Image and Vision Computing 102 (2020) 103981

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

An attention-based deep learning model for multiple pedestrian attributes recognition

Ehsan Yaghoubi [a,b,*], Diana Borza [c], João Neves [d], Aruna Kumar [c], Hugo Proença [a,b]

[a] IT: Instituto de Telecomunicações, Portugal
[b] University of Beira Interior, Portugal
[c] Technical University of Cluj-Napoca, Romania
[d] TomiWorld, Portugal

ARTICLE INFO

Article history:
Received 3 April 2020
Accepted 9 July 2020
Available online 16 July 2020

Keywords:
Pedestrian attributes recognition
Multi-task learning
Visual surveillance

ABSTRACT

The automatic characterization of pedestrians in surveillance footage is a tough challenge, particularly when the data is extremely diverse with cluttered backgrounds, and subjects are captured from varying distances, under multiple poses, with partial occlusion. Having observed that the state-of-the-art performance is still unsatisfactory, this paper provides a novel solution to the problem, with two-fold contributions: 1) considering the strong semantic correlation between the different full-body attributes, we propose a multi-task deep model that uses an element-wise multiplication layer to extract more comprehensive feature representations. In practice, this layer serves as a filter to remove irrelevant background features, and is particularly important to handle complex, cluttered data; and 2) we introduce a weighted-sum term to the loss function that not only relativizes the contribution of each task but also is crucial for performance improvement in multiple-attribute inference settings. Our experiments were performed on two well-known datasets (RAP and PETA) and point for the superiority of the proposed method with respect to the state-of-the-art. The code is available at https://github.com/Ehsan-Yaghoubi/MAN-PAR-.

1. Introduction

The automated inference of pedestrian attributes is a long-lasting goal in video surveillance and has been the scope of various research works [1,2]. Commonly known as *pedestrian attribute recognition* (PAR), this topic is still regarded as an open problem, due to extremely challenging variability factors such as occlusions, viewpoint variations, low-illumination, and low-resolution data (Fig. 1(a)).

Deep learning frameworks have repeatedly been improving the state-of-the-art in many computer vision tasks, such as object detection and classification, action recognition and soft biometrics inference. In the PAR context, several models have been also proposed [3,4], with most of these techniques facing particular difficulties to handle the heterogeneity of visual surveillance environments.

Researchers have been approaching the PAR problem from different perspectives [5]; [6–8] proposed deep learning models based on *full-body* images to address the data variation issues, while [9–12] described *body-part* deep learning networks to consider the fine-grained features of the human body parts. Other works focused particularly on the *attention mechanism* [11,13,14], and typically performed additional operations in the output of the mid-level and high-level convolutional layers. However, learning a comprehensive feature representation of pedestrian data, as the backbone for all those approaches, still poses some challenges, mostly resulting from the multi-label and multi-task intrinsic properties of PAR networks.

In opposition to previous works that attempted to jointly extract local, global and fine-grained features from the *input image*, in this paper, we propose a multi-task network that processes the *feature maps* and not only considers the correlation among the attributes but also captures the foreground features using a hard attention mechanism. The attention mechanism yields from the element-wise multiplication between the feature maps and a foreground mask that is included as a layer on top of the backbone feature extractor. Furthermore, we describe a weighted binary cross-entropy loss, where the weights are determined based on the number of categories (e.g., gender, ethnicity, age, ...) in each task. Intuitively, these weights control the contribution of each category during training and are the key to avoid the predominance of some labels over the others, which was one of the problems we identified in our evaluation of the previous works. In the empirical validation of the proposed method, we used two well-known PAR datasets (PETA and RAP) and three baseline methods considered to represent the state-of-the-art.

* Corresponding author at: IT: Instituto de Telecomunicações, Portugal.
E-mail address: Ehsan.Yaghoubi@ubi.pt (E. Yaghoubi).

https://doi.org/10.1016/j.imavis.2020.103981
0262-8856/

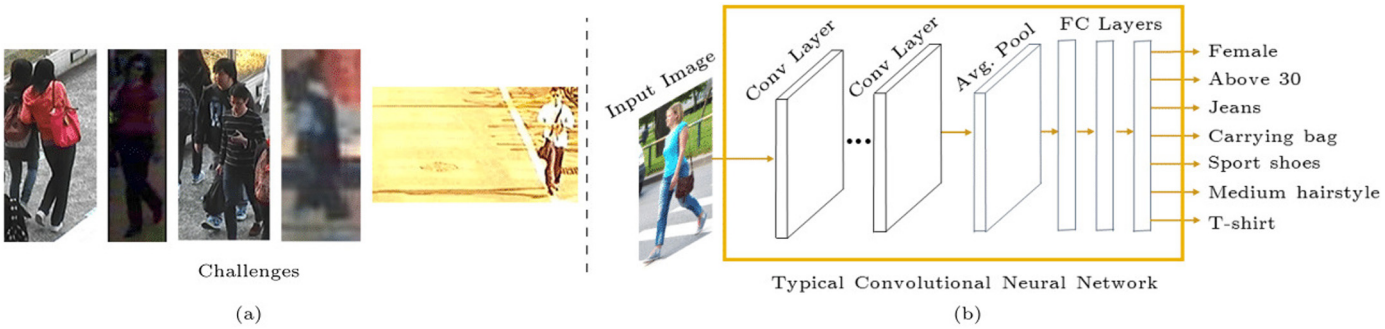

Fig. 1. (a) Examples of some of the challenges in the PAR problem: crowded scenes, poor illumination conditions, and partial occlusions. (b) Typical structure of PAR networks, which receive a single image and perform labels inference.

The contributions of this work can be summarized as follows:

1. We propose a multi-task classification model for PAR that its main feature is to focus on the foreground (human body) features, attenuating the effect of background regions in the feature representations (Fig. 2);
2. We describe a weighted sum loss function that effectively handles the contribution of each category (e.g., gender, body figure, age, etc.) in the optimization mechanism, which prohibits some of the categories to predominate over the others during the inference step;
3. Inspired by the attention mechanism, we implement an element-wise multiplication layer that simulates hard attention in the output of the convolutional layers, which particularly improves the robustness of feature representations in highly heterogeneous data acquisition environments.

The remainder of this paper is organized as follows: Section 2 summarizes the PAR-related literature, and Section 3 describes our method. In Section 4, we provide the empirical validation details and discuss the obtained results. Finally, conclusions are provided in Section 5.

2. Related work

The ubiquity of CCTV cameras has been rising the ambition of obtaining reliable solutions for the automated inference of pedestrian attributes, which can be particularly hard in case of crowded urban environments. Given that face close-shots are rarely available at far distances, PAR upon full-body data is of practical interest. In this context, the earlier PAR methods focused individually on a single attribute and used handcrafted feature sets to feed classifiers such as SVM or AdaBoost [17–19]. More recently, most of the proposed methods were based on deep learning frameworks, and have been repeatedly advancing the state-of-the-art performance [20–23].

