Let’s Observe Them Over Time: An Improved Pedestrian Attribute Recognition Approach

Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim

1Indian Institute of Technology, Bhubaneswar, Odisha, 752050, India
2Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea
3Yonsei-KIST Convergence Research Institute, Yonsei University, Seoul 03722, Republic of Korea

Abstract

Despite poor image quality, occlusions, and small training datasets, recent pedestrian attribute recognition (PAR) methods have achieved considerable performance. However, leveraging only spatial information of different attributes limits their reliability and generalizability. This paper introduces a multi-perspective approach to reduce over-dependence on spatial clues of a single perspective and exploits other aspects available in multiple perspectives. In order to tackle image quality and occlusions, we exploit different spatial clues present across images and handpick the best attribute-specific features to classify. Precisely, we extract the class-activation energy of each attribute and correlate it with the corresponding energy present across other images using the proposed Self-Attentive Cross Relation Module. In the next stage, we fuse this correlation information with similar clues accumulated from the other images. Lastly, we train a classification neural network using combined correlation information with two different losses. We have validated our method on four widely used PAR datasets, namely Market1501, PETA, PA-100k, and Duke. Our method achieves superior performance over most existing methods, demonstrating the effectiveness of a multi-perspective approach in PAR.

1. Introduction

Pedestrian attribute recognition (PAR) not only facilitates individual-level analysis but also enables broader applications, such as person-re-identification [14, 17, 32], human identification [23], face recognition [11], and person search [1]. The ability to automatically infer attributes such as age, gender, clothing style, and accessories from pedestrian images holds immense potential for safety and surveillance. It can also play a pivotal role in enhancing the capability of security systems deployed for detection of attributes like handbags, cellphones, backpacks etc. This will be useful in lost-and-found situations too. However, despite substantial advancements, pedestrian attribute recognition is still a critical challenge due to the following factors: (i) Pedestrians exhibit a wide range of appearances due to camera parameter variations. (ii) Pedestrian attributes can be highly correlated, making it difficult to distinguish between them solely based on visual cues. (iii) Pedestrians in real-world scenarios are often partially occluded by objects or other individuals. (iv) High-resolution images and larger training datasets are unavailable. Due to these factors, it is challenging to train a robust pedestrian attribute recognition model.

Several notable attempts have been made to recognize pedestrian attributes [3, 5, 10, 14, 15, 19–21, 29, 31, 32, 36]. These attempts can be safely grouped into two broad categories: feature-centric and attribute-correlation approaches. The feature-centric methods are engineered to extract image features that represent specific attributes. Few early at-
tempts mainly leverage hand-crafted features such as texture and color [9]. Recently, deep-learning-based approaches extract features using region proposal [2, 8], attention-mechanism [3, 19, 36], and pose estimation [14]. However, these attempts fail to distil attribute correlations. Correlation-based approaches employ sequential networks such as RNN [36] and ConvLSTM [20] to establish correlation. However, these approaches fall short in the following scenarios: (i) Due to occlusions and poor image quality, important visual cues are obscured or distorted. (ii) Attributes may exhibit ambiguity due to limited viewpoint, e.g. inferring age when the pedestrian’s back region is only visible. (iii) Mining attribute correlation is not always advantageous, especially in smaller, environment-specific datasets. For example, the CUHK subset from PETAPI [6] dataset is exclusively captured within a university campus environment, predominantly featuring students and teachers. Market1501 [37] is dominated by accessories like bags, handbags, hats, backpacks, shorts, etc. The trained models pose the potential challenge of overfitting the specific attributes and characteristics prevalent in the dataset. Moreover, a handful of PAR datasets [6, 17, 37] include multiple instances of the same person. Fig. 2 depicts a few such samples.

