Jointly human semantic parsing and attribute recognition with feature pyramid structure in EfficientNets

Mahnaz Moghaddam | Mostafa Charmi | Hossein Hassanpoor

1 Department of Electrical Engineering, Faculty of Engineering, University of Zanjan, Zanjan, Iran
2 Department of Computational Neuroscience, Dade Pardazi Shenakht Mehrvar Atynegar, Tehran, Iran

Correspondence
Mostafa Charmi, Department of Electrical Engineering, Faculty of Engineering, University of Zanjan, University Blvd. 45371-38791 Zanjan, Iran.
Email: Charmi.Mostafa@znu.ac.ir

Abstract
Pedestrian attributes recognition is an important issue in computer vision and has a special role in the field of video surveillance. The previous methods presented to solve this issue are mainly based on multi-label end-to-end deep neural networks. These methods neglect to apply attributes for defining local feature areas and they suffer from the problems of the bounding box presence. A new framework for jointly human semantic parsing and pedestrian attribute recognition to achieve effective attribute recognition is proposed. By extracting human parts via semantic parsing, both semantic and spatial information can be explored with eliminating of background. The framework also uses multi-scale features to employ rich details and contextual information through proposed attribute recognition-bidirectional feature pyramid network. For baseline network that has a significant impact on the performance, EfficientNet-B3 is selected as a baseline network from The EfficientNet family which provides an appropriate trade-off between the three factors of CNNs scaling (depth/width/resolution). Finally, the proposed framework is tested on datasets PETA, RAP and PA-100k. Experimental results show that our method has superior performance in both mean accuracy and instance-based metrics compared to state-of-the-art results.

1 | INTRODUCTION

In recent years, detection of pedestrian attributes such as, gender, age, backpack, clothing style etc. has attracted much attention due to their high potential for use in video surveillance applications such as person re-identification [1,2], attribute-based person retrieval [3,4] and face verification [5]. In fact, the goal is to recognize pedestrian attributes or collect the person’s attributes when the target image is given. The extracted attributes contain searchable human semantic descriptions and can be utilized as soft-biometrics in the mentioned applications.

Many attempts have been made to detect pedestrian attributes. However, this issue remains unsolved due to challenges such as viewpoint changes, brightness changes and low-resolution images. Traditional pedestrian attribute recognition methods usually focus on developing robust feature representation using handcrafted features and robust classifiers. These methods typically extract the attributes by utilizing a set of low-level features such as HOG, LBP and SIFT and using a large number of separate classifiers for each attribute. However, because of the low level of the features and also the use of separate classifiers for each class, which ignores the relationship between the attributes, these algorithms have very low performance for real-world applications.

With the advent of deep learning and especially convolutional neural networks (CNNs) that extract features automatically and efficiently using multi-layer non-linear transformations, many CNN-based methods have been developed to extract attributes. These methods perform multi-label attribute extraction in an end-to-end manner, achieving significant improvement over traditional methods [1,6][1-3]. Early algorithms for the recognition of pedestrian attributes with deep learning networks only consider the whole input image [7,8]. These algorithms try to learn very powerful feature representations in multi-task learning. Although these models are simple and perform well for practical applications, they do not use local information that is crucial for fine-grained attribute classification. Also, they do not consider the relationship between the attribute and the part where the attribute appears. For example, any attribute may only appear on a specific part, such as a hat or hair length that appears...
on the head, pants or skirts that appear on the lower-body, or glasses cannot appear on the foot.

Therefore, part-based algorithms that use local and global information jointly were introduced to improve pedestrian attribute recognition performance and consider spatial relationships between attributes. The localization of different parts of the body in these algorithms is accomplished by an external part localization module such as part detection [9], pose estimation [9] and region proposals [10]. These methods extract features from body parts that have been localized. These parts are usually different image patches that provide a simple representation of the person, and most are horizontal strips that aggregate the extracted features in these parts, which are part-based features with global features. However, despite the sophisticated calculations to localize the parts, these methods still have difficulty in showing the correspondence between the attribute and its particular region. To explore spatial context accurately, some methods [11–13] proposed attention mechanisms to extract attribute-specific local features for image representation. However, these methods can not completely cover the informative pixels due to lack of pixel-level annotations. To balance between part and attention-based approaches, flexible attribute localization modules (ALMs) [14] are proposed to constraining the attention regions to several bounding boxes. However, this method has limited capability to describe human body structures due to extracting spatial context from bounding box.

Mentioned methods generally use bounding boxes to identify different parts of the body. Bounding boxes are coarse and cannot cover the flexible nature of the human body and properly extract fine-grained information. Also, due to the presence of other objects in the background, the person may be occluded. In addition, background noise and clutters are other problems in the direct recognition of pedestrian attributes from the bounding box that reduce the accuracy and performance of the system.

Therefore, to tackle the above problems, deleting the background and the bounding box, as well as, the precise localization of the attributes can increase the accuracy and improve the performance of the pedestrian attribute recognition. Thus, this paper aims to precisely and specifically localize each attribute using human semantic parsing. To this end, we use human semantic parsing to extract contours for different parts of the body under severe pose changes, and to extract fine-grained features. Instead of using a separate pre-trained model, we jointly localize the semantic parts and extract the attributes. This work is done in ref. [15] using two separate CNNs and the use of knowledge distillation. However, using two separate CNNs will expand the network and increase the number of parameters and training time. Therefore, to avoid the above mentioned problems, we employ a pre-trained semantic parsing model to generate pixel-level annotation for attributes recognition datasets that include only image-level annotation. Hence, with the two above mentioned annotations of a dataset, we can train a backbone network to localize parts and attribute recognition in a multi-task and end-to-end manner.