In the context of deep learning, [24] proposed a multi-label model composed of several CNNs working in parallel, and specialized in segments of the input data. [6] compared the performance of single-label versus multi-label models, concluding that the semantic correlation between the attributes contributes to improve the results. [7] proposed a parameter sharing scheme over independently trained models. Subsequently, inspired by the success of Recurrent Neural Networks, [25] proposed a Long Short-Term Memory (LSTM) based model to learn the correlation between the attributes in low-quality pedestrian images. Other works also considered information about the subjects pose [26], body-parts [27] and, viewpoint [9,14], claiming to improve performance by obtaining better feature representations. In this context, by aggregating multiple feature maps from low, mid, and high-level layers of the CNN, [28] enriched the obtained feature representation. For a comprehensive overview of the existing human attribute recognition approaches, we refer the readers to [5].

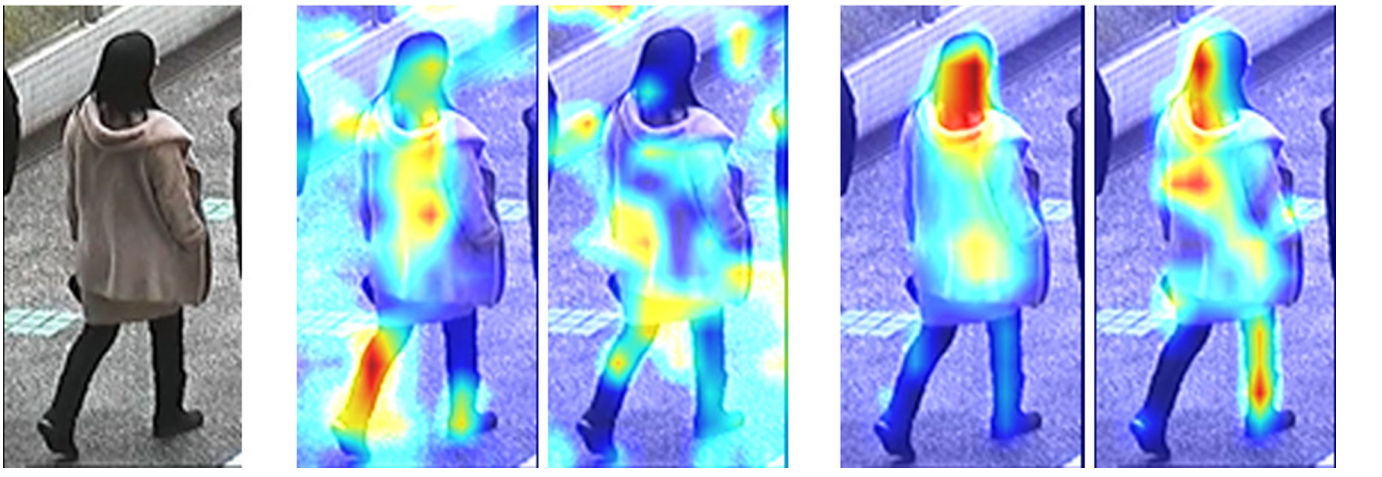

Fig. 2. Comparison between the attentive regions obtained typically by previous methods [15,16] and ours solution, while inferring the *Gender* attribute. Note the less importance given to background regions by our solution with respect to previous techniques.

3. Proposed method

As illustrated in Fig. 2, our primary motivation is to provide a PAR pipeline that is robust to background-based irrelevant features, which should contribute for improvements in performance, particularly in crowded scenes that partial occlusions of human body silhouettes occur (Figs. 1(a) and 2).

3.1. Overall architecture

Fig. 3 provides an overview of the proposed model, inferring the complete set of attributes of a pedestrian at once, in a single-shot paradigm. Our pipeline is composed of four main stages: 1) the *convolutional layers*, as general feature extractors; 2) the *body segmentation module*, that is responsible for discriminating between the foreground/background regions; 3) the *multiplication layer*, that in practice implements the attention mechanism; and 4) the *task-oriented branches*, that avoid the predominance of some of the labels over others in the inference step.

At first, the input image feeds a set of convolutional layers, where the local and global features are extracted. Next, we use the body segmentation module to obtain the binary mask of the pedestrian body. This mask is used to remove the background features, by an element-wise multiplication with the feature maps. The resulting features (that are free of background noise) are then compressed using an average pooling strategy. Finally, for each *task*, we add different fully connected layers on top of the network, not only to leverage the useful information from other tasks but also to improve the generalization performance of the network. We have adopted a multi-task network, because the shared convolutional layers extract the *common* local and global features that are necessary for all the tasks (i.e., behavioral attributes, regional attributes, and global attributes) and then, there are separate branches that allow the network to focus on the most important features for each task.

3.2. Convolutional building blocks

The implemented convolution layers are based on the concept of residual block. Considering $\boldsymbol{x}$ as the input of a conventional neural network, we want to learn the true distribution of the output $H(\boldsymbol{x})$. Therefore, the difference (residual) between the input and output is $R(\boldsymbol{x}) = H(\boldsymbol{x}) - \boldsymbol{x}$, and can be rearranged to $H(\boldsymbol{x}) = R(\boldsymbol{x}) + \boldsymbol{x}$. In other words, traditional network layers learn the true output $H(\boldsymbol{x})$, whereas residual network layers learn the residual $R(\boldsymbol{x})$. It is worth mentioning that it is easier to learn the residual of the output and input, rather than only the true output [29]. In fact, residual-based networks have the degree of freedom to train the layers in residual blocks or skip them. As the optimal number of layers depends on the complexity of the problem under study, adding skip connections makes the neural network active in training the useful layers.

There are various types of residual blocks made of different arrangements of the Batch Normalization (BN) layer, activation function, and convolutional layers. Based on the analysis provided in [30], the forward and backward signals can directly propagate between two blocks, and optimal results will be obtained when the input $\boldsymbol{x}$ is used as skip connection (Fig. 4).

3.3. Foreground human body segmentation module

We used the Mask R-CNN [31] model to obtain the full-body human masks. This method adopts a two-stage procedure after the convolutional layers: *i*) a Region Proposal Network (RPN) [32] that provides several possibilities for the object bounding boxes, followed by an alignment layer; and *ii*) a Fully Convolutional Network (FCN) [33] that infers the bounding boxes, class probabilities, and the segmentation masks.