Motivated by the above observations, we have introduced a multi-perspective approach that analyzes multiple images of a single pedestrian and predicts the attributes using more robust and reliable cues. Precisely, we extract the class-activation map of each attribute using a baseline network. It highlights the discriminative regions within an image. This predicts a specific attribute with higher intensities, suggesting a more pronounced association between the highlighted regions and the attribute of interest. In addition to this, we also obtain the confidence score of each attribute. We extract these cues across multiple images and establish a correlation between activation energy and confidence score contrary to the attribute correlation approach. In the final stage, we fuse these cues and train a neural network using combined correlation information precisely distil from multiple images. This has been illustrated in Fig. 1. For example, if the handbag attribute is not visible from one of the images, our model extracts relevant spatial features of handbag from the remaining images and combines them with a final set of features to predict handbag. This is difficult to predict using existing PAR methods. We have experimentally validated that the attribute-specific class-activation energy plays a vital role along with the spatial information to predict the attribute accurately. To the best of our knowledge, this is the first approach to address the pedestrian attribute recognition problem using a multi-perspective approach. Moreover, this unique approach traces occlusions opportunely and minimizes the prediction error caused by the limited viewpoints. To achieve this, we have made the following technical contributions:

- We introduce a multi-perspective approach to PAR problem that predicts attributes using stronger class-activation energy across images, thus tackling occlusions and limited viewpoints together.
- We propose Self-Attentive Cross Relation Module (S-ACRM) that mines the correlation between an attribute’s activation energy and confidence score.
- The proposed method has been trained on multiple non-consecutive images of the same pedestrian. It has been extensively evaluated on four widely used PAR datasets, namely PA-100k, Market1501, PETA and Duke.

The rest of the paper is organized as follows. In Sec. 3, we formulate the PAR problem and provide a detailed description of the proposed method. In Sec. 4, we have presented experiments and results. Sec. 5 discusses future directions and concludes the work.

2. Related Work

Pedestrian attribute recognition (PAR) has gained significant attention in computer vision. Earlier methods [9, 11] rely on handcrafted features and separate classifiers for each attribute.

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have shown promising results in PAR. Wang et al. [33] have reviewed deep learning-based methods, highlighting their success in capturing attribute co-occurrence dependencies. Some holistic approaches have been proposed, such as DeepCAMP [7].

Figure 2. Multi-perspective and Occluded Samples: Some of the important PAR datasets contain occluded objects and persons in multiple perspectives. The first two rows are from the PETAPI-3DPeS [6] and Market1501 [37] datasets depicting a person captured from multiple perspectives. The last row depicts a few images containing occluded samples.
Figure 3. **Proposed Architecture:** It consists of three stages: initial predictions by the baseline $\phi$, generation of the correlation matrix $R$ using proposed S-ACRM, and attribute classification stage. In stage one, we accumulate three vectors, class-activation energy, confidence score, and attribute prediction. In the second stage, we employ a self-attention mechanism to learn the correlation between activation energy and confidence score. In the third stage, we combine these relationship matrices and train the fully connected layer $F$ using $L_A$ and $L_G$ losses.

which extracts attribute features from image blocks, and Localization Guided Networks [29], which focus on attribute-related local features. Attention-based methods, like the multi-directional attention model proposed by Liu et al. [19], have been introduced to improve recognition performance.

Another line of research explores modelling the relationships between attributes. JRL [31] utilizes Recurrent Neural Networks (RNNs) to capture the dependencies between attribute labels. Graph-based methods such as VC-GCN [15] and A-AOG [22] incorporate conditional random fields and graphical models to represent attribute correlations. Li et al. [10] have proposed a graph reasoning network to model spatial and semantic relationships between regions and attributes jointly. JLAC [27] employs attribute-relationship and contextual relationship modules to discover and capture attribute and contextual relationships. In addition to this, a multi-view approach is particularly advantageous as it captures diverse perspectives, ensuring robust recognition even in occluded scenarios. Chen et al. [4] have proposed a method utilizing view information and attention mechanisms to accurately localize attributes. QuadNet [35] addresses the challenges of baggage re-identification (ReID) by leveraging multi-view sampling and view-aware attentional features.

It is worth noting that the aforementioned methods often require large-scale labelled training data with high-quality images as inputs, which may not apply to real-world surveillance scenarios with small training data and poor image quality. Therefore, recent research has focused on addressing these challenges. For instance, Li et al. [16] have incorporated human pose guidance to extract local region features. At the same time, the proposed JRL model [31] introduces multiple time steps attention mechanisms to model relations between images and attributes.

In summary, pedestrian attribute recognition has witnessed significant progress with the advent of deep learning techniques, particularly CNNs. Various approaches have been explored including holistic designing, attention-based, graph-based, contextual information, and multi-view cues. However, they fail to perform satisfactorily in the presence of occlusion, poor image quality, and smaller training datasets.