On the other hand, the baseline network also has a significant impact on system performance. To select a suitable base-line network, we focus on the problem of scaling on CNNs. There are generally 3 dimensions, that is, depth, width and resolution for scaling in CNNs. Depth actually indicates the number of layers and the depth of the network. The width shows the number of channels in each layer of the convolution. The resolution reflects the resolution of the images applied to CNN. For example, firstly increasing depth by scaling-up from ResNet-50 to ResNet-200 allows for richer and more sophisticated details, or more semantic information. However, it reduces resolution which is an important factor in the accuracy of attribute recognition. Secondly, more fine-grained features can be achieved by increasing the network width. However, due to the saturation of accuracy, the width of the network cannot be increased too much. Thirdly, increasing the input image resolution increases the resolution of feature maps and extracts more detailed information. However, pedestrian attributes recognition images don’t have high resolution, and increasing their resolution also impedes accuracy saturation. Therefore, the three mentioned factors should be considered when selecting a network. The EfficientNet family [16] allows for simultaneous scaling of these three factors. Therefore, given the limitations of increasing depth, width and resolution for attribute recognition applications, we will examine the EfficientNet family to select a backbone network that has an appropriate trade-off between three factors mentioned above for attribute recognition applications.

Since the features in the deeper layers have coarse resolution and fine-grained details may not be found in these layers. So, we use pyramid structures to enrich feature maps with fine-grained details extracted from the lower layers. However, there is only one top-down path in pyramid structures implemented as FPNs which are used for attribute recognition in ref. [14]. To use low-level features efficiently, we came up with bidirectional feature pyramid network (BiFPN) [17] which provides a bi-directional path including top-down and bottom-up. We then proposed attribute recognition-BiFPN (AR-BiFPN) module to improve attributes recognition performance with having jointly low-level details and high-level semantics.

The contributions of this paper are as follows:

1. To achieve the reinforcement of the features representations and localization, human semantic parsing has been jointly used with pedestrian attribute recognition in a network, which causes to learning the relationship between each attribute and its region accurately without background.
2. Baseline network resolution plays an important role in the efficiency of attribute recognition and network resolution depends on two parameters, that is, network depth and width. So, we have examined the EfficientNet family to choose an efficient backbone network.
3. To use low-level details and high-level semantics, AR-BiFPN pyramid structure is presented in the output of EfficientNet, which has led to accurate localization of attributes due to simultaneous attention to semantic parts and multi-scale features.
2 RELATED WORKS

2.1 Pedestrian attribute recognition

In early works on pedestrian attribute recognition, low-level features have been adopted and multiple binary classifiers have been trained independently for each attribute. Layne et al. [1] have proposed training of support vector machines classifiers for attributes recognition and then employ the results to assist person re-identification. Zhu et al. [18] applied the Gentle AdaBoost algorithm for attribute recognition by extracting multiple features, for example, HSV and HOG histograms from the upper-body and lower-body parts. AdaBoost algorithm was utilized for feature selection and a weighted k-NN classifier for classification. Recently, deep learning is commonly used in the literature to solve the problem of pedestrian attribute recognition because of its great power in feature learning. As shown in Table 1, deep learning approaches for pedestrian attribute recognition can be divided into five domains, including: global-based, attention-based, part-based, flexible localization-based and parsing-based methods.

2.1.1 Global-based

Early works used a holistic CNN model for jointly classification of multiple attributes. These models did not address local features. In ref. [19], the authors introduce two DeepSAR and DeepMAR algorithms based on global features. In DeepSAR, they use AlexNet as the base network for feature extraction and adapt the Softmax loss classification to calculate the final cost. Since DeepSAR could not account for the relationship between the attributes, they proposed DeepMAR which receives images and labels simultaneously and jointly considers all attributes via weighted sigmoid cross-entropy. Zhou et al. [20] have provided a pyramidal spatial pooling for jointly consideration of attribute recognition and localization. In ref. [21], a multi-task learning algorithm called MTCNN is presented for estimating human attributes using rich group information. MTCNN allows CNN models to share visual knowledge between different groups of attributes. However, the performance of these methods is limited due to the lack of prior information on the attributes and fine-grained recognition.

2.1.2 Part-based

To overcome global-based approach problems, some research have introduced part-based algorithms that share global and local information jointly. Zhu and et al. [22] first subdivide the human body into 15 rigid parts to extract local features. Then, a separate CNN model is trained for each part and the prior fully connected layer is used to merge the different parts. Diba and et al. [23] proposed a new CNN called DeepPattern that extracts mid-level image patches. To learn distinct patch groups, they specifically train proposed CNN to fused local patch features with global features. Liu and et al. [10] discovered human attributes in a weakly supervised manner. They considered the proposals produced by EdgeBoxes [24] as the parts of each attribute. This method is not consistent and cannot be trained in end-to-end manner. Part-based approaches either rely on predefined rigid parts or utilize a complex part localization mechanism based on bounding boxes and the bounding boxes are coarse annotations, thus may have limited capability to describe some fine-grained details.