3.4. Hard attention: element-wise multiplication layer

The idea of an attention mechanism is to provide the neural network with the ability to focus on a feature subset. Let $\boldsymbol{I}$ be an input image, $\boldsymbol{F}$ the corresponding feature maps, $\boldsymbol{M}$ an attention mask, $f_\phi(\boldsymbol{I})$ an attention network with parameters ϕ, and $\boldsymbol{G}$ an attention glimpse (i.e., the result of applying an attention mechanism to the image $\boldsymbol{I}$). Typically, the attention mechanism is implemented as $F = f_\phi(\boldsymbol{I})$, and $\boldsymbol{G} = \boldsymbol{M} \odot \boldsymbol{F}$, where $\odot$ is an element-wise multiplication. In soft attention, features are multiplied with a mask of values between zero and one, while in the hard attention variant, values are binarized and - hence - they should be fully considered or completely disregarded.

In this work, as we produce the foreground binary masks, we applied a hard attention mechanism on the output of the convolutional layers. To this end, we used an element-wise multiplication layer that receives a set of feature maps $\boldsymbol{F}_{H\times W\times D}$ and a binary mask $\boldsymbol{M}_{H\times W\times D}$, and returns a set of attention glimpses $\boldsymbol{G}_{H\times W\times D}$, in which H, W, and D are the height, weight, and the number of the feature maps, respectively.

3.5. Multi-task CNN architecture and weighted loss function

We consider multiple soft label *categories* (e.g., gender, age, lower-body clothing, ethnicity and hairstyle), with each of these including two or more *classes*. For example, the category of *lower-body clothing* is composed of 6 classes: {'pants', 'jeans', 'shorts', 'skirt', 'dress', 'leggings'}. As stated above, there are evident semantic dependencies between most of the labels (e.g., it is not likely that someone uses a 'dress' and 'sandals' at the same time). Hence, to model these relations between the different categories, we use a hard parameter sharing strategy [34] in our multi-task residual architecture. Let T, C_t, K_c, N_k be the number of tasks, the number of categories (labels) in each task, the number of classes in each category, and the number of samples in each class, respectively.

During the learning phase, the model $\mathcal{H}$ receives one input image $\boldsymbol{I}$, its binary mask $\boldsymbol{S}$, the ground truth labels $\boldsymbol{Y}$, and returns $\hat{\boldsymbol{Y}}$ as the predicted attributes (labels):

$$\hat{\boldsymbol{Y}} = \left\{ \begin{array}{c} \hat{y}_{t,c_t,k_t} \mid t \in \{1,\ldots,T\}, c \in \{1,\ldots,C_t\}, k \in \{1,\ldots,K_c\}, \\ T, C_t, K_c \in \mathbb{N}, \hat{y}_i \in 1,0, \end{array} \right\} \tag{1}$$

in which $\hat{\boldsymbol{y}}_{t,c,k}$ denotes the predicted attributes.

The key concept of the learning process is the loss function. In the single attribute recognition [35] setting, if the n-th image $\boldsymbol{I}_n$, $(n = 1,\ldots,N)$ is characterized by the m-th attribute, $(m = 1,\ldots,M)$, then $\boldsymbol{y}_{nm} = 1$; otherwise, $\boldsymbol{y}_{nm} = 0$. In case of having multiple attributes (multi-task), the predicting functions are in the form of $\Phi = \{\Phi_1, \Phi_2, \ldots, \Phi_m, \ldots, \Phi_M\}$, and $\Phi_m(\boldsymbol{I}) \in \{1,0\}$. We define the minimization of the loss function over the training samples for the mth attribute as:

$$\Psi_m = argmin_{\Psi_m} \sum_{n=1}^{N} \mathcal{L}(\Phi_m(I_n, \Psi_m), y_{nm}), \tag{2}$$

where Ψ_m contains a set of optimized parameters related to the m-th attribute, while $\Phi_m(I_n, \Psi_m)$ returns the predicted label ($\hat{\boldsymbol{y}}_{\boldsymbol{nm}}$) for the m-th attribute of the image I_n. Besides, $\mathcal{L}(.)$ is the loss function that measures the difference between the predictions and ground-truth labels.

Considering the interconnection between attributes, one can define a unified multi-attribute learning model for all the attributes. In this case, the loss function jointly considers all the attributes:

$$\Psi = argmin_{\Psi} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathcal{L}(\Phi_m(I_n, \Psi_m), y_{nm}), \tag{3}$$

in which Ψ contains the set of optimized parameters related to all attributes.

Fig. 3. Overview of the major contributions (C*i*) in this paper. C1) the element-wise multiplication layer receives a set of feature maps $\boldsymbol{F}_{H\times W\times D}$ and a binary mask $\boldsymbol{M}_{H\times W\times D}$, and outputs a set of *attention glimpses*. C2) The multitask-oriented architecture provides to the network the ability to focus on the local (e.g., head accessories, types of shoes), behavioral (e.g., talking, pushing), and global (e.g., age, gender) features (visual results are given in Fig. 7). C3) a weighted cross-entropy loss function not only considers the interconnection between the different attributes, but also handles the contribution of each label in the inference step. RCB is the abbreviation for Residual Convolutional Block, illustrated in Fig. 4. RPN, FCN, and FCL stand for Region Proposal Network, Fully Connected Network, and Fully Connected layer, respectively.

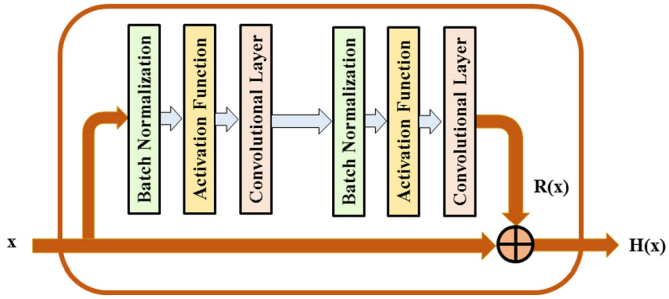

Fig. 4. Residual convolutional block in which the input **x** is considered a skip connection.

In opposition to the above-mentioned functions, in order to consider the contribution of each category in the loss value, we define a weighted sum loss function:

$$\Psi = argmin_{\Psi} \sum_{t=1}^{T} \sum_{c=1}^{C_t} \sum_{k=1}^{K_c} \sum_{n=1}^{N_k} \frac{1}{\mathcal{R}_c} \mathcal{L}(\Phi_{tck}(I_n, \Psi_{tck}), y_{tckn}), \tag{4}$$

where $\mathcal{R}_c \in \{R_1, \ldots, R_{C_t}\}$ are scalar values corresponding to the number of classes in the categories 1, ..., C_t.

Using the *sigmoid* activation function for all classes in each category, we can formulate the *cross-entropy* loss function as:

$$Loss = -\sum_{t=1}^{T} \sum_{c=1}^{C_t} \sum_{k=1}^{K_c} \sum_{n=1}^{N_k} \frac{1}{n\mathcal{R}_c} (y_{tckn}\ log(\hat{p}_{tckn}) + (1-y_{tckn})\ log(1-\hat{p}_{tckn})), \tag{5}$$

where $\boldsymbol{y}_{tckn}$ is the binary value that relates the class label k in category c. The ground-truth label for observation n and $\hat{\boldsymbol{p}}_{tckn}$ is the predicted probability of the observation n.