### 3. Proposed Method

We start by formulating the Pedestrian Attribute Recognition (PAR) problem. Next, we present the proposed method that exploits the spatial relationship between non-consecutive pedestrian images to detect attributes. Lastly, we describe the incorporated loss functions, model training, and inference.

#### 3.1. Problem Definition

Following the work presented in [5], we define the PAR problem as a multi-label classification problem, where given
a pedestrian image, our goal is to learn an attribute recognition model that predicts the attributes.

Assume the set of attributes of a pedestrian image is denoted by \( \Pi = \{ \pi_1, \pi_2, \ldots, \pi_M \} \), where \( M \) is the number of attributes available in the dataset. Let \( \{(I_1, Y_1), (I_2, Y_2), \ldots, (I_N, Y_N)\} \) be \( N \) samples in the training set, where \( I_i \) is the \( i \)-th pedestrian image sample and \( Y_i \in \Pi \). Specifically, \( Y \) is a human-annotated binary vector with 0 and 1 specifying the absence and presence of an attribute in image \( I \). Given this setup, our goal is to design a PAR model \( H(.) \) that generates the probability \( p_i \) for each attribute \( \pi_i \) in \( \Pi \), i.e. \( H(I, \Pi) = [p_1, p_2, \ldots, p_M] \).

The probability \( p \) has been used to calculate the loss during training and generate prediction results during inference.

### 3.2. Overall Architecture

An overview of the proposed pipeline is depicted in Fig. 3. It consists of three steps, namely i) the generation of initial attribute prediction by backbone network \((\phi)\), ii) the generation of cross-relation matrix using the proposed Self-Attentive Cross Relation Module (S-ACRM), and iii) the attribute classification stage. Our motivation stems from the understanding that selecting the most informative pedestrian evidence among multiple images enhances attribute detection. Hence, the proposed pipeline generates initial predictions and then attempts to establish a relationship between confidence and activation energy of each attribute to find substantial evidence. In the first step, we obtain three prediction vectors using \( \phi \), namely \( V_c, V_e, \) and \( V_p \). \( V_c, V_e \in \mathbb{R}^{M \times 1} \), where \( V_c \) is the confidence score generated by the FC layer of network \( \phi \), \( V_e \) be the activation energy vector obtained through target layer of \( \phi \), and \( M \) is the number of attributes present in the dataset. \( V_p \) is the binary prediction vector for \( M \) attributes. We have used these vectors to find robust evidence across \( K \) input images and generate the final predictions. In the subsequent sections, we provide detailed descriptions of these stages.

### 3.3. Initial Predictions Using Baselines

Given multiple images of the same pedestrian, we have employed the baseline \( \phi \) to generate predictions of \( K \) images such that \( V_p^i = \phi_s(I^i, W) \) and \( V_e^i = \phi_t(I^i, W) \), where \( V_p^i, V_e^i \) are prediction \((0,1)\) and class-activation energy vectors of \( i \)-th image, \( W \) is pre-trained weights of the baseline, \( \phi_t \) and \( \phi_s \) are target and last sigmoid layer, respectively.

The pipeline aims to predict attributes using stronger evidence across \( K \) images. Hence, we exploit attribute-specific class-activation information to determine the quality of the evidence. Following this [30], we estimate how much class-activation energy falls into the targeted bounding box. To obtain this energy of an attribute, we first obtain the class-activation feature map using the intermediate target layer \( \phi_t \) and calculate the proportion suggested by [30]. To accomplish this, we have divided the input image into two halves (upper and lower body) and assumed these halves as bounding boxes for the attributes. For example, upper body is the bounding box for the hat attribute. Next, the amount of class-activation energy within the box is divided by the total energy of the saliency map to generate a proportion. This energy-based estimation is inspired from the following observation: pedestrian attributes tend to appear in specific regions of the image, e.g. head being predominantly located in the upper part or shoes are in the lower part. Fig. 4 depicts two examples of class-activation energy. For every \( i \)-th attribute \( \pi_i \), we obtain scalar value \( V_p^i \), where high values represent stronger evidence. In addition to this, we also collect confidence score \( V_c^i \) and prediction vector \( V_p^i \) for every \( i \)-th attribute.

