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

Kamalakar Vijay Thakare[1], Debi Prosad Dogra[1], Heeseung Choi[2,3], Haksub Kim[2] and Ig-Jae Kim[2,3]

[1]Indian Institute of Technology, Bhubaneswar, Odisha, 752050, India

[2]Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea

[3]Yonsei-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.

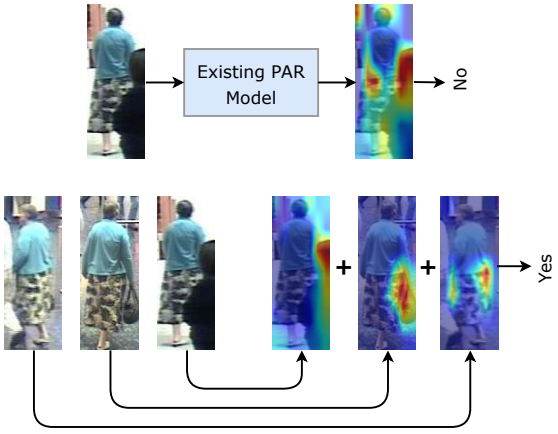

Figure 1. **"Is this person holding or carrying a bag?"** The existing PAR models process a single image and predict solely based on spatial cues. The occlusion and limited viewpoint often hinder prediction. Our proposed multi-perspective approach exploits class-activation information across images to accurately predict the attribute.

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-

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 PETA-3DPeS [6] and Market1501 [37] datasets depicting a person captured from multiple perspectives. The last row depicts a few images containing occluded samples.

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 PETA [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 *hand bag* attribute is not visible from one of the images, our model extracts relevant spatial features of *hand bag* from the remaining images and combines them with a final set of features to predict *hand bag*. 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],