4. Experiments and discussion

The proposed PAR network was evaluated on two well-known datasets: the PETA [17] and the Richly Annotated Pedestrian (RAP) [15], with both being among the most frequently used benchmarks in PAR experiments.

4.1. Datasets

RAP [15] is the largest and the most recent dataset in the area of surveillance, pedestrian recognition, and human re-identification. It was collected at an indoor shopping mall with 25 HD cameras (spatial resolution 1,280 × 720) during one month. Benefiting from a *motion detection and tracking algorithm*, authors have processed the collected videos, which resulted in 84,928 human full-body images. The resulting bounding boxes vary in size from 33 × 81 to 415 × 583. The annotations provide information about the viewpoint ('front', 'back', 'left-side', and 'right-side'), body occlusions, and body-part pose, along with a detailed specification of the train-validation-test partitions, person ID, and 111 binary human attributes. Due to the unbalanced distribution of the attributes and insufficient data for some of the classes, only 55 of these binary attributes were selected [15]. Table 1 shows the categories of these attributes. It is worth mentioning that, as the annotation process is performed per subject instance, the same identity may have different attribute annotations in distinct samples.

PETA [17] contains ten different pedestrian image collections gathered in outdoor environments. It is composed of 19,000 images corresponding to 8705 individuals, each one annotated with 61 binary attributes, from which 35 were considered with enough samples and selected for the training phase. Camera angle, illumination, and the resolution of images are the particular variation factors in this set.

4.2. Evaluation metrics

PAR algorithms are typically evaluated based on the standard classification accuracy per attribute, and on the mean accuracy ($\overline{mA}$) of the attribute. Further, the mean accuracy over all attributes was also used [36, 37]:

$$\overline{mA} = \frac{1}{2M} \sum_{m=1}^{M} \left(\frac{\hat{\mathcal{P}}_m}{\mathcal{P}_m} + \frac{\hat{\mathcal{N}}_m}{\mathcal{N}_m} \right), \tag{6}$$

where m denotes one attribute, and M is the total number of attributes. For each attribute m, P_m, N_m, $\hat{\mathcal{P}}_m$, and $\hat{\mathcal{N}}_m$ stand for the number of positive samples, negative samples, correctly recognized as positive samples, correctly identified as negative samples.

4.3. Preprocessing

RAP and PETA samples vary in size, with each image containing exclusively one subject annotated. Therefore, to have constant ratio images, we first performed a zero-padding and then resized them into 256 × 256. It worth mentioning that, after each residual block, the input size is divided by 2. Therefore, as we have implemented the backbone with 4 residual stages, to multiply the binary mask and feature maps with a size of 16 × 16, the input size should be 256 × 256. Note that the sharp edges caused by these zero pads do not affect the network due to the presence of the *multiplication layer* before the classification layers.

To assure a fair comparison between the tested methods, we used the same train-validation-test splits as in [15]: 50,957 images were used for learning, 16,986 for validation purposes, and the remaining 16,985 images used for testing. The same strategy was used for the PETA dataset. Table 2 shows the parameter settings of our multi-task network.

4.4. Implementation details

Our method was implemented using Keras 2.2.5 with Tensorflow 1.12.0 backend [38], and all the experiments were performed on a machine with an Intel Core i5 — 8600K CPU @ 3.60 GHz (Hexa Core | 6 Threads) processor, NVIDIA GeForce RTX 2080 Ti GPU, and 32 GB RAM.

Table 1
RAP dataset annotations.

Branch	Annotations
Soft Biometrics	Gender, Age, Body figure, Hairstyle, Hair color
Clothing Attributes	Hat, Upper body clothes style and color, Lower body clothes style and color, Shoe style
Accessories	Glasses, Backpack, Bags, Box
Action	Telephoning, Talking, Pushing, Carrying, Holding, Gathering

Table 2
Parameter Settings for the experiment on RAP dataset.

Parameter	Value
Image input shape	256 × 256 × 3
Mask input shape	16 × 16 × 3
Learning rate	$1 \times e^{-4}$
Learning decay	$1 \times e^{-6}$
Number of epochs	200
Drop-out probability	0.7
Batch size	8

Table 3
Task specification policy for PETA and RAP datasets.

Dataset	Task 1 (full body)	Task 2 (head)	Task 3 (upper body)	Task 4 (lower body)	Task 5 (foot wears)	Task 6 (accessories)	Task 7 (action)
PETA	Female, Male, AgeLess30, AgeLess45, AgeLess60, AgeLarger60	Hat, LongHair, Scarf, Sunglasses, Nothing	Casual, Formal, Jacket, Logo, Plaid, ShortSleeves, Strip, Tshirt, Vneck, Other	Casual, Formal, Jeans, Shorts, ShortSkirt, Trousers	LeatherShoes, Sandals, FootwearShoes, Sneaker	Backpack, MessengerBag, PlasticBags, CarryingNothing, CarryingOther	–
RAP	Female, Male, AgeLess16, Age17–30, Age31–45, Age46–60, BodyFat, BodyNormal, BodyThin, Customer, Employee	BaldHead, LongHair, BlackHair, Hat, Glasses	Shirt, Sweater, Vest, TShirt, Cotton, Jacket, SuitUp, Tight, ShortSleeves, Others	LongTrousers, Skirt, ShortSkirt, Dress, Jeans, TightTrousers	Leather, Sports, Boots, Cloth, Casual, Other	Backpack, ShoulderBag, HandBag, Box, PlasticBag, PaperBag, HandTrunk, Other	Calling, Talking, Gathering, Holding, Pushing, Pulling, CarryingByArm, CarryingByHand

The proposed CNN architecture was fulfilled as a dual-step network. At first, we applied the body segmentation network (i.e., Mask R-CNN, explained in the next subsection) to extract the human full-body masks. We then trained a two-input multi-task network that receives the preprocessed masks and the input data. It is worth mentioning that, on account of the spreading or gathering nature of the attributes features in the full-body human images, we intuitively clustered all the binary attributes into 7 and 6 groups for the experiments on RAP and PETA, respectively, as given in Table 3.

As above stated, we used the pre-trained Mask R-CNN [39] to obtain all the foreground masks in our experiments. The used segmentation model was trained in the MS-COCO dataset [40]. Table 4 provides the details of our implementation settings.

Table 4
Mask R-CNN parameter settings.

Parameter	Value
Image input dimension	$1024 \times 1024 \times 3$
RPN anchor scales	32, 64, 128, 256, 512
RPN anchor ratio	0.5, 1, 2
Number of proposals per image	256

Table 5
Comparison between the results observed in the PETA dataset (mean accuracy percentage). The highest accuracy values per attribute among all methods appear in bold.