### 3.4. Self-Attentive Cross Relation Module

Higher class-activation energy represents stronger support for the presence or absence of the attribute. Similarly, the confidence score indicates the level of certainty regarding the predicted attribute. To model the relationship between the pair of energy and the confidence score of an attribute, we propose an Attentive Cross Relation Module (S-ACRM). It leverages self-attention to capture and model the interdependencies between energy and confidence, ultimately producing a cross-relation matrix that encapsulates associations. We input \( V_c \) and \( V_p \) to S-ACRM and obtain correlation matrix \( R \) with size \( M \times M \) (\( M \) is the number of attributes).
Specifically, \( V_s \) and \( V_c \) correspond to query and key vectors. This module aims to derive attention weights that emphasize the relevance of the energy vector when considering the confidence score for each attribute. This weighting scheme ensures that the attributes with stronger evidence in the energy vector receive heightened attention during the confidence score computation. The resulting attention weights are utilized to construct a cross-relation matrix \( R \) that encapsulates the learned associations between \( V_s \), \( V_c \).

\[
R_i = s(V_i^e \times V_c^T) \\
\in \mathbb{R}^{M \times M}, V_s, V_c \in \mathbb{R}^{M \times 1}
\]

(1)

We first transpose \( V_c \) and multiply it with \( V_e \), and the relation matrix \( R \) is obtained after the softmax operation. This formulation is for every i-th image depicted using Eq. (1), where \( s \) is the softmax operation.

The single row from the matrix \( R \) represents the correlation between the attribute’s energy score and the other attribute’s confidence score. It means, more evident attributes receive increased attention during the prediction. Hence, in S-ACRM, \( V_s \) and \( V_c \) correspond to key and query vectors, respectively.

3.5. Training using Correlation Matrix

The output of the proposed S-ACRM is a correlation matrix \( R \) for each input image. In this stage, we fuse them to generate a combined matrix \( R_c \) using attribute-wise max operation, formulated in Eq. (2).

\[
R_c = \max(R_1, R_2, \ldots, R_K), \quad K \in \{3, 4, \ldots\}
\]

(2)

The i-th row of the matrix \( R_c \) presents a stronger class-activation energy across \( K \) images for the i-th attribute. For example, if the boots of the pedestrian are well-predicted (intense energy) in any input image, then the row belonging to this attribute has higher correlation values even though the model fails to detect it more confidently across the remaining images. In addition to this, we also combine binary prediction vector \( V_p \) as depicted in Eq. (3) using attribute-wise max, where \( K \in \{3, 4, \ldots\} \).

\[
\hat{Y} = \max(V_p^1, V_p^2, \ldots V_p^K)
\]

(3)

Each element in \( \hat{Y} \) represents an attribute’s presence or absence predicted by the baseline model across \( K \) images. In addition to this, we also combine the human-annotated ground truth \( Y \) using a similar attribute-wise max operation as shown in Eq. (4).

\[
Y_G = \max(Y_1, Y_2, \ldots Y_K)
\]

(4)

where \( Y_i \) is the annotated attribute set of the i-th image. Finally, we feed combined correlation-matrix \( R_c \) to the Fully-connected layer \( F \) and train using \( \hat{Y} \) and \( Y_G \). This is depicted in Eq. (5). Specifically, we employ two losses \( L_A \) and \( L_G \) to train the \( F \).

\[
L_A = BCE(F(R_C), \hat{Y}) \quad L_G = BCE(F(R_C), Y_G)
\]

(5)

It has been observed that there is a significant imbalance between pedestrian attributes [5, 12]. To alleviate this imbalance distribution, we have employed a weighted binary cross-entropy (BCE) loss function as given in Eq. (6).

\[
L_A = \sum_{i=1}^{M} \omega_i (\hat{y}_i \log (p_i) + (1 - \hat{y}_i) \log (1 - p_i))
\]

(6)

\[
L_G = \sum_{i=1}^{M} \omega_i (y_i \log (p_i) + (1 - y_i) \log (1 - p_i))
\]

(7)

where \( \hat{y}_i \in \hat{Y}, y_i \in Y_G, p_i \) probability generated by the model and \( \omega_i \) is the imbalance weight of the attribute \( \pi_i \) and calculate using Eq. (8).

\[
\omega_j = \begin{cases} 
1 - r_i, & y_i = 1 \\
e^{-r_i}, & y_i = 0 
\end{cases}
\]

(8)

where \( r_i \) is the positive sample ratio of attribute \( \pi_i \) in the training set. The final loss function consists of the sum of the loss functions depicted in Eq. (6) and Eq. (7), that is, 

\[ L = L_A + L_G \]

3.6. Inference

We train the proposed model using multiple images of the same person to emphasize more on attribute-specific evidence. The trained model can easily predict a person’s attributes using a single image during the inference stage. Firstly, \( V_s \) and \( V_c \) are generated by the baseline \( \phi \). Secondly, these vectors are utilized to obtain the correlation matrix \( R \). Lastly, they are fed to the trained fully connected model \( F \) to generate the final attribute predictions.