2.1.3 Attention-based

Liu and et al. [11] have presented a multi-directional attention model that includes a main CNN network and an attentive feature net. Attentive feature net contains several branches of multi-directional attention modules to apply at different levels of semantic features. Sarfraz and et al. [13] used view cues to better estimate attributes. The authors believed that visual cues that point to attributes could be highly localized to better infer human attributes such as hair, backpacks, shorts and so on. The algorithm presented in ref. [12] for attribute recognition can be considered an ensemble method that involves multi-scale visual attention and weighted focal cost for deep imbalanced classification. The authors adapt feature maps from different layers for this task. Although, the accuracy of the attributes recognition has improved with these methods, they could fail to consider specific information for each attribute. Indeed, they couldn’t completely consider the informative pixels due to they have only access to image-level annotations.
2.1.4 | Flexible localization-based

In the following, Tang et al. [14] proposed an algorithm in an attribute-specific manner while attempted to balance between part and attention-based approaches by constraining the attention regions to several bounding boxes. This algorithm performs localization without region annotations and proposes a flexible ALM that can automatically discover discriminative regions. They also embed multiple ALMs at different feature levels and introduce a feature pyramid network (FPN) by integrating high-level semantics to reinforce the attribute. This method by applying ALMs properly localizes attributes at different feature levels and achieves significant improvement on recognition of fine-grained attributes, for example, baldhead and hat. However, coarse attributes such as the body fat and types of clothing that take the big area have a lot of background information. Thus, this method has a limited capability to describe fine-grained details in recognition of coarse attributes.

2.1.5 | Parsing-based

To overcome the limited capability of mentioned methods in describing human body structures and eliminating background noise, Li et al. [15] introduced a parsing-based method to recognize attributes by joint visual-semantic reasoning (JV-SR) and knowledge distillation. They train two separate models for human parsing and attribute recognition in a graph-based reasoning which models the dependencies between attribute groups. However, using two separate CNNs will expand the network and increase the number of parameters and training time. Also, this method does not apply lower layers features and use just high-level features, thus it cannot capture rich details.

Our proposed method is inspired by the JV-SR method, however we employ a pre-trained semantic parsing model to generate pixel-level annotation for attributes recognition datasets, then train a backbone network to jointly localize attributes parts by semantic parsing and recognize attributes.

We implement this issue in a multi-task and end-to-end manner. Moreover, inspired by BiFPN [17], we introduce a new feature pyramid architecture to reinforce the features.

2.2 | Human semantic parsing

In the literature, various aspects of human semantic parsing have been discussed. Some of them considered pose estimation and human body parsing as a multi-task learning problem [25,26]. In ref. [27], authors have contributed edge prediction with human semantic parsing to predict the boundary area accurately. Many previous works in this field assume that ground truth labels are well interpreted and correct. Because of the time-consuming and costly production of ground truth, there may inevitably exist a lot of label noise.

With this intuition, the paper [28] proposes a self-correction structure to solve this problem and provides a framework for human semantic parsing. This structure uses boundary information along with parsing information to refine the labels and has achieved very good results. Therefore, we will use the pretrained model of this structure to generate pixel-level annotations for the human attribute recognition datasets. Due to its accuracy and high performance, using this model to extract pixel-level annotation can be both time-saving and somewhat accurate compared to manual extraction.

2.3 | Multi-scale feature representations

FPN is one of the pioneering works to combine multi-scale features, which offers a top-down path for utilizing different scale features. In the following of this idea, PNet [29] is presented which adds an extra bottom-up path aggregation network to the top of the FPN. M2det [30] is another proposed method in this field that proposes a U-shaped module to fuse multi-scale features. The NAS-FPN [31] leverages neural architecture search for automated feature network topology design. Although this method achieves better performance, it requires thousands of hours of GPUs during the search, and the resulting features network is irregular and difficult to interpret. Recently, BiFPN [17] has been proposed for object detection that uses a two-way, bottom-up and top-down path to represent features in multi-scale. So, we adapt BiFPN to apply multi-scale features in pedestrian attribute recognition and recommend AR-BiFPN.

3 | METHODOLOGY

3.1 | Framework review

This paper proposes a framework for pedestrian attributes recognition that explores both semantic and spatial information. It also uses multi-layer outputs for employing rich details and contextual information through feature pyramid architectures. As is shown in Figure 1, the proposed structure has two branches, including human parsing and multi-scale feature representation. A common backbone network, one of the EfficientNet family models, was used to extract feature maps in both branches. After the input image is fed to the network, feature maps first are extracted. Then, using these feature maps, the parsing branch generates six semantic regions of the body, namely the foreground, head, upper-body, hands, lower-body and shoes. Finally, attributes are recognized in the semantic regions using the pyramid feature maps. We have used EfficientNet as a backbone network to extract feature maps because it achieves a good trade-off between the three factors of depth, width and resolution, which significantly impact on improving the accuracy of attribute recognition. In the following, we first briefly describe the EfficientNet family architecture, parsing and AR-BiFPN branches, then explain how to combine feature maps extracted by AR-BiFPN with semantic regions.
3.2 EfficientNet structure

CNNs are usually expanded at the cost of fixed resources and then scaled to achieve higher accuracy. For example, ResNet can be scaled up by using more layers from ResNet-18 to ResNet-200. The scaling methods are mainly implemented by using high-resolution images, and increasing their depth and width. Even though with improved accuracy, these methods always require tedious manual regulation, but have sub-maximum performance.

In ref. [16], the authors have proposed a scaling method for using CNNs in a more structured way that utilizes a simple compound coefficient. This method scales each dimension with a set of scaling coefficients. While scaling individually improves model performance, but balancing grid dimensions (width, depth and resolution) with respect to available resources improves overall performance in the best possible way. In the first step of the compound scaling method, a grid search is performed to find the relationship between various scaling dimensions of the baseline network under constant source constraint (e.g. 2x more FLOPS). This search determines the appropriate scaling coefficient for each dimension and then uses these coefficients to scale up the base network as much as the target model or computational budget. Using this new scaling method, a family of models, called EfficientNets [16], has been proposed which is extremely accurate with better efficiency (smaller and faster).