Attributes	DeepMar [15]	MLCNN [16]	Proposed
Male	89.9	84.3	**91.2**
AgeLess30	**85.8**	81.1	85.3
AgeLess45	81.8	79.9	**82.7**
AgeLess60	86.3	92.8	**93.9**
AgeLarger60	94.8	97.6	**98.6**
Head-Hat	91.8	96.1	**97.4**
Head-LongHair	88.9	88.1	**92.3**
Head-Scarf	96.1	97.2	**98.2**
Head-Nothing	85.8	86.1	**90.7**
UB-Casual	84.4	89.3	**93.4**
UB-Formal	85.1	91.1	**94.6**
UB-Jacket	79.2	92.3	**95.0**
UB-ShortSleeves	87.5	88.1	**93.4**
UB-Tshirt	83.0	90.6	**93.8**
UB-Other	**86.1**	82.0	84.8
LB-Casual	84.9	90.5	**93.7**
LB-Formal	85.2	90.9	**94.0**
LB-Jeans	85.7	83.1	**86.7**
LB-Trousers	**84.3**	76.2	78.9
Shoes-Leather	87.3	85.2	**89.8**
Shoes-Footwear	**80.0**	75.8	79.8
Shoes-Sneaker	78.7	81.8	**86.6**
Backpack	82.6	84.3	**89.2**
MessengerBag	82.0	79.6	**86.3**
PlasticBags	87.0	93.5	**94.5**
Carrying-Nothing	83.1	80.1	**85.9**
Carrying-Other	77.3	**80.9**	78.8
Average of 27 Att.	85.4	86.6	**90.0**
Average of 35 Att.	82.6	–	**91.7**

Table 6
Comparison of the results observed in the RAP dataset (mean accuracy percentage). The highest accuracy values per attribute among all methods appear in bold.

Attributes	ACN [7]	DeepMar [15]	Proposed
Female	94.06	**96.53**	96.28
AgeLess16	77.29	77.24	**99.25**
Age17–30	69.18	69.66	**69.98**
Age31–45	66.80	66.64	**67.19**
Age46–60	52.16	59.90	**96.88**
BodyFat	58.42	61.95	**87.24**
BodyNormal	55.36	58.47	**78.20**
BodyThin	52.31	55.75	**92.82**
Customer	80.85	82.30	**96.98**
Employee	85.60	85.73	**97.67**
BaldHead	65.28	80.93	**99.56**
LongHair	89.49	92.47	**94.67**
BlackHair	66.19	79.33	**94.94**
Hat	60.73	84.00	**99.02**
Glasses	56.30	84.19	**96.76**
UB-Shirt	81.81	**85.86**	83.93
UB-Sweater	56.85	64.21	**92.66**
UB-Vest	83.65	89.91	**96.91**
UB-TShirt	71.61	75.94	**77.17**
UB-Cotton	74.67	79.02	**89.48**
UB-Jacket	78.29	**80.69**	71.93
UB-SuitUp	73.92	77.29	**97.18**
UB-Tight	61.71	68.89	**96.10**
UB-ShortSleeves	88.27	90.09	**90.79**
UB-Others	50.35	54.82	**97.91**
LB-LongTrousers	86.60	**86.64**	84.88
LB-Skirt	70.51	74.83	**97.37**
LB-ShortSkirt	73.16	72.86	**98.10**
LB-Dress	72.89	76.30	**97.34**
LB-Jeans	90.17	89.46	**91.56**
LB-TightTrousers	86.95	87.91	**94.71**
Shoes-Leather	71.92	80.50	**84.00**
Shoes-Sports	62.59	71.58	**80.68**
Shoes-Boots	85.03	91.37	**96.68**
Shoes-Cloth	68.74	72.31	**98.67**
Shoes-Casual	54.57	64.58	**77.74**
Shoes-Other	52.42	61.56	**92.00**
Backpack	68.87	80.61	**98.03**
ShoulderBag	69.30	82.52	**93.29**
HandBag	63.95	76.45	**97.64**
Box	66.72	76.18	**96.30**
PlasticBag	61.53	75.20	**97.78**
PaperBag	52.25	63.34	**99.07**
HandTrunk	79.01	84.57	**97.74**
Other	66.14	**76.14**	71.54
Calling	74.66	86.97	**97.13**
Talking	50.54	54.65	**97.54**
Gathering	52.69	58.81	**95.47**
Holding	56.43	64.22	**97.71**
Pushing	80.97	82.58	**99.15**
Pulling	69.00	78.35	**98.24**
CarryingByArm	53.55	65.40	**97.77**
CarryingByHand	74.58	82.72	**87.57**
Other	54.83	58.79	**99.13**
Average	68.92	75.54	**92.23**

Table 7
Ablation studies. The first row shows our baseline system with a multi-label architecture and binary-cross-entropy loss function, while the other rows indicate the proposed system with various settings.

Multi-task architecture	Multiplication Layer	Weighted Loss (Binary-cross-entropy)	mAP (%)
–	–	–	81.11
✓	–	–	89.18
✓	–	✓	89.35
✓	✓	–	89.73

By feeding the input images to the convolutional building blocks, we obtain a set of feature maps that will be multiplied by the corresponding mask, using the element-wise multiplication layer. This layer receives two inputs with the same shapes. Transferring the input data with shape of $256 \times 256 \times 3$ into a 4-residual block backbone, we obtain a $16 \times 16 \times 1{,}024$-shaped output. Also, masks are resized to have the same size as the corresponding feature maps. Therefore, as a result of multiplying the binary mask and feature maps, we obtain a set of attention glimpses with the $16 \times 16 \times 1{,}024$ shape. These *glimpses* are downsampled to 1,024 features using a global average pooling layer to decrease the sensitivity of the locations of the features in the input image [41]. Afterward, in the interest of training one classifier for each task, a *Dense*[*ReLU*] → *DropOut* → *Dense*[*ReLU*] → *DropOut* → *Dense*[*ReLU*] → *Dense*[*Sigmoid*] architecture is stacked on top of the shared layers for each task.

4.5. Comparison with the state-of-the-art

We compared the performance attained by our method to three baselines, that were considered to represent the state-of-the-art: ACN [7], DeepMar [15], and MLCNN [16] on the RAP and the PETA datasets. These methods have been selected for two reasons: 1- in a way similar to our method, ACN and DeepMar are global-based methods (i.e., they extract features from the full-body images) 2- Authors of these methods have reported the results for all the attributes in a separate way, assuring a fair comparison between the performance of all methods.

As the solution proposed in this paper, the ACN [7] method analyzes the full-body images and jointly learns all the attributes without relying on additional information. DeepMar [15] is a global-based end-to-end CNN model that provides all the binary labels for the input image, simultaneously. In [16], authors propose a multi-label convolutional neural network (MLCNN) that divides the input image into overlapped parts and fuses the features of each CNN to provide the binary labels for the pedestrians. Tables 5 and 6 provide the obtained results observed for the three methods considered in the PETA and RAP datasets.

