IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 January 2023, accepted 29 January 2023, date of publication 1 February 2023, date of current version 6 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241334

RESEARCH ARTICLE

Efficient Deep Learning Approach to Recognize Person Attributes by Using Hybrid Transformers for Surveillance Scenarios

S. RAGHAVENDRA[1], RAMYASHREE[1], S. K. ABHILASH[2], VENU MADHAV NOOKALA[2], AND S. KALIRAJ[1]

[1]Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India
[2]PathPartner Technology Private Ltd., Bengaluru 560075, India

Corresponding author: S. Kaliraj (kaliraj.s@manipal.edu)

ABSTRACT Numerous deep perception technologies and methods are built on the foundation of pedestrian feature identification. It covers various fields, including autonomous driving, spying, and object tracking. A recent study area is the identification of personality traits that has attracted much interest in video surveillance. Identifying a person's distinct areas is complex and plays an incredibly significant role. This paper presents a current method applied to networks of primary convolutional neurons to locate the area connected to the Person attribute. Using Individual Feature Identification, the features of a person, such as gender, age, fashion sense, and equipment, have received much attention in video surveillance analytics. This Article adopted a Conv-Attentional image transformer that broke down the most discriminating Attribute and region into multiple grades. The feed-forward system and conv-attention are the components of serial blocks, and parallel blocks have two attention-focused tactics: direct cross-layer attention and feature interpolation. It also provides a flexible Attribute Localization Module (ALM) to learn the regional aspects of each Attribute are considered at several levels, and the most discriminating areas are selected adaptively. We draw the conclusion that hybrid transformers outperform pure transformers in this instance. The extensive experimental results indicate that the proposed hybrid technique achieves higher results than the current strategies on four unique private characteristic datasets, i.e., RapV2, RapV1, PETA, and PA100K.

INDEX TERMS Attribute recognition, CNN, deep neural network, image classification, transformers.

I. INTRODUCTION

Analysis of Considering pedestrian characteristics is a growing topic because there is a need for intelligent video surveillance with the aid of deep neuronal networks. Numerous real-world applications, such as autonomous driving, robotic navigation, and video survival, require the features of pedestrian attributes. It is one of the most comprehensive computer vision research projects to date and has found substantial improvement. However, modern-day strategies are not extraordinarily strong, and their performance varies across datasets. Some widely used pedestrian attribute recognition techniques go through everything from over-fitting to supplying datasets, particularly when it comes to autonomous driving.

Consequently, one of the future potentials for person detection and its characteristics is improved individual detection through recognition will lead to improved, refined factors and fine-grained recognition characteristics. PAR (Person Attribute Recognition) ambitions to decide the attributes of the focused character image. Identification of a person's features, such as age, range, gender, dress style, shoes, etc., is done through person attribute recognition. A multidimensional, deep neural network approach-based algorithm called Person Attribute Recognition (PAR) was developed. Our principal interest is in video analytics such as Person Re-Id [1], [2], Person Searching with attributes [3], [4] and Person retrieval [5]. It is helpful in specific situations

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

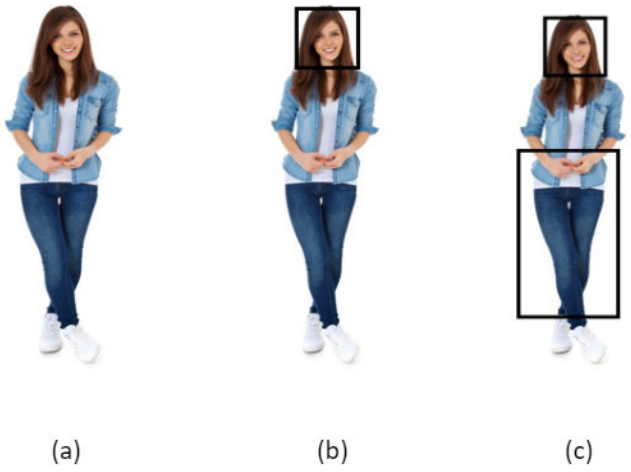

FIGURE 1. Recognizing the attribute Longhair, different approaches provide distinct attentive zones. (a)Input image. (b) selection based on attribute (c) Par-based method to select Body parts.

to discover criminals through video surveillance; numerous machines and vision duties combine algorithms associated with the attribute information into their algorithms to get higher performance in character Re-Identification and detection [6]. The unique elements that want to be considered in PAR are lighting fixtures prerequisites in which the Person's view can be distinct in one-of-a-kind situations, location of the individual, the taking pictures place and based on resolution. All modern PAR datasets for autonomous driving have limitations. Firstly, it contains less frequent attributes. Secondly, they have low pedestrian density; Thirdly, these datasets have limited variety.

In many real-world applications, It's important to obtain an expressive representation from the multi-view data.Semi-supervised learning framework that combines deep metric learning with density clustering can be effective for utilising the information included in unlabeled data [7].