Mobile inverted bottlenecked convolution (MBConv) is utilized in the EfficientNets base structure similar to MobileNetV2 [32], which is larger here due to the increased FLOP budget. Using this block, seven operational steps are implemented and the base network, which is the EfficientNet-B0 model, is obtained. Then, the base network is scaled up to produce all models of the EfficientNets family. As is shown in Figure 1, we obtained feature maps by applying one of the EfficientNet models. Afterwards, we use the final output of the five last operational steps, namely p3 to p7, to generate rich feature maps that include both detail and high levels semantics. We also used p4 and p5 to generate 6 semantic regions of the human body.

3.3 Human semantic parsing model

As described in Section 2.1, several methods attempt to extract local features by applying external part localization modules or predefined rigid parts to exploit local visual cues. However, these methods are not the optimal solution because they neglect the importance of specific localization for each attribute. Therefore, we use probability maps of 6 different semantic regions of the body, namely the foreground, head, upper-body, hands, lower-body and shoes, to search for each attribute in its specific location and apply local visual cues. These maps are produced using the human semantic parsing branch. Obviously, localization of parts using semantic parsing would be superior to applying parts extracted from bounding boxes. Because the background is eliminated in the semantic parsing and the pixel-level precision is performed. Also, it is more robust against pose variation and occlusion.

Due to limitations of the resolution in the surveillance images and the experiment results, we select EfficientNet-B3 as the backbone. Since there are seven MBConv blocks in the EfficientNets structure, we use feature maps produced by the sixth and seventh stages of EfficientNet-B3, for human semantic parsing, which have the same resolution. The sixth stage is the last stage where the resolution is reduced. If the resolution of input image is $300 \times 300$, the output resolution of stages 6 and 7 will be $9 \times 9$. Since the quality of semantic parsing depends heavily on the final activation being of sufficient resolution, so
we first up-sample the concatenation result of \( p_6 \) and \( p_7 \) that we call \( p_{67}^p \), and then apply it to semantic parsing branch. According to the standard approach \[33\], we implement a semantic branch with an atrous spatial pyramid pooling (rates = 1, 3, 6, 9) and follow it by a 1 × 1 convolution, and finally we extract six semantic regions by applying a pixel-level Softmax function.

### 3.4 Proposed feature pyramid structure

In ref. \[14\], a method inspired by FPN models (Figure 2(a)) has been introduced for applying the feature pyramid architecture to improve pedestrian attributes recognition.

This method performs localization of features and features learning over the region in a mutually reinforcing way. A conventional FPN aggregates multi-scale features in a top-down manner and is constrained by the one pathway flow of information. But the BiFPN (Figure 2(b)), proposed in ref. \[17\], provides a bottom-up and top-down bidirectional pathway for feature extraction on multi-scale. Therefore, to obtain rich multi-scale features for pedestrian attributes recognition, we draw on the BiFPN approach and implement a bottom-up and top-down pathway to aggregate basic network output maps. We call our proposed structure AR-BiFPN (Figure 2(c)) which can capture more fine-grained features.

As stated in the previous section, there are seven operating stages in the EfficientNets family, where resolution of feature maps remains constant at each stage. In BiFPN (Figure 2(b)) proposed in ref. \[17\] for object detection, the output feature maps produced in the last block of stages 3 to 7 namely \( p_3^{in}, ..., p_7^{in} \) are used to construct the feature pyramid. The reason for choosing the last blocks is that their features are stronger. To build AR-BiFPN (Figure 2(c)), we first concatenate the feature maps from stage 4 with stage 5, as well as, stage 6 with stage 7. These stages have the same resolution, so they don’t need to be up-sampled for concatenation with each other. It also reduces the computation cost and simplifies the pyramid structure. Then, we use concatenation operation to aggregate features at different levels instead of the element-wise addition used in BiFPN, which causes some information to disappear.

Finally, we create the feature pyramid in three levels along the top-down and bottom-up paths.

The top-down pathway consists of 2 top-down and 3 lateral connections. Each feature map is fused with its adjacent feature map through top-down connections. If resolutions are different at adjacent levels, the up-sampling operation is also performed. Thus, the fusing of features in the top-down pathway is described as follow:

\[
\begin{align*}
\hat{p}_{67}^{mid} &= \text{MBConv}(f(p_{67}^{in})) \\
\hat{p}_4^{mid} &= \text{MBConv}(p_{4,5}^{in}), \text{up}(p_{67}^{mid}) \\
p_3^{mid} &= \text{MBConv}(f(p_3^{in}))
\end{align*}
\]

where \( f \) is a 1 × 1 convolution and “up” is also an up-sampling operation using the nearest neighbour interpolation. The up-sampling operation makes similar the resolution of the two feature maps that will be fused together. Next, the bottom-up pathway is implemented to fuse the features obtained from the top-down pathway. In this pathway, at each level, the features obtained in the top-down pathway are fused with their higher level features. Also at each level, an additional edge from the original input is added to the output node to fuse more features without the need for high cost.