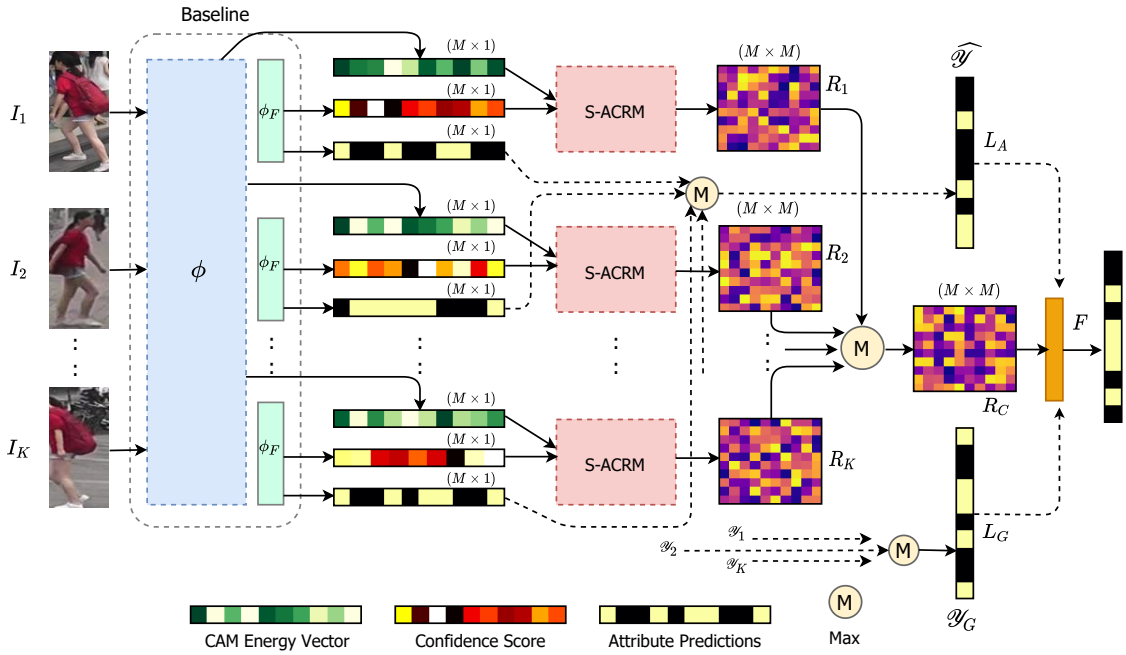

Figure 3. **Proposed Architecture:** It consists of three stages: initial predictions by the baseline ϕ, generation of correlation matrix R using proposed S-ACRM, and attribute classification stage. In stage one, we accumulate three vectors, class-activation energy, confidence score, and attribute prediction. 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 $\{(\mathcal{I}_1, \mathcal{Y}_1), (\mathcal{I}_2, \mathcal{Y}_2), \ldots (\mathcal{I}_N, \mathcal{Y}_N)\}$ be N samples in the training set, where $\mathcal{I}_i$ is the i-th pedestrian image sample and $\mathcal{Y}_i \in \Pi$. Specifically, $\mathcal{Y}$ is a human-annotated binary vector with 0 and 1 specifying the absence and presence of an attribute in image $\mathcal{I}$. Given this setup, our goal is to design a PAR model $\mathcal{H}(.)$ that generates the probability p_i for each attribute π_i in Π, i.e. $\mathcal{H}(\mathcal{I}, \Pi) = [p_1, p_2, \ldots p_M]$. 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 (ϕ), 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 ϕ, namely $\mathcal{V}_e$, $\mathcal{V}_c$, and $\mathcal{V}_p$. $\mathcal{V}_e, \mathcal{V}_c \in \mathbb{R}^{M \times 1}$, where $\mathcal{V}_c$ is the confidence score generated by the FC layer of network ϕ, $\mathcal{V}_e$ be the activation energy vector obtained through target layer of ϕ, and M is the number of attributes present in the dataset. $\mathcal{V}_p$ 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 ϕ to generate predictions of K images such that $\mathcal{V}_p^i = \phi_s(\mathcal{I}^i, \mathbf{W})$ and $\mathcal{V}_e^i = \phi_t(\mathcal{I}^i, \mathbf{W})$, where $\mathcal{V}_p^i$, $\mathcal{V}_e^i$ are prediction (0,1) and class-activation energy vectors of i-th image, $\mathbf{W}$ is pre-trained weights of the baseline, ϕ_t and ϕ_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 ϕ_t and calculate the *proportion* suggested by [30]. To

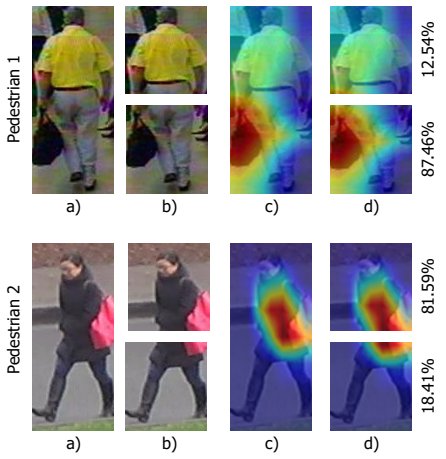

Figure 4. **Class-activation Energy:** Pedestrian 1 is an example from PETA-GRID [6] and pedestrian 2 is a sample from DukeMTMC-reID [17]. Each image sample is represented using four parts: a) input sample, b) two body parts (upper and lower), c) class-activation map for a specific attribute, and d) energy map division according to the body parts. In the Pedestrian 1 sample, for *bag* prediction, the energy of lower body (87.46%) is higher than the upper body (12.54%). In contrast, in Pedestrian 2 for the *upper_cloth_color*, energy of upper body (81.59%) is higher.

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 π_i, we obtain scalar value $\mathcal{V}_e^i$, where high values represent stronger evidence. In addition to this, we also collect confidence score $\mathcal{V}_c^i$ and prediction vector $\mathcal{V}_p^i$ 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 $\mathcal{V}_c$ and $\mathcal{V}_e$ to S-ACRM and obtain correlation matrix R with size $M \times M$ (M is the number of attributes).