4. Experiments

In this section, we present implementation details, datasets, evaluation metrics, comparisons of the proposed method with recent state-of-the-art PAR methods, qualitative results, and ablation experiments.

4.1. Implementation Details

We employ DeepMAR [12] architecture as a baseline network \( \phi \) to acquire initial prediction vectors. The authors
of DeepMAR have constructed a 5-layer CNN. However, we have utilized ResNet50 pre-trained on ImageNet as a feature extractor. The whole network has been trained using sigmoid cross entropy loss as suggested by the authors in [12, 29]. We have also carried out ablation experiments related to this backbone. The fully-connected layer $\phi_F$ consists of 4096, 512, and $M$ output neurons, where $M$ is the number of attributes in the training set. We have extracted the class-activation map using Grad-CAM [38] and calculated the energy of the map as suggested in Score-CAM [30]. $M = 27$, 35, and 26 for Market1501 [37], PETA [6], and PA-100k [19], respectively. The final fully-connected network $F$ has been trained with Adam optimizer with the initial learning rate set to $1 \times 10^{-3}$ for the first 20 epochs, and then a linear decrease by a factor of 0.1 as the epochs increase. The batch size has been set to 64 and 128.

### 4.2. Datasets and Evaluation Metrics

The PA-100k [19] dataset comprises with 100,000 pedestrian images captured across 598 outdoor scenes. Each image was annotated with 26 widely utilized attributes. The dataset has been partitioned into training, validation, and test sets to facilitate research, maintaining an 8:1:1 ratio.

The PETA dataset [6] encompasses 8,705 pedestrians captured in 19,000 images with varying resolutions ranging from $17 \times 39$ to $169 \times 365$. Each pedestrian was annotated with 61 binary attributes and four multi-class attributes. However, as per the established protocol, only 35 attributes are instance-based, namely **accuracy, precision, recall, and F1 score**.

### 4.3. Comparisons with State-of-the-art

We have compared our method with recent PAR methods on four aforementioned datasets. Tab. 1 presents the performance of recent notable methods [12, 13, 15, 18, 19, 29] on PA-100k dataset.

#### Table 1. Performance comparisons (in %) of the state-of-the-art PAR methods over the PA-100k [19] dataset. Since the PA-100k dataset does not contain multiple appearances of the person, the value of $K$ is set to 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Visual Encoder</th>
<th>mA</th>
<th>Accu</th>
<th>Prec</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepMAR [12]</td>
<td>CaffeNet</td>
<td>84.61</td>
<td>80.74</td>
<td>88.38</td>
<td>87.92</td>
<td>88.70</td>
</tr>
<tr>
<td>HP-Net [19]</td>
<td>CaffeNet</td>
<td>81.77</td>
<td>75.04</td>
<td>83.88</td>
<td>82.46</td>
<td>83.14</td>
</tr>
<tr>
<td>VS-GCN [19]</td>
<td>CaffeNet</td>
<td>81.15</td>
<td>73.66</td>
<td>84.06</td>
<td>81.26</td>
<td>82.64</td>
</tr>
<tr>
<td>PGDM [13]</td>
<td>CaffeNet</td>
<td>81.15</td>
<td>73.66</td>
<td>84.06</td>
<td>81.26</td>
<td>82.64</td>
</tr>
<tr>
<td>LG-Net [18]</td>
<td>CaffeNet</td>
<td>81.15</td>
<td>73.66</td>
<td>84.06</td>
<td>81.26</td>
<td>82.64</td>
</tr>
<tr>
<td>ALM [29]</td>
<td>BN-Inception</td>
<td>80.58</td>
<td>77.43</td>
<td>84.21</td>
<td>82.41</td>
<td>83.84</td>
</tr>
<tr>
<td>Ours</td>
<td>ResNet50</td>
<td>82.26</td>
<td>77.14</td>
<td>84.36</td>
<td>82.73</td>
<td>83.37</td>
</tr>
<tr>
<td>Ours</td>
<td>VGG16</td>
<td>81.45</td>
<td>75.36</td>
<td>84.10</td>
<td>85.34</td>
<td>86.55</td>
</tr>
</tbody>
</table>