At this level, the feature fusion is as follows:

\[
\begin{align*}
p_3^{out} &= \text{up}(\hat{p}_3^{mid}), (p_3^{mid}) \\
p_4^{out} &= \text{up}(\hat{p}_4^{out}), f(p_{4,5}^{in}), p_4^{mid} \\
p_{6,7}^{out} &= \{\text{up}(p_{4,5}^{out}), p_{6,7}^{mid}\}
\end{align*}
\]

### 3.5 Attribute recognition in semantic region

To take advantage of local visual cues and search for each attribute in its specific location, we use feature maps of six different semantic regions of the body. The semantic parsing
model is used to generate these feature maps. We first pool outputs activations of the AR-BiFPN module several times by utilizing these generated feature maps. This can be performed as a weighted sum function between AR-BiFPN outputs and feature maps of semantic regions where feature maps are used as weights. Such a method produces six feature vectors in each level, each vector representing exclusively one region of the human body. Next, we perform element-wise max operation in the head, upper-body, lower-body, hands and shoes representations, and then concatenate the obtained result with the foreground representation in each levels. Subsequently, three fully-connected layers are applied to estimate the prediction of attribute in three levels. Finally, outputs of three fully-connected layers from region aware attribute recognition part and one fully-connected layer from main net are aggregated through an element-wise maximum operation for the most accurate attribute recognition.

3.6 Cost function

The entire network is trained in end-to-end manner using two types of training cost functions. The first relates to the pedestrian attribute recognition and the second relates to human semantic parsing. For the first cost function, we apply the weighted binary cross-entropy loss function [19] and the deep Supervision mechanism [34] for training. As can be seen in Figure 1, there are 4 individual prediction vectors, 3 of which belong to region aware attribute recognition part, and one to the main network. The maximum voting scheme is utilized to select the best predictions among these 4 levels of attribute recognition. Therefore, we use four weighted binary cross-entropy loss function for each of the individual prediction vectors to be directly supervised by ground truth labels. These loss functions formulated as follow:

\[
I_{\text{cls}}(\hat{y}, y) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{a \in \{1, 2, 3, 4\}} \left( \hat{y}^a_n \log \left( \sigma \left( \hat{y}^a_n \right) \right) + (1 - \hat{y}^a_n) \log \left( 1 - \sigma \left( \hat{y}^a_n \right) \right) \right),
\]

where \(N\), \(\hat{y}^a_n\) and \(a_n\) are the number of attributes, the loss weight for the \(n\)-th attribute and the prior class distribution of the \(n\)-th attribute, respectively. The \(i\)-th branch is represented by \(i \in \{1, 2, 3, 4\}\). The sigmoid activator is denoted by \(\sigma\). With sum of these 4 loss functions \(I_{\text{att}}\) is calculated as follow:

\[
I_{\text{att}} = \sum_{i=1}^{4} I_{\text{cls}}(\hat{y}, y)
\]

where \(I_{\text{cls}}\) is the weighted binary cross-entropy loss function for each prediction vector.

The second loss function is related to the semantic parsing branch that we use negative log-likelihood. If for an image \(I\), the ground truth label of human semantic parsing be \(h^m_{\text{hm}}\) and \(\hat{h}^m_{\text{hm}}\) is assumed prediction of semantic parsing, where \(m\) is the number of pixels for the \(k\)-th class. We define the pixel-wise classification loss as follows.

\[
I_{\text{cls}} = -\frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \hat{y}_{mk}^m \log p \left( \hat{h}_{mk}^m \right),
\]

where \(M\) is the number of pixels in the image \(I\) and \(K\) is the number of classes. Thus, the final loss function is calculated as follow:

\[
L = \lambda_1 I_{\text{att}} + \lambda_2 I_{\text{cls}}
\]

That \(\lambda_1\) and \(\lambda_2\) are hyper-parameters to control the share of each loss. By minimizing \(L\), we jointly train the model for human semantic parsing and attribute recognition in an end-to-end manner.

4 EXPERIMENTS

4.1 Datasets and evaluation metrics

Our proposed methods are evaluated on three public datasets as: PETA [6], RAP [35] and PA-100K [11]. PETA: The dataset contains 19,000 images with 61 binary attributes and 4 multi-class attributes. The total dataset is randomly partitioned into 3 non-overlapping subsets, which include: 9500 images for training, 1900 for validation and 7600 images for testing. In this dataset, 35 attributes which the positive ratio is higher than 5% are used for evaluation. RAP: The dataset contains 41585 images collected from an indoor camera network. Each image is labelled with 69 binary attributes and 3 multi-classes attributes. According to the official protocol [35], the total dataset is partitioned into 33,268 training images and 8317 test images. PA-100k: The dataset is by far the largest dataset to recognize pedestrian attributes, with a total of 100,000 pedestrian images collected from outdoor surveillance cameras. Each image has 26 common attributes. According to the official protocol [11], the total dataset is randomly partitioned into 80,000 training images, 10,000 validation images and 10,000 test images.

4.1.1 Implementation details

As mentioned before, our proposed structure uses a network to jointly parsing human semantics parts and to extract attributes in semantic segments. So to train the network for two tasks simultaneously, in addition to attribute recognition annotation, we need to generate segmentation annotations for the three PETA, PAP and PA-100k datasets so that we can use these annotations to calculate the segmentation cost expressed in Section 3.6. Therefore, we have used the structure presented in the paper [28], the SCHP model, which has recently achieved very good results in the field of semantic segmentation, as a benchmark for producing segmentation annotations for three attribute recognition datasets. This structure is evaluated on two single human parsing datasets including LIP and PASCAL-
Table 2: Extracting pedestrian attributes with different EfficientNet models and comparing it with ResNet50 and Inception-v3 (The best results are bolded and the second best ones are marked with an asterisk)

| Metric | #Params | #FLOPS | Input resolution | Model       |
|--------|---------|--------|------------------|-------------|
| F1     | Recall  | Pre    | Acc              | mA          |
| 85.95  | 85.52   | 86.46  | 78.16            | 83.95       |
| 86.23  | 86.01   | 86.57  | 78.63            | 84.84       |
| 86.25  | 85.98   | 86.62  | 78.88            | 84.94       |
| 87.36  | 86.19   | 86.63  | 78.97            | 85.06       |
| 87.41* | 86.21   | 87.88* | 79.07            | 85.03       |
| 85.68  | 81.90   | 89.70  | 77.73            | 81.50       |
| 85.90  | 84.52   | 87.32  | 78.05            | 83.49       |
| 87.46  | 86.25   | 87.93  | 79.01*           | 85.15       |

Person-Part. We selected the mode set for LIP dataset. This dataset contains pixel-level annotations and 19 semantic human parts labels. So, we have applied human attribute recognition images to this structure and it generated 19 semantic human parts labels for each image, which we combined some parts with each other such as hands and arms to reduce the training problems, and produced 6 semantic human parts labels. In the following, we briefly looked at the annotations produced and manually corrected the annotations that needed to be corrected, and finally produced the final ground truth for segmentation.