Table 5 shows the evaluation results of the DeepMar and MLCNN methods, including our model on the PETA dataset. According to this table, our model shows superior recognition rates for 22 (out of 27) attributes, concluded to more than 3% improvement in total accuracy. If we consider 35 attributes, the proposed network achieves a 91.7% recognition rate while this value for the DeepMar approach is 82.6%.

The experiment carried out without considering image augmentation (i.e., 5-degree rotation, horizontal flip, 0.02 width and height shift

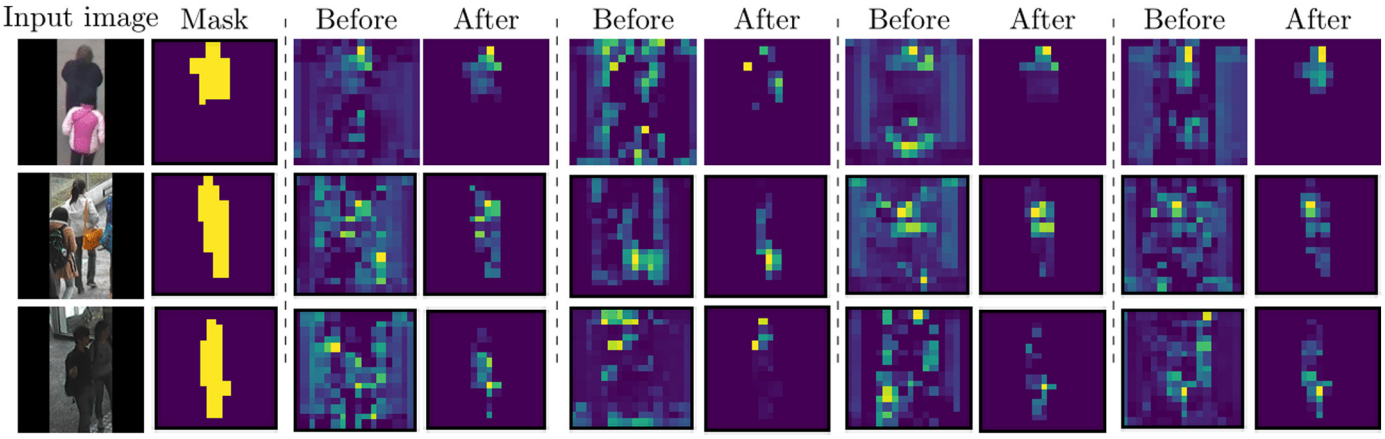

Fig. 5. The effectiveness of the *multiplication layer* on filtering the background features from the feature maps. The far left column shows the input images to the network, the *Mask* column presents the ground truth binary mask (the first input of the multiplication layer), the columns with *Before* label (the second input of the multiplication layer) display the feature maps before applying the multiplication operation, and the columns with *After* label show the output of the multiplication layer.

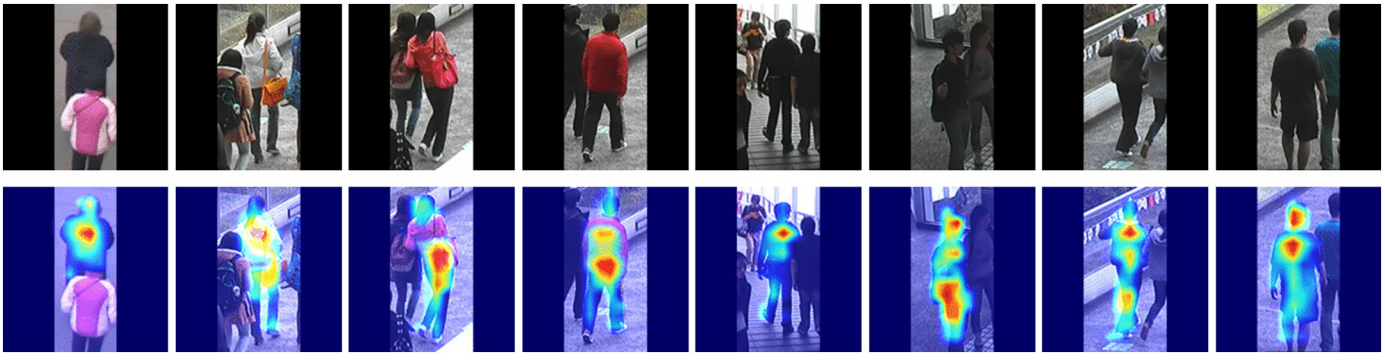

Fig. 6. Illustration of the effectiveness of the *multiplication layer* upon the focus ability of the proposed model in case of partial occlusions. Samples regard the PETA dataset, with the network predicting the *age* and *gender* attributes.