Among the compared methods, DeepMAR [12] with CaffeNet as the visual encoder achieves mean accuracy of 72.70%, demonstrating considerable performance. HP-Net [19] with Inception as the visual encoder achieves a slightly higher value (74.21%), indicating improved accuracy across multiple attributes. Without explicitly relying on the visual encoder, the VS-GCN [15] method shows remarkable results with an mA of 79.52%. This method achieves higher precision (89.40%) and recall (87.15%), indicating its effectiveness using graph networks. ALM [29] with BN-Inception as the visual encoder achieves an impressive mA of 80.68%. This method excels in recall (88.44%), indicating its ability to capture attributes accurately. Significant precision and recall have been observed for all methods due to the availability of a sufficient number of training samples in the PA-100K dataset. However, a few of them still under-perform due to large occlusions. In contrast, our proposed methods using ResNet50 and VGG16 as visual encoders achieve competitive performance with mA values of 82.26% and 81.45%, respectively, indicating its robustness against the high rate of occlusions.

#### Table 2. Performance comparisons (in %) of the state-of-the-art PAR methods over the PETA [6] dataset. It contains at least two non-consecutive appearances for each pedestrian, hence $K = 2$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Visual Encoder</th>
<th>mA</th>
<th>Accu</th>
<th>Prec</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSAR [12]</td>
<td>VGG16</td>
<td>81.30</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DeepMAR [12]</td>
<td>CaffeNet</td>
<td>82.89</td>
<td>75.07</td>
<td>83.68</td>
<td>83.14</td>
<td>83.41</td>
</tr>
<tr>
<td>ACN [28]</td>
<td>Inception</td>
<td>81.15</td>
<td>73.66</td>
<td>84.06</td>
<td>81.26</td>
<td>82.64</td>
</tr>
<tr>
<td>JRL [31]</td>
<td>-</td>
<td>85.67</td>
<td>86.03</td>
<td>85.34</td>
<td>85.42</td>
<td>85.42</td>
</tr>
<tr>
<td>HP-Net [19]</td>
<td>Inception</td>
<td>81.77</td>
<td>76.13</td>
<td>84.92</td>
<td>83.24</td>
<td>84.07</td>
</tr>
<tr>
<td>VeSPA [25]</td>
<td>-</td>
<td>83.45</td>
<td>77.73</td>
<td>86.18</td>
<td>84.81</td>
<td>85.49</td>
</tr>
<tr>
<td>MsVNA [23]</td>
<td>BN-Inception</td>
<td>84.59</td>
<td>78.56</td>
<td>86.79</td>
<td>86.12</td>
<td>86.46</td>
</tr>
<tr>
<td>VC-GCN [35]</td>
<td>-</td>
<td>85.21</td>
<td>81.82</td>
<td>89.43</td>
<td>88.42</td>
<td>88.42</td>
</tr>
<tr>
<td>Ours</td>
<td>ResNet50</td>
<td>88.45</td>
<td>82.21</td>
<td>90.12</td>
<td>87.46</td>
<td>88.70</td>
</tr>
<tr>
<td>Ours</td>
<td>VGG16</td>
<td>86.18</td>
<td>80.60</td>
<td>87.56</td>
<td>86.23</td>
<td>86.55</td>
</tr>
</tbody>
</table>

Tab. 2 represents the performance of recent notable methods [12, 15, 19, 24–26] on the PETA dataset. Among the compared methods, our approach with ResNet50 as the visual encoder achieves the highest mA of 88.45%, indicating its superior performance in accurately recognizing pedestrian attributes. Our method demonstrates high precision (90.12%).
and a balanced recall (87.46%), showcasing its effectiveness in capturing attributes using multiple cues during the training. In contrast, other methods such as DeepSAR [12] and JRL [31] do not provide specific values for certain metrics, indicating limited performance analysis. ACN [26] and HPNet [19] achieve moderate mA values: 81.15% and 81.77%, respectively, with relatively lower precision and recall values. VeSPA and MsVV A show improved performance with higher mA values: 83.45% and 84.59%, respectively. However, they still exhibit lower precision and recall than the proposed method. VC-GCN [15] achieves a high mA of 85.21%, but its precision, recall, and F1 score are identical, suggesting a bias in its attribute predictions.