Specifically, $\mathcal{V}_c$ and $\mathcal{V}_e$ 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 $\mathcal{V}_c$, $\mathcal{V}_e$.

$$\begin{aligned} R_i &= s(\mathcal{V}_e^i \times \mathcal{V}_c^{iT}) \\ R \in \mathbb{R}^{M\times M}, &\mathcal{V}_c, \mathcal{V}_e \in \mathbb{R}^{M\times 1} \end{aligned} \tag{1}$$

We first transpose $\mathcal{V}_c$ and multiply it with $\mathcal{V}_e$, 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, $\mathcal{V}_e$ and $\mathcal{V}_c$ correspond to key and query vectors, respectively.

3.5. Trainig 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).

$$R_c = max(R_1, R_2, ..., R_K), \quad K \in \{3, 4, ...\} \tag{2}$$

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 $\mathcal{V}_p$, as depicted in Eq. (3) using attribute-wise *max*, where $K \in \{3, 4, \ldots\}$.

$$\hat{\mathcal{Y}} = max(\mathcal{V}_p^1, \mathcal{V}_p^2, \ldots \mathcal{V}_p^K) \tag{3}$$

Each element in $\hat{\mathcal{Y}}$ 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 $\mathcal{Y}$ using a similar attribute-wise *max* operation as shown in Eq. (4),

$$\mathcal{Y}_G = max(\mathcal{Y}_1, \mathcal{Y}_2, \ldots \mathcal{Y}_K) \tag{4}$$

where $\mathcal{Y}_i$ 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 $\hat{\mathcal{Y}}$ and $\mathcal{Y}_G$. This is depicted in Eq. (5). Specifically, we employ two losses $\mathcal{L}_A$ and $\mathcal{L}_G$ to train the F.

$$\mathcal{L}_A = BCE(F(R_C), \hat{\mathcal{Y}}) \quad \mathcal{L}_G = BCE(F(R_C), \mathcal{Y}_G) \tag{5}$$

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),

$$\mathcal{L}_A = \sum_{i=1}^{M} \omega_i \left(\hat{y}_i \log\left(p_i\right) + \left(1 - \hat{y}_i\right) \log\left(1 - p_i\right)\right) \tag{6}$$

$$\mathcal{L}_G = \sum_{i=1}^{M} \omega_i \left(y_i \log\left(p_i\right) + \left(1 - y_i\right) \log\left(1 - p_i\right)\right) \tag{7}$$

where $\hat{y}_i \in \hat{\mathcal{Y}}$, $y_i \in \mathcal{Y}_G$, p_i probability generated by the model and ω_i is the imbalance weight of the attribute π_i and calculate using Eq. (8),

$$\omega_j = \begin{cases} e^{1-r_i}, & y_i = 1 \\ e^{r_i}, & y_i = 0 \end{cases} \tag{8}$$

where r_i is the positive sample ratio of attribute π_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, $\mathcal{L} = \mathcal{L}_G + \mathcal{L}_A$.

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, $\mathcal{V}_e$ and $\mathcal{V}_c$ are generated by the baseline ϕ. 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 ϕ 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 ϕ_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 1×10^{-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 × 39 to 169 × 365. Each pedestrian was annotated with 61 binary attributes and four multi-class attributes. However, as per the established protocol, only 35 attributes with a positive label ratio exceeding 5% have been utilized in the present analysis.

The **Market-1501** [37] dataset was collected before a supermarket at Tsinghua University. The original dataset contains 751 pedestrians for training and 750 for testing. The attributes were annotated at the pedestrian level; thus, the dataset contains 28 x 751 samples for training and 28 x 750 for testing.

The **Duke Attribute** dataset [17] was also labelled at the identity level. It contains 34,183 images from 1812 identities, and each image was annotated with 8 binary attributes and 2 multi-class attributes.

We have used **mean accuracy** (mA) metric to evaluate label-based predictions. It involves calculating the accuracy of each attribute across all samples, regardless of their positive or negative labels. The mA is then obtained by averaging the accuracy values of all attributes. The alternate metrics 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.

The top two results are shown in red and blue.