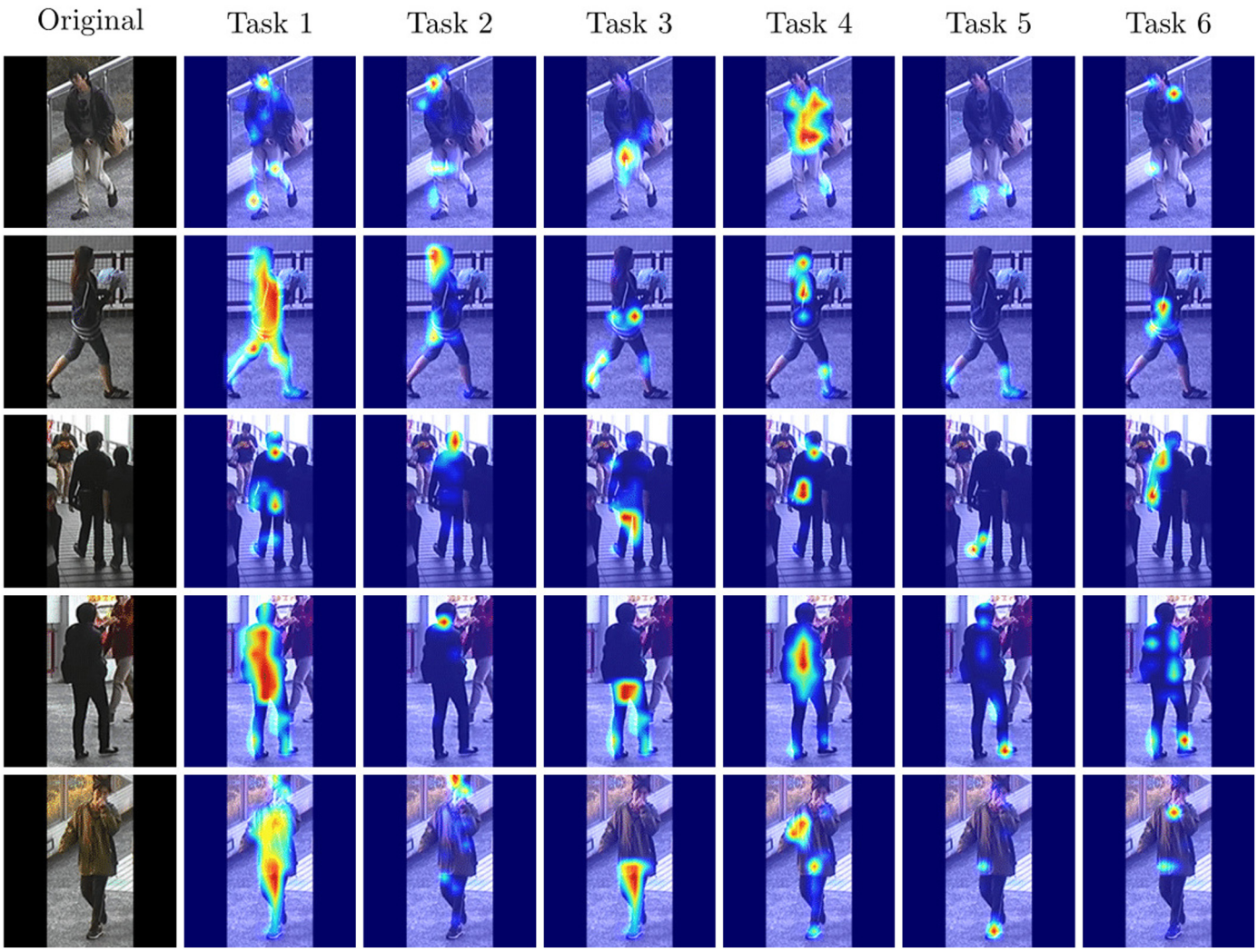

Fig. 7. Visualization of the heat maps resulting of the proposed multi-task network. Sample regard the PETA dataset. The leftmost column shows the original samples, the column *Task 1* (i.e., recognizing *age* and *gender*) presents the effectiveness of the network focus on the human full-body, and the remaining columns display the ability of the system on region-based attribute recognition. The task policies are given in Table 3.

range, 0.05 shear range, 0.08 zoom range and changing the brightness in the interval [0.9,1.1]), showed 85.5% and 88.2% average accuracy for 27 and 35 attributes, respectively. We augmented the images randomly, and after the visualization of some images, we determined the values in augmentations.

As shown in Table 6, the average recognition rates for the ACN and DeepMar methods respectively were 68.92% and 75.54%, while our approach achieved more than 92%. In particular, excluding five attributes (i.e., *Female*, *Shirt*, *Jacket*, *Long Trousers*, and *Other* class in attachments category), our PAR model provides notoriously better results than the DeepMar method, and better than the ACN model in all cases.

The proposed method shows superior results in both datasets; however, in 22 attributes of the RAP benchmark, the recognition percentage is yet less than 95%, and in 7 cases, this rate is even less than 80%. The same interpretation is valid for the PETA dataset as well, which indicates the demands of more research works in the PAR field of study.

4.6. Ablation studies

In this section, we study the effectiveness of the mentioned contributions in Fig. 3. To this end, we trained and tested a light version of the network (with three residual blocks and input image size 128×128) on the PETA dataset with similar initialization, but different settings (Table 7). The first row of Table 7 shows the performance of a network, constructed from three residual blocks with four shared fully connected layers on top, plus one fully connected layer for each attribute. In this architecture, as the system cannot decide on each task independently, the performance is poor (81.11%), and the network cannot predict the uncorrelated attributes (e.g., behavioral attributes versus appearance attributes) effectively. However, the results in the second row of Table 7 show that repeating the fully connected layers for each task independently (while keeping the rest of the architecture unchanged), improves the results by around 8%. Furthermore, equipping the network with the proposed weighted loss function (Table 7, row 3) and adding the *Multiplication layer* (Table 7, row 4) showed further improvements in the performance to 89.35% and 89.73%, respectively.

4.6.1. Feature map visualization

Neural networks are known as poorly interpretable models. However, as the internal structures of the CNNs are designed to operate upon two-dimensional images, they preserve the spatial relationships for what it is being learned [42]. Hence, by visualizing the operations on each layer, we can understand the behavior of the network. As a result of slicing the small linear filters over the input data, we obtain the

Table 8
Performance of the network trained with different loss functions on PETA dataset.

Loss function	mAP (%)
Binary focal loss function [43]	79.30
Weighted BCE loss function [6]	90.19
Proposed weighted loss function (with BCE)	90.34
Proposed weighted loss function (with binary focal loss)	89.27

activation maps (feature maps). To analyze the behavior of the proposed *multiplication layer* (Fig. 3), we visualized the input and output feature maps in Fig. 5, such that the columns labeled as *Mask* and *Before* refer to the inputs of the layer, and the columns labeled as *After* show the multiplication results of the two inputs. As it is evident, unwanted features resulting from the partial occlusions were filtered from the feature map, which improved the overall performance of the system.

4.6.2. Where is the network looking at?

As a general behavior, CNNs infer what could be the optimal local/global features of a training set and generalize them to decide on unseen data. Here, partial occlusions can easily affect this behavior and decrease the performance, being helpful to understand where the model is actually looking at in the prediction phase. To this end, we plot some heat maps to investigate the effectiveness of the proposed *multiplication layer* and *task-oriented architecture*. Heat maps are easily understandable and highlight the regions on which the network focuses while making a prediction.

Fig. 6 shows the behavior of the system regarding the examples with partial occlusions. As it is seen, the proposed network is able to filter the harmful features of the distractors effectively, while focusing on the target subject. Moreover, Fig. 7 shows the model behavior during the attribute recognition in each task.

4.6.3. Loss function

Table 8 provides the performance of the proposed network, when using different loss functions suitable for binary classification. Focal loss [43] forces the network to concentrate on hard samples, while the weighted Binary Cross-Entropy (BCE) loss [6] allocates a specific binary weight to each class. Training the network using binary focal loss function showed 79.30% accuracy in the test phase, while this number was 90.19% for the weighted BCE loss (see Table 8).

The proposed weighted loss function uses the BCE loss function, while recommends different weights for each *class*. We further trained the proposed model with the binary focal loss function using the proposed weights. The results in Table 8 indicate a slight improvement in the performance when we train the network using the proposed weighted loss function with BCE (90.34%).

5. Conclusions

Complex background clutter, viewpoint variations, and occlusions are known to have a noticeable negative effect on the performance of person attribute recognition (PAR) methods. According to this observation, in this paper, we proposed a deep-learning framework that improves the robustness of the obtained feature representation by directly discarding the background regions in the fully connected layers of the network. To this end, we described an element-wise multiplication layer between the output of the residual convolutional layers and a binary mask representing the human full-body foreground. Further, the refined feature maps were down-sampled and fed to different fully connected layers, that each one is specialized in learning a particular task (i.e., a subset of attributes). Finally, we described a loss function that weights each category of attributes to ensure that each attribute receives enough attention, and there are not some attributes that bias the results of others. Our experimental analysis on the PETA and RAP datasets pointed for solid improvements in the performance of the proposed model with respect to the state-of-the-art.