4.1.2 Evaluation metrics

For the evaluation process, some metrics from the label-based and instance-based metrics are used in this paper. The mean Accuracy (mA) as a label-based metric is the mean value of positive and negative accuracies for each attribute. It is formulated as follows:

$$mA = \frac{1}{2N} \sum_{i=1}^{N} \left( \frac{TP_i}{P_i} + \frac{TN_i}{N_i} \right),$$

where \(N\) and \(S\) are the number of attributes and examples, respectively. \(P_i\) is the number of positive examples and \(TP_i\) is correctly predicted positive examples of the \(i\)-th attribute; \(N_i\) and \(TN_i\) are defined similarly.

The considered instance-based metrics are accuracy, recall rate, precision and F1 score. For accuracy, precision and recall rate, the scores of the predicted attributes against the ground truth are first computed for each instance and then averaged over all test images. The F1-score is computed based on precision and recall.

4.2 Experiment results

4.2.1 Baseline network selection

The EfficientNet structure allows for simultaneous scaling of the three factors of resolution, depth and width of the network for better performance. Therefore, we consider the EfficientNet family to choose the baseline network. Due to the low resolution of surveillance images, we only perform the recognition of attributes with the first 5 models of the EfficientNet family, that is, B0 to B4. Thus we train an EfficientNet based DeepMAR model with conventional input resolution is used in other attribute recognition papers, that is, the length of the input image is twice its width for fair comparison. We used their ImageNet-pretrained checkpoints. In addition, to compare EfficientNet results with other CNNs, we performed human attribute recognition with Inception-V3 and ResNet-52. These two models give better results in human attribute recognition than almost any other CNN with less than 30 M parameters. The results of the human attribute recognition by these models are shown in Table 2. The best results are bolded and the second best ones are marked with an asterisk (*). As can be seen in this table, in all metrics except precision, EfficientNets perform better, especially EfficientNet-B3 achieves 1.88% better result in the mA than ResNet-50, while almost the number of parameters and FLOPs in EfficientNet-B3 is half ResNet-50.

Also, in 3 Instance-base metrics, namely, accuracy, Recall and F1, the EfficientNet-B3 model compared to ResNet-50 have 1.18%, 1.97% and 1.67% better performance, respectively. Compared to Inception-V3, only the precision metric is lower, which is due to the high cost of calculations in Inception-V3, so that the FLOPs on this network is almost 3 times higher than EfficientNet-B3. As this table shows, using EfficientNet-B3, the number of parameters and FLOPS in the backbone network and then the whole structure will be less. This is because we used depth-wise separable convolutions instead of the standard convolution in implementing the AR-BiFPN and the atrous spatial pyramid pooling structures. The computational cost of a standard convolution for a feature map with dimensions \(D_s \times D_v \times P\) and Q filters is \(D_s^2PD_v^2Q\), where \(D_s, D_v\) and \(P\) are height, weight and channel, respectively. While the computational cost of the depth-wise convolution is \(D_s^2D_v^2P + PD_v^2Q\), this is significantly lower than the standard convolution.

On the other hand, in semantic parsing, the resolution of the feature map must be high, so we set the input image resolution for EfficientNet-B3, 300 x 300, in other words, we doubled the width of the input image and recognized the human attribute once again with this resolution. As is shown in Table 3, the
### TABLE 3
Investigating the change in resolution on attribute recognition performance with EfficientNet-B3 (The best results are bolded)

| Metric  | F1    | Recall | Pre | Acc | mA    | Input resolution | Model       |
|---------|-------|--------|-----|-----|-------|------------------|-------------|
|         | 87.36 | 86.19  | 86.63 | 78.97 | 85.06 | 300 × 150        | EfficientNet-B3 |
|         | 87.46 | 86.25  | 87.93 | 79.01 | 85.15 | 300 × 300        |             |

### FIGURE 3
The variations of mA and accuracy on the validation set of PETA when parameter $\lambda_1$ varies while $\lambda_2$ was set to 1

performance will be improved slightly with this resolution, but doubling the width of the input image will improve the accuracy of the semantic parsing. Therefore, to implement the proposed structure, we set the input image resolution to $300 \times 300$.

#### 4.2.2 Ablative study of parameters

To find the optimal value of two key parameters $\lambda_1$ and $\lambda_2$, we set $\lambda_2 = 1$ and then train the proposed model by varying $\lambda_1$ to get the best value for it. Only PETA dataset is utilized for training the model due to save the computation cost. Attribute recognition results on the validation set of PETA is presented in Figure 3. When $\lambda_1$ is small, the impact of semantic parsing is high so that the network reduces to the semantic parsing model while when $\lambda_1$ becomes larger, person attribute recognition will exert more influence, and thus can approximate Baseline without semantic parsing branch. There is a good trade-off between mA and Accuracy values when $\lambda_1 = 7$, so we set the $\lambda_1 = 7$ and $\lambda_2 = 1$ to evaluate the proposed model and compare it with state-of-the-art works.