Method	Visual Encoder	mA	Accu	Prec	Recall	F1
DeepMAR [12]	CaffeNet	72.70	70.39	82.24	80.42	81.32
HP-Net [19]	Inception	74.21	72.19	82.97	82.09	82.53
VS-GCN [19]	-	79.52	**80.58**	**89.40**	87.15	86.26
PGDM [13]	CaffeNet	74.95	73.08	84.36	82.24	83.29
LG-Net [18]	-	76.96	75.55	**86.99**	83.17	85.04
ALM [29]	BN-Inception	80.68	77.08	84.21	**88.84**	86.46
Ours	ResNet50	**82.26**	**77.19**	86.36	**87.92**	**87.32**
Ours	VGG16	**81.45**	75.36	84.10	85.34	**86.55**

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 (88.84%), 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 underperform 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 $K = 2$.

Method	Visual Encoder	mA	Accu	Prec	Recall	F1
DeepSAR [12]	VGG16	81.30	-	-	-	-
DeepMAR [12]	CaffeNet	82.89	75.07	83.68	83.14	83.41
ACN [26]	Inception	81.15	73.66	84.06	81.26	82.64
JRL [31]	-	85.67	–	86.03	85.34	85.42
HP-Net [19]	Inception	81.77	76.13	84.92	83.24	84.07
VeSPA [25]	-	83.45	77.73	86.18	84.81	85.49
MsVVA [24]	BN-Inception	84.59	78.56	86.79	86.12	86.46
VC-GCN [15]	-	85.21	**81.82**	**88.43**	**88.42**	**88.42**
Ours	ResNet50	**88.45**	**82.21**	**90.12**	**87.46**	**88.70**
Ours	VGG16	**86.18**	80.69	87.56	86.23	86.55

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 HP-Net [19] achieve moderate mA values: 81.15% and 81.77%, respectively, with relatively lower precision and recall values. VeSPA and MsVVA 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.

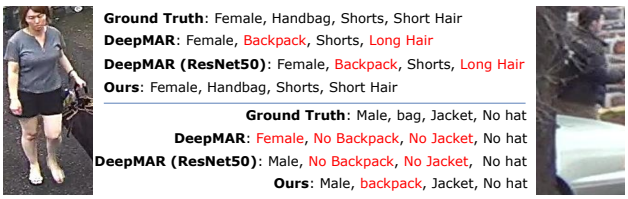

Figure 5. **Prediction Quality:** Qualitative analysis of the proposed method, DeepMAR [12], DeepMAR with ResNet50 as visual encoder. Wrong predictions are in red; right predictions are in black.

Figure 6. **Attribute-specific Localization:** Visualization of the prediction of specific attributes in the image generated by the proposed model. The first-row image and attributes are obtained from Market1501 [37], whereas the second and third rows belong to subset PETA-CUHK [6] and Duke dataset [17], respectively.

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 %.

K	PETA [6]	Market-1501 [37]	Duke-Attribute [17]
2	86.46	82.46	87.33
3	91.67	82.93	89.10
4	92.71	86.57	92.79
5	95.55	89.41	95.22

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.

Method	Market-1501 [37]	Duke-Attribute [17]
PedAttriNet [17]	84.64	80.07
APR [17]	85.33	80.12
JLPLS-PAA [28]	87.88	85.24
EALC [34]	**88.41**	**85.76**
Ours	**91.43**	**89.16**

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 83.56% 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.

Component	mA (%)	F1 (%)
DeepMAR (CNN) [12]	82.89	83.41
DeepMAR(ResNet50)	83.56	84.12
DeepMAR + S-ACRM	84.67	84.81
DeepMAR + S-ACRM + L_A	84.98	85.09
DeepMAR + S-ACRM + L_G	**85.43**	**87.65**
DeepMAR + S-ACRM + $(L_A + L_G)$	**86.45**	**88.70**

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

[1] Surbhi Aggarwal, R. Venkatesh Babu, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2606–2614, 2020. 1

[2] Haoran An, Haonan Fan, Kaiwen Deng, and Hai-Miao Hu. Part-guided network for pedestrian attribute recognition. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019. 2