Declaration of Competing Interest

None.

Acknowledgments

This research is funded by the “Fundo Europeu de Desenvolvimento Regional (FEDER), Fundo de Coesao (FC) and Fundo Social Europeu (FSE)” under the “PT2020 - Portugal 2020” Program, “IT: Instituto de Telecomunicações” and “TOMI: City's Best Friend” with reference POCI-01-0247-FEDER-033395. Also, the work is funded by Fundação para a Ciência e a Tecnologia / Ministério da Educação e Ciência (FCT/MEC) through national funds and, when applicable, co-funded by the FEDER PT2020 partnership agreement under the project UID/EEA/50008/2019).

References

[1] A.B. Mabrouk, E. Zagrouba, Abnormal behavior recognition for intelligent video surveillance systems: a review, Expert Syst. Appl. 91 (2018) 480–491.
[2] J. Kumari, R. Rajesh, K. Pooja, Facial expression recognition: a survey, Proc. Comput. Sci. 58 (2015) 486–491.
[3] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[4] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26.
[5] X. Wang, S. Zheng, R. Yang, B. Luo, J. Tang, Pedestrian attribute recognition: a survey, arXiv (preprint arXiv:1901.07474).
[6] D. Li, X. Chen, K. Huang, Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios, 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE 2015, pp. 111–115.
[7] P. Sudowe, H. Spitzer, B. Leibe, Person attribute recognition with a jointly-trained holistic cnn model, Proceedings of the IEEE International Conference on Computer Vision Workshops 2015, pp. 87–95.
[8] A.H. Abdulnabi, G. Wang, J. Lu, K. Jia, Multi-task cnn model for attribute prediction, IEEE Trans. Multimed. 17 (11) (2015) 1949–1959.
[9] P. Liu, X. Liu, J. Yan, J. Shao, Localization guided learning for pedestrian attribute recognition, arXiv (preprint arXiv:1808.09102).
[10] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 2470–2478.
[11] Y. Li, C. Huang, C.C. Loy, X. Tang, Human attribute recognition by deep hierarchical contexts, European Conference on Computer Vision, Springer 2016, pp. 684–700.
[12] Y. Chen, S. Duffner, A. STOIAN, J.-Y. Dufour, A. Baskurt, Pedestrian attribute recognition with part-based CNN and combined feature representations, VISAPP2018, Funchal, Portugal, 2018https://hal.archives-ouvertes.fr/hal-01625470.
[13] N. Sarafianos, X. Xu, I.A. Kakadiaris, Deep imbalanced attribute classification using visual attention aggregation, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 680–697.
[14] M. S. Sarfraz, A. Schumann, Y. Wang, R. Stiefelhagen, Deep view-sensitive pedestrian attribute inference in an end-to-end model, arXiv (preprint arXiv:1707.06089).
[15] D. Li, Z. Zhang, X. Chen, K. Huang, A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios, IEEE Trans. Image Process. 28 (4) (2018) 1575–1590.
[16] J. Zhu, S. Liao, Z. Lei, S.Z. Li, Multi-label convolutional neural network based pedestrian attribute classification, Image Vis. Comput. 58 (2017) 224–229.
[17] Y. Deng, P. Luo, C.C. Loy, X. Tang, Pedestrian attribute recognition at far distance, Proceedings of the 22Nd ACM International Conference on Multimedia, MM '14, ACM, New York, NY, USA 2014, pp. 789–792, https://doi.org/10.1145/2647868.2654966http://doi.acm.org/10.1145/2647868.2654966.
[18] J. Zhu, S. Liao, Z. Lei, D. Yi, S. Li, Pedestrian attribute classification in surveillance: database and evaluation, Proceedings of the IEEE International Conference on Computer Vision Workshops 2013, pp. 331–338.
[19] R. Layne, T.M. Hospedales, S. Gong, Attributes-based re-identification, Person Re-identification, Springer 2014, pp. 93–117.
[20] Z. Tan, Y. Yang, J. Wan, H. Wan, G. Guo, S. Z. Li, Attention based pedestrian attribute analysis, IEEE Trans. Image Process.
[21] Q. Li, X. Zhao, R. He, K. Huang, Visual-semantic graph reasoning for pedestrian attribute recognition, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8634–8641.
[22] X. Zhao, L. Sang, G. Ding, J. Han, N. Di, C. Yan, Recurrent attention model for pedestrian attribute recognition, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 9275–9282.

[23] M. Lou, Z. Yu, F. Guo, X. Zheng, Mse-net: pedestrian attribute recognition using mlsc and se-blocks, International Conference on Artificial Intelligence and Security, Springer 2019, pp. 217–226.
[24] J. Zhu, S. Liao, D. Yi, Z. Lei, S.Z. Li, Multi-label cnn based pedestrian attribute learning for soft biometrics, 2015 International Conference on Biometrics (ICB), IEEE 2015, pp. 535–540.
[25] J. Wang, X. Zhu, S. Gong, W. Li, Attribute recognition by joint recurrent learning of context and correlation, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 531–540.
[26] D. Li, X. Chen, Z. Zhang, K. Huang, Pose guided deep model for pedestrian attribute recognition in surveillance scenarios, 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE 2018, pp. 1–6.
[27] L. Yang, L. Zhu, Y. Wei, S. Liang, P. Tan, Attribute recognition from adaptive parts, arXiv (preprint arXiv:1607.01437).
[28] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, Hydraplus-net: attentive deep features for pedestrian analysis, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 350–359.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 770–778.
[30] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, European Conference on Computer Vision, Springer 2016, pp. 630–645.
[31] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 2961–2969.
[32] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems 2015, pp. 91–99.
[33] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition 2015, pp. 3431–3440.
[34] S. Ruder, An Overview of Multi-task Learning in Deep Neural Networks, CoRR abs/1706.05098, arXiv:1706.05098 http://arxiv.org/abs/1706.05098.
[35] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 3730–3738.
[36] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, X. Xue, Adaptively weighted multi-task deep network for person attribute classification, Proceedings of the 25th ACM International Conference on Multimedia, ACM 2017, pp. 1636–1644.
[37] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, Y. Yang, Improving person re-identification by attribute and identity learning, Pattern Recogn.
[38] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16) 2016, pp. 265–283.
[39] W. Abdulla, Mask r-cnn for Object Detection and Instance Segmentation on Keras and Tensorflow, https://github.com/matterport/Mask_RCNN 2017.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, European Conference on Computer Vision, Springer 2014, pp. 740–755.
[41] M. Lin, Q. Chen, S. Yan, Network in network, arXiv (preprint arXiv:1312.4400).
[42] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 2980–2988.