#### 4.2.3 Comparison with different state-of-the-art methods

In this subsection, performance of the proposed method is compared with 8 state-of-the-art methods. DeepMAR [19] is based on CNN models and receives images and labels simultaneously and jointly considers all the attributes using the weighted sigmoid cross-entropy loss in the learning process. GRL [36] proposes a grouping recurrent learning model that uses intra-group mutual exclusion and inter-group correlation. HP-Net [11] includes a standard CNN main network and an attentive Feature Net. Attentive feature net is made up of several branches of attention modules with several directions for applying semantic features at different levels. VeSPA [13] uses viewpoint cues to better estimate attributes. In a weakly supervised manner, LG-Net [10] detects attribute parts and uses pre-generated proposals from EdgeBoxes as attributes areas. PGDM [9] uses an external pose estimation module to localize different parts. MS-ALM [14] proposes an ALM to discover regional features and introduces feature pyramid architecture. JV-SR [15] provides a graphical global reasoning framework for join modelling of visual-semantic relations and the use of knowledge distillation to guide learning. Table 4 shows the results of attribute recognition in three datasets with our proposed method and 8 state-of-the-art methods. The results suggest that our proposed structure in all three datasets yields superior performance over the state-of-the-art works in both label-based and instance-based metrics. In all datasets, label-based metric, that is, mA, has the best results, so that compared to the second best results, which is GRL in PETA and MS-ALM datasets in RAP and PA-100k datasets, it improves by 1.15%, 0.6% and 0.9%, respectively. Also, the our proposed structure in PETA dataset performs best in all Instance-based metrics except precision, and the second best result in Acc, Recall and F1 metrics is for PGDM and GRL and JV-SR respectively. The best result in the precision in the PETA dataset is for the JV-SR method, while our method has the second best results. This is because of the JV-SR uses a separate network to extract body parts and learn the dependencies between attribute groups. So, its precision is slightly higher than our model, which is obtained in exchange for increasing the parameters and higher computational cost. By the way, due to the lack of multi-scale features, mA of this method is relatively low. Furthermore, the two parameters precision and recall are inversely related to each other, so that one decreases with the increase of the other. As can be seen in Table 4, the improvement of precision in the JV-SR method has caused recall to drop to 4th place. In both RAP and PA-100k datasets, the results are similar to PETA datasets. In these datasets, the best results in accuracy, recall and F1 metrics is related to our proposed method and the best second results in accuracy and F1 is related to JV-SR and in recall is related to MS-ALM. By embedding ALMs at different feature levels, MS-ALM achieved a significant improvement on recognition of fine-grained attributes thus it has good result in label-based metric. However, this method cannot completely
describe human non-rigid structure due to using bounding boxes. Also, using multiple bounding boxes in multi-scale manner increase background problems, so that the instance-based metrics especially precision has not good result in this work.

4.2.4 Effect of attribute recognition-bidirectional feature pyramid network

In this section, we examine the effect of the proposed feature pyramid architecture, AR-BiFPN, by comparing it with FPN and BiFPN. The Baseline network that these three pyramid architectures are constructed in its output is DeepMAR, which has been implemented with EfficientNet-B3. In order construction of AR-BiFPN, we first concatenated the same resolution feature maps, that is, p7 with p6 and p5 with p4, and then implemented the pyramid structure. In the following, we have used concatenate instead of addition element-wise to transfer feature maps to adjacent levels compared to BiFPN.

As Table 5 shows, the most improvement is for the label-base metric, mA with the AR-BiFPN improved by 2.32% compared to baseline network. It also has been improved by 1.03% and 0.69% compared to FPN and BiFPN, respectively. Because regular FPNs only add multi-scale features from top to bottom, so they are limited by the one-way flow of information. In BiFPN, although the features are aggregated in two directions, top-down and bottom-up, some essential information disappears due to the use of element-wise addition to aggregate the features at the adjacent level.

5 CONCLUSION

In this paper, we investigated the effect of scaling factor (depth/width/resolution) on attribute recognition performance by examine EfficientNets family for this task. Then, we
proposed a framework for jointly human semantic parsing and pedestrian attribute recognition with EfficientNets. By this framework, we can explore both semantic and spatial information with eliminating of background, also we proposed AR-BiFPN to utilize multi-scale features. Experimental results on PETA, RAP and PA-100K datasets suggest that the proposed framework can significantly surpass many existing methods. In the future works, we will discover more efficient solution for using the human semantic parsing knowledge to further improve pedestrian attribute recognition.