The number of samples in Market1501 [37] and Duke-Attribute [17] is less. Hence, a couple of works [17, 34] have reported new criteria mFive, which is the mean accuracy over the five criteria (mA, Accu, Prec, Recall, and F1). Tab. 4 compares the performance of a few PAR methods using this criterion on two datasets: Market-1501 [37] and Duke-Attribute [17]. It can be observed that, in terms of accuracy, our method achieves the best performance with 91.43% accuracy on the Market-1501 dataset and 89.16% accuracy on the Duke-Attribute dataset. This indicates the significance of the activation energy extracted from multiple images. Another reason is, Market1501 and Duke datasets have been captured using multiple cameras, resulting in more complex pedestrian appearances.

4.4. Qualitative Analysis

Fig. 6 depicts attribute-specific localization results obtained by the proposed method on three PAR datasets, namely Market-1501 [37], PETA [6], and Duke [17]. It can be observed that the model correctly predicts attributes based on their appearances in the image. Moreover, we have also compared the attribute predictions with DeepMAR as the baseline. It is shown in Fig. 5. It can be seen from the first example that DeepMAR and advanced DeepMAR cannot recognise the handbag’s presence due to partial occlusion. However, our method predicts it correctly. On the other hand, for the second example, due to heavy occlusion and similarity between the colour of the jacket and bag, no method (including ours) can recognize the bag. However, our method mistakenly realizes it as a backpack, which is a reasonable prediction. Moreover, due to colour similarity, both DeepMAR and advanced DeepMAR fail to distinguish the jacket to which our method predicts it correctly. Both Figs. 5 and 6 witness the advantage of leveraging class-activation clues to predict attributes even in partial occlusion accurately. More such prediction examples have been added to the supplementary document.

4.5. Effect of Number of Input Image

A handful of PAR datasets [6, 17, 37] include multiple instances of the same person. Also, the number of instances per pedestrian varies throughout the datasets. Moreover, their pose, lighting condition, and back-front appearances vary according to the scenarios. Since $K$ number of images of the same pedestrian have been used during the training, the following question arises: What is the minimum number of non-consecutive images of the same pedestrian required to learn all attributes? To find an answer to this question, we have trained the proposed model with a fixed number of non-consecutive images of the same pedestrian and reported the mA values on three chosen datasets, namely PETA [6], Market-1501 [37], and Duke-attribute [17]. Note that these are custom experiments. We have used only selected datasets with pedestrian images with at least 2 different appearances of the same person. Moreover, during the specific $K$ value experiment, we selected all training samples with at least $K$ appearances and discarded the remaining samples. Tab. 3 shows how mean accuracy varies with different $K$ values. From the results, we safely conclude that the number of appearances provided during the training plays a critical
role in learning a diverse range of attributes. For example, two appearances are not provided with a significant gain in accuracy due to low dissimilarity (same pose or lighting conditions). However, when the number of different appearances increases, the model learns more robust features and can avoid occlusions or other limiting parameters.

Table 3. Effect of $K$ images of the same pedestrian used during the training on three datasets. All values are in %.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>86.46</td>
<td>82.46</td>
<td>87.33</td>
</tr>
<tr>
<td>3</td>
<td>91.67</td>
<td>82.93</td>
<td>89.10</td>
</tr>
<tr>
<td>4</td>
<td>92.71</td>
<td>86.57</td>
<td>92.79</td>
</tr>
<tr>
<td>5</td>
<td>95.55</td>
<td>89.41</td>
<td>95.22</td>
</tr>
</tbody>
</table>

Table 4. Performance comparisons (in %) using the PAR methods over the Market1501 [37] and Duke [17] Attribute dataset. The accuracy reported in this table is derived using mFive metric.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PedAttriNet [17]</td>
<td>84.64</td>
<td>80.07</td>
</tr>
<tr>
<td>APR [17]</td>
<td>85.33</td>
<td>80.12</td>
</tr>
<tr>
<td>JLPLS-PAA [28]</td>
<td>87.88</td>
<td>85.24</td>
</tr>
<tr>
<td>EALC [34]</td>
<td>88.41</td>
<td>85.76</td>
</tr>
<tr>
<td>Ours</td>
<td>91.43</td>
<td>89.16</td>
</tr>
</tbody>
</table>