[3] Haoran An, Hai-Miao Hu, Yuanfang Guo, Qianli Zhou, and Bo Li. Hierarchical reasoning network for pedestrian attribute recognition. *IEEE Transactions on Multimedia*, 23:268–280, 2021. 1, 2

[4] Wei-Chen Chen, Xin-Yi Yu, and Lin-Lin Ou. Pedestrian attribute recognition in video surveillance scenarios based on view-attribute attention localization. *Machine Intelligence Research*, pages 153–168, 2022. 3

[5] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Trans. Circuit Syst. Video Technol.*, 32(10):6994–7004, 2022. 1, 3, 5, 8

[6] Yubin DENG, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM International Conference on Multimedia*, page 789–792, 2014. 2, 4, 6, 7, 8

[7] Ali Diba, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Deepcamp: Deep convolutional action and attribute mid-level patterns. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3565, 2016. 2

[8] Daiheng Gao, Zhenzhi Wu, and Weihao Zhang. Safe-net: Solid and abstract feature extraction network for pedestrian attribute recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1655–1659, 2019. 2

[9] Emad Sami Jaha and Mark S. Nixon. Soft biometrics for subject identification using clothing attributes. In *IEEE International Joint Conference on Biometrics*, pages 1–6, 2014. 2

[10] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 942–951, 2021. 1, 3

[11] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372, 2009. 1, 2

[12] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115, 2015. 5, 6, 7, 8

[13] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018. 6

[14] Huafeng Li, Shuanglin Yan, Zhengtao Yu, and Dapeng Tao. Attribute-identity embedding and self-supervised learning for scalable person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 3472–3485, 2020. 1, 2

[15] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Visual-semantic graph reasoning for pedestrian attribute recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8634–8641, Jul. 2019. 1, 3, 6, 7

[16] Ye Li, Lei Wu, Ziyang Chen, Guangqiang Yin, Xinzhong Wang, and Zhiguo Wang. Identity-assisted network for pedestrian attribute recognition. In *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 1–6, 2022. 3

[17] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 1, 2, 4, 6, 7, 8

[18] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. In *British Machine Vision Conference*, 2018. 6

[19] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, page 1–9, 2017. 1, 2, 3, 6, 7

[20] Yuan Liu, Maoqing Tian, Jun Hou, Shuai Yi, and Zhiping Lin. Pentadent-net: Pedestrian attribute recognition with distance refinement and correlation mining. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2211–2215, 2020. 1, 2

[21] Yan Wang M. Saquib Saquib, Arne Schumann and Rainer Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 134.1–134.13, 2017. 1

[22] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1555–1569, 2018. 3

[23] Daniel A. Reid and Mark S. Nixon. Using comparative human descriptions for soft biometrics. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–6, 2011. 1

[24] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 708–725, 2018. 6

[25] M. Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In *British Machine Vision Conference*, 2017. 6

[26] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 329–337, 2015. 6, 7

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

convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12055–12062, 2020. 3

[28] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z. Li. Attention-based pedestrian attribute analysis. *IEEE Transactions on Image Processing*, pages 6126–6140, 2019. 8

[29] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4997–5006, 2019. 1, 3, 6

[30] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 4, 6

[31] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 531–540, 2017. 1, 3, 6, 7

[32] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018. 1

[33] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, page 108220, 2022. 2

[34] Dunfang Weng, Zichang Tan, Liwei Fang, and Guodong Guo. Exploring attribute localization and correlation for pedestrian attribute recognition. *Neurocomputing*, pages 140–150, 2023. 7, 8

[35] Hao Yang, Xiuxiu Chu, Li Zhang, Yunda Sun, Dong Li, and Stephen J Maybank. Quadnet: Quadruplet loss for multi-view learning in baggage re-identification. *Pattern Recognition*, 126:108546, 2022. 3

[36] Xin Zhao, Liufang Sang, Guiguang Ding, Jungong Han, Na Di, and Chenggang Yan. Recurrent attention model for pedestrian attribute recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9275–9282, 2019. 1, 2

[37] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 2, 6, 7, 8

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2921–2929, 2016. 6