REFERENCES

1. Layne, R., et al.: Person re-identification by attributes. British Mach. Vis. Conf. 2(3), 8 (2012)
2. Wang, J., et al.: Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-identification. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 2275–2284 (2018)
3. Feris, R., et al.: Attribute-Based People Search: Lessons Learnt from a Practical Surveillance System. Proc. Of Int. Conf. on Multimedia Retrieval (ICMR), pp. 153–160 (2014)
4. Siddique, B., Feris, R.S., Davis, L.S.: Image Ranking and Retrieval Based on Multi-Attribute Queries. Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Colorado, CO, USA, pp. 801–808 (2011)
5. Kumar, N., et al.: Attribute and Simile Classifiers for Face Verification. IEEE 12th Int. Conf. on Computer Vision, Kyoto, Japan, pp. 365–372 (2009)
6. Deng, Y., et al.: Pedestrian Attribute Recognition at Far Distance. Proc. of the 22nd ACM Int. Conf. on Multimedia, New York, USA, pp. 789–792 (2014)
7. Li, D., Chen, X., Huang, K.: Multi-Attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios. IAPR Asian Conf. on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, pp. 111–115 (2015)
8. Sudoke, P., Spitzer, H., Leibe, B.: Person Attribute Recognition with a Jointly-Trained Holistic CNN Model. 2015 IEEE Int. Conf. on Computer Vision Workshop (ICCVW), Santiago, Chile, pp. 87–95 (2015)
9. Li, D., et al.: Pose Guided Deep Model for Pedestrian Attribute Recognition in Surveillance Scenarios. IEEE Int. Conf. on Multimedia and Expo (ICME), San Diego, CA, USA, pp. 1–6 (2018)
10. Liu, P., et al.: Localization Guided Learning for Pedestrian Attribute Recognition. British Mach. Vis. Conf., (BMVC) (2018)
11. Liu, X., et al.: Hydraplus-Net: Attentive Deep Features for Pedestrian Analysis. IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, pp. 350–359 (2017)
12. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep Unbalanced Attribute Classification Using Visual Attention Aggregation. European Conf. on Computer Vision, Munich, Germany, pp. 708–725 (2018)
13. Sarfraz, M.S., et al.: Deep View Sensitive Pedestrian Attribute Inference in an End-to-End Model. arXiv preprint arXiv:1707.06089 (2017)
14. Tang, C., et al.: Improving Pedestrian Attribute Recognition with Weakly-Supervised Multi-Scale Attribute-Specific Localization. IEEE/CVF Int. Conf. on Computer Vision (ICCV), Seoul, Korea (South), pp. 4996–5005 (2019)
15. Li, Q., et al.: Pedestrian Attribute Recognition by Joint Visual-Semantic Reasoning and Knowledge Distillation. Proc. of the Twenty-Eighth Int. Joint Conf. on Artificial Intelligence, pp. 833–839 (2019)
16. Tan, M., Le, Q.V.: EfficientDet: Scalable and Efficient Object Detection. Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 10781–10790 (2020)
17. Zhu, J., et al.: Pedestrian Attribute Classification in Surveillance: Database and Evaluation. IEEE Int. Conf. on Computer Vision Workshops, Sydney, NSW, Australia, pp. 331–338 (2013)
18. Liu, X., et al.: Pedestrian Attribute Recognition in Surveillance Scenarios. IEEE Trans. on Pattern Analysis and Mach. Intell. (2020)
19. Li, D., Chen, X., Huang, K.: Multi-Attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios. 2015 3rd IAPR Asian Conf. on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, pp. 111–115 (2015)
20. Zhou, Y., et al.: Weakly Supervised Learning of Mid-Level Features for Pedestrian Attribute Recognition and Localization. Proc. of the British Mach. Vis. Conf. (BMVC) (2017)
21. Abdulnabi, A.H., et al.: Multi-Task CNN Model for Attribute Prediction. IEEE Trans. on Multimedia, 17(11), 1949–1959 (2015)
22. Zhu, J., et al.: Multi-Label CNN Based Pedestrian Attribute Learning for Soft Biometrics. 2015 Int. Conf. on Biometrics (ICB), Phuket, Thailand, pp. 535–540 (2015)
23. Diba, A., et al.: DeepCAM: Deep Convolutional Action & Attribute Mid-Level Patterns. IEEE Conf. On Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 3557–3565 (2016)
24. Zitni, C.L., Dollár, P.: Edge Boxes: Locating Object Proposals from Edges. Thirteenth European Conf. on Computer Vision (ECCV), Zurich, pp. 391–405 (2014)
25. Li, D., Chen, X., Huang, K.: Multi-Attribute Learning for Pedestrian Attribute Recognition and Localization. Proc. of the IEEE Trans. on Pattern Recognition and Mach. Intell. 41(4), 871–885 (2018)
26. Xia, F., et al.: Joint Multi-Person Pose Estimation and Semantic Part Segmentation. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, pp. 6769–6778 (2017)
27. Ruan, T., et al.: Devil in the Details: Towards Accurate Single and Multiple Human Parsing. AAAI Conf. on Artificial Intelligence 33(1), 4814–4821 (2019)
28. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-Correction for Human Parsing. IEEE Trans. on Pattern Anal. and Mach. Intell. (2020)
29. Liu, S., et al.: Path Aggregation Network for Instance Segmentation. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 8759–8768 (2018)
30. Zhao, Q., et al.: M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network. The Thirty-Third AAAI Conf. on Artificial Intelligence (AAAI-19), Hilton Hawaiian Village, Honolulu, Hawaii, USA, 33(1), pp. 9259–9266 (2019)
31. Ghiasi, G., et al.: NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 7036–7045 (2019)
32. Sandler, M., et al.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 4510–4520 (2018)
33. Chen, L.C., et al.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. on Pattern Recognition and Mach. Intell. 40(4), 834–848 (2017)
34. Wang, Y., et al.: Resource Aware Person Re-Identification across Multiple Resolutions. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 8042–8051 (2018)
35. Li, D., et al.: A Richly Annotated Dataset for Pedestrian Attribute Recognition. arXiv preprint arXiv:1603.07054 (2016)
36. Zhao, X., et al.: Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning. Proc. Twenty-Seventh International Joint Conf. on Artificial Intelligence, Stockholm, Sweden, pp. 3177–3183 (2018)

How to cite this article: Moghaddam M, Charmi M, Hassanpoor H. Jointly human semantic parsing and attribute recognition with feature pyramid structure in EfficientNets. IET Image Process. 2021;15:2281–2291. https://doi.org/10.1049/ipt2.12195