4.6. Performance Comparisons on Occluded Samples

Occlusions can lead to misinterpretations, reducing the reliability and performance of recognition models as they obscure critical features and details. In order to understand its effectiveness, we have tested the proposed method, recent transformer-based method [5], and DeepMAR [12] on 500 occluded images from PETA [6] datasets. We have observed 92.14%, 87.57% and 78.43% mA values, respectively. Our method demonstrates superior performance demonstrating its robustness and adaptability in handling occlusions. Fig. 5 depicts such an occluded example shown on its right side.

4.7. Ablation Studies

Table 5 presents the results of the ablation study conducted on the proposed model using the PETA dataset. The model proposed in DeepMAR [12] inspires the baseline model. Here, a 5-layer CNN has been trained with the FCN layer. It achieves an mA score of 82.89% and an F1 score of 83.41%. Adding the proposed S-ACRM block improves mA score to 85.36% and an F1 score to 84.12%, indicating the importance of class-activation cues. Further enhancements are observed when incorporating additional components. Combining S-ACRM with different training losses, $L_G$ and $L_A$, yields mA scores of 85.43% and 84.98%, respectively. Finally, incorporating two losses during training has achieved improved results with mA score of 86.45% and an F1 score of 88.70%. These results demonstrate the effectiveness of the S-ACRM module and combined losses in improving the method’s performance.

Table 5. Ablation study of the proposed method on the PETA [6] dataset.

<table>
<thead>
<tr>
<th>Component</th>
<th>mA (%)</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepMAR (CNN) [15]</td>
<td>82.89</td>
<td>83.41</td>
</tr>
<tr>
<td>DeepMAR (ResNet50)</td>
<td>83.56</td>
<td>84.12</td>
</tr>
<tr>
<td>DeepMAR + S-ACRM</td>
<td>84.67</td>
<td>84.81</td>
</tr>
<tr>
<td>DeepMAR + S-ACRM + $L_A$</td>
<td>84.98</td>
<td>85.09</td>
</tr>
<tr>
<td>DeepMAR + S-ACRM + $L_G$</td>
<td>85.43</td>
<td>87.65</td>
</tr>
<tr>
<td>DeepMAR + S-ACRM + ($L_A + L_G$)</td>
<td>86.45</td>
<td>88.70</td>
</tr>
</tbody>
</table>

5. Conclusion and Future Work

In conclusion, this research introduces a novel multi-perspective approach for pedestrian attribute recognition (PAR) to overcome the challenges often posed by the poor image quality, occlusions, and limited training datasets. While previous PAR methods have shown significant progress, they often rely solely on spatial information from a single perspective, limiting their reliability and generalizability. The proposed approach addresses this limitation by leveraging multiple perspectives and exploiting various aspects available in the images. It employs a Self-Attentive Cross Relation Module (S-ACRM) to correlate the class-activation energy of each attribute across different images, effectively capturing attribute-specific features. Additionally, the method incorporates a fusion of correlation information and utilizes two different losses for training a classification neural network. Experimental evaluations on widely used PAR datasets, including Market1501, PETA, PA-100k, and Duke demonstrate the superior performance of the multi-perspective approach as compared to existing methods. The results highlight the effectiveness of considering multiple perspectives in PAR, reducing overdependence on spatial clues from a single viewpoint.

Furthermore, this research opens up a few interesting avenues for future work. Firstly, exploring more advanced fusion techniques for incorporating correlation information from multiple perspectives can potentially enhance the overall performance of the PAR system. Moreover, investigating methods to leverage additional contextual information, such as scene context or temporal cues, may improve attribute recognition accuracy.

Acknowledgement

This work was supported in part by the Korea Institute of Science and Technology (KIST) Institutional Program under Project 2E32301 and in part by the National Research Foundation (NRF) Project (Grant No. 2018M3E3A1057288) executed at IIT Bhubaneswar with project code CP438.
References


[27] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z. Li. Relation-aware pedestrian attribute recognition with graph...


