^1 f_a^{n_2} \\ f_n^2 f_a^1 & f_n^2 f_a^2 & \cdots & f_n^2 f_a^{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ f_n^{n_1} f_a^1 & f_n^{n_1} f_a^2 & \cdots & f_n^{n_1} f_a^{n_2} \end{bmatrix} \in \mathbb{R}^{n_1 \times n_2}$$

Tensor Sketch

$$\Psi(f_n \otimes f_a, h, s) = \Psi(f_n, h, s) * \Psi(f_a, h, s)$$

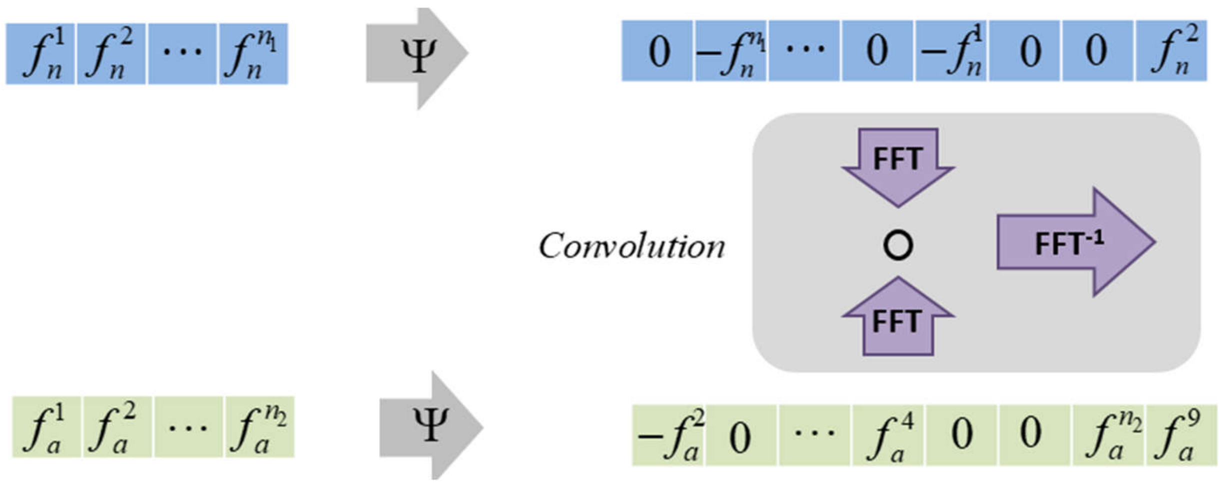
$$\Psi = [0, \ldots, 0] \qquad \Psi[h[i]] = \Psi[h[i]] + s[i] \cdot f[i]$$

$$s \in \{-1, 1\}^p \qquad h \in \{1, 2, \cdots, d\}^p$$

IEEE Xplore® Digital Library
ISCAS 2020
2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020
CAS

Meanwhile, the convolution theorem indicates that convolution in the time domain is equivalent to element-wise product in the frequency domain. This can be effectively extended to the convolution operation of the tensor sketch.

$$f_n * f_a = FFT^{-1}\left(FFT(f_n) \odot FFT(f_a)\right)$$



（2）Spatial Pooling：

$$f_{fusion} = \frac{1}{M}\sum_{i=1}^{M}(f_n * f_a)$$

Fusion Loss：

$$L_{fusion}(x_i, y(x_i), \theta) = -\frac{1}{N_s}\sum_{i=1}^{N_s} p\left(x_i^N,\ y\left(x_i^N\right)\right)\log \hat{p}(x_i)$$

where $\hat{p}(x_i)$ is predicted probability of f_{fusion}

Final Loss：

$$L_{sup} = L_{fusion} + \lambda_1 L_n + \lambda_2 L_a$$

IEEE Xplore® Digital Library

ISCAS 2020

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

CAS

Algorithm 1 Bilinear Pooling Embedding.

Input:

The Attention branch feature vector $f_n \in R^{n_1}$

The Attribute branch feature vector $f_a \in R^{n_2}$

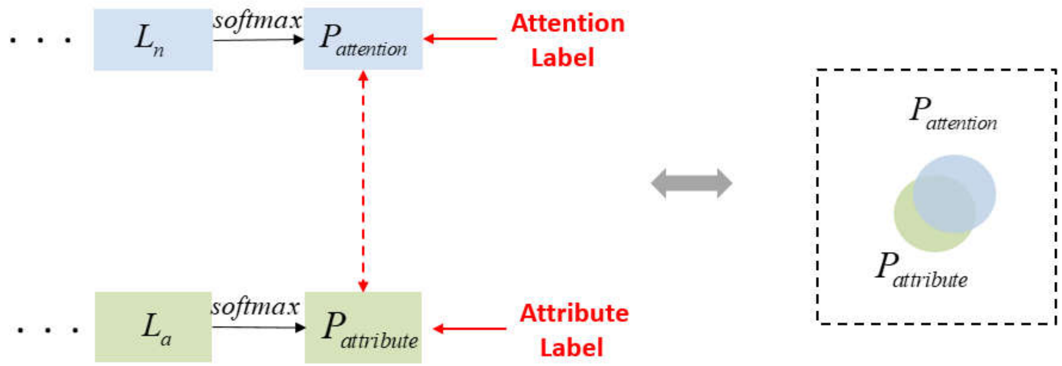
Output:

Feature map $\Phi(f_n, f_a) \in R^d$

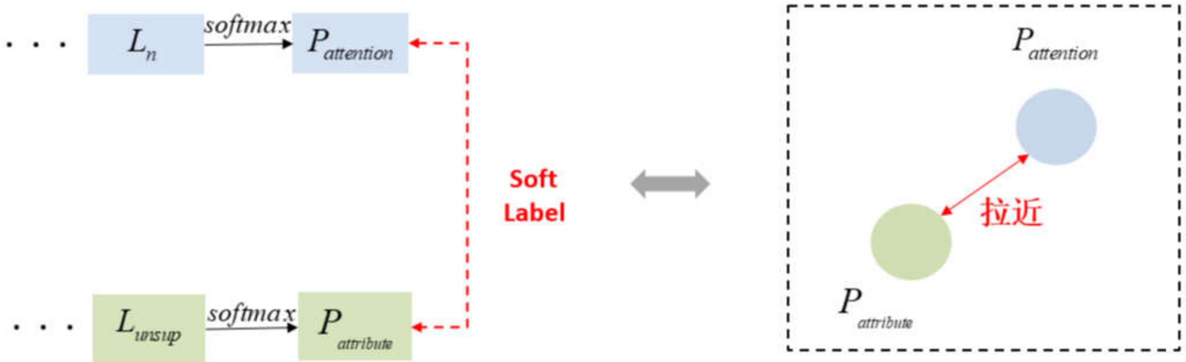
Procedure: Bilinear Aggregation

for $t = 1$ to $max - iteration$ **do**
    Random generate $h_k \in N^c$ from $\{1, \ldots, d\}^c$
    Random generate $S_k \in N^k$ from $\{+1, -1\}^k$
    Compute sketch function $\Psi\left(f_n \otimes f_a, h, s\right)$
    for $i = 1$ to n **do**
        $\Psi = [0, \ldots, 0]$
        $\Psi[h[i]] = \Psi[h[i]] + s[i] \cdot f[i]$
        return Ψ
    end for
    $\Phi = FFT^{-1}\left(FFT\left(f_n\right) \odot FFT\left(f_a\right)\right)$
end for

Supervised Learning:



Unsupervised Attribute-related Learning:



$$L_{unsup}\left(x_i, y\left(x_i\right), a_{i,j}^*, \theta\right) = -\frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{j=1}^{k}\left(a_{i,j}^*\log\left(p^*\left(x_i\right)\right)\right.$$
$$\left.+\left(1-a_{i,j}^*\right)\log\left(1-p^*\left(x^j\right)\right)\right)$$

4 Experimental result

Method	Market-1501				DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
gBiCov [27]	8.28	—	—	2.23	—	—	—	—
HistLBP [28]	9.62	—	—	2.72	—	—	—	—
LOMO [29]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
Bow [10]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
RKSL [30]	33.9	—	—	11.0	—	—	—	—
UMDL [31]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
OIM [32]	38.0	58.0	66.3	14.0	24.5	38.8	46.0	11.3
PTGAN [21]	38.6	—	66.1	—	27.4	—	50.7	—
DADM [4]	39.4	—	—	19.6	—	—	—	—
ISR [33]	40.3	—	—	14.3	—	—	—	—
PUL [23]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
UJ-AAN	**41.0**	**64.4**	**73.5**	**22.1**	**27.8**	**45.3**	**54.0**	**15.2**

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

In this paper, we presented a novel Unsupervised Joint Attention-Attribute Network (UJ-AAN) for joint learning of person re-identification attention selection and semantic attribute in an end-to-end unsupervised fashion.

Email:

gqxiao@swu.edu.cn

yangrui1994@email.swu.edu.cn

Laboratory of Digital Media and Communications of Southwest University, Chongqing, China