While holistic methods typically rely on global characteristics, regional characteristics are essential for categorizing attributes. It makes sense that qualities could be localized in pertinent areas of an image. As shown in Fig1-b, it makes sense to concentrate on the regions associated with the head while identifying Longhair. Recent methods that try to take advantage of attention localization promote the learning of discriminative features for attribute identification. Traditional deep learning models suffer from issues like excessive calculation and slow timeliness when applied for multiple object recognition. A lightweight improvement technique based on the YOLOv4 algorithm can be deployed to address these issues and improve effectiveness [8]. Object detection, which tries to reliably detect targets with few visual cues in the image, has long struggled with small target detection, particularly in drone captured settings. Better in this case is YOLOv5 [9]. Using the visual attention mechanism to capture the essential elements is a well-liked approach [10]. In order to extract the attentive features, these techniques create attention masks from particular layers and multiply them with associated feature maps. These techniques extract regional features from the localized body components, such as the head, chest, and legs, as shown in Figure 1.

In the years before 2012, manually created feature selection methods. The distinctive techniques incorporate both global and regional elements. The process starts with feature extraction, then moves on to categorizing attributes. Approaches using Support Vector Machines (SVM) for classification have been widely used. But, according to the findings, applying these algorithms falls short of the acceptable performance standards for real-world applications. Convolutional neural networks are used in place of SVMs (Support Vector Machines) to address this issue. Convolutional Neural Networks (CNN) [12] were not initially able to localize the characteristics of various worldwide attire, such as hair and shirts. A person's features can now be learned, and their outfit can be localized via non-linear mappings with the help of recent developments in convolutional neural networks. These developments improved the results when Convolutional Neural Networks were used as the foundation. The Recurrent Neural Network (RNN) is an expanded network capable of performing the same function. The network architecture development activities involved in Natural Language Processing were substituted by RNNs (Recurrent Neural networks) that were produced. Superior outcomes. Transformer architectures are replacing CNN in this use case because they can overcome CNN's limitations because of the enormous advancements made in those designs. With the advances Several person attribute recognition methods that use multi-label classification and solely extract features from the entire input photos have been proposed for Convolutional Neural Networks. RNNs have taken the position of the network architecture's evolution in Natural Language Processing workloads since they produced better results [6]. Transformer architectures have undergone significant advancements and now outperform earlier versions [13], [14]. It's time to start employing transformers for classification and detection jobs. Compared to comparable CNN and image/vision Transformers on ImageNet, modest CoaT models achieve better classification results. Along with the analysis and evaluation of the Person Attribute Recognition on various architectures, the PAR datasets serve as an additional means of demonstrating the success of CoaT's foundation. The effectiveness of CoaT's backbone is also illustrated in the PAR datasets, along with the detailed survey.

An attribute-specific Attribute Localization Module (ALM) can automatically identify the exclusionary sections and extract region-based feature representations. The ALM comprises a small channel attention sub-network to fully leverage the inter-channel interdependence of the input features and a spatial transformer [15] to localize the attribute-specific regions adaptively. Below is a summary of the contributions made to work:

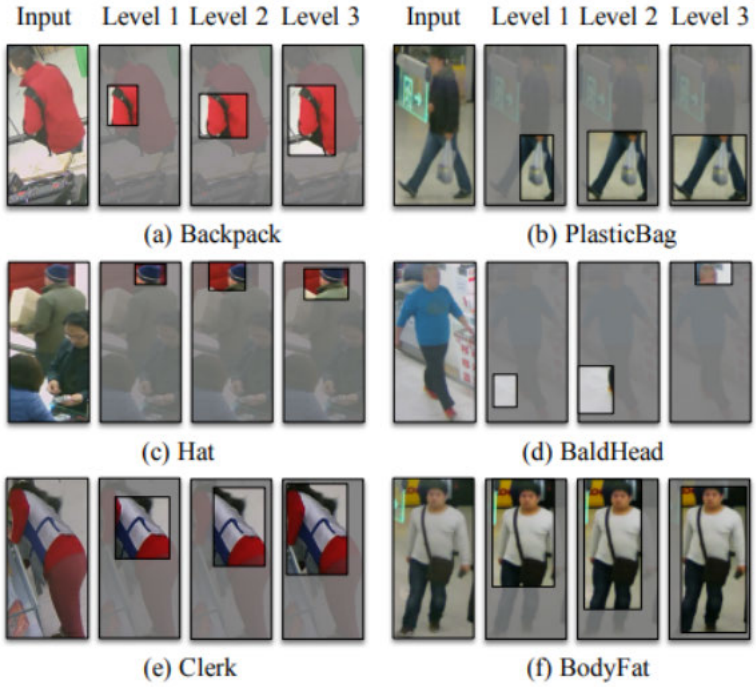

FIGURE 2. Visualization results [11].

- Compared with CNN and state-of-the-art techniques, transformers performed better than both in our extensive trials on transformers and Convolutional Neural Networks.
- This research proved that When compared to pure transformers, hybrid transformers deliver superior performance.
- Demonstrated that a lightweight transformer model with fewer parameters can produce results virtually identical to hybrid designs with more parameters.

The following breaks down the content into sections. The related study of person attribute recognition is presented in Section II. Sections III and IV covered the proposed hybrid methodology, and the experiment's analysis and findings were presented. The conclusion is found in Section V.

II. RELATED WORK

This individual's attribute detection (PAR) concentrates on the datasets and different techniques. They also contrasted the various attention models used to this specific use case.

A. PEDESTRIAN ATTRIBUTE RECOGNITION

The localization of characteristics serves as the basis of attribute-based recognition models. The limitations of attribute recognition are discussed in [16], and one of them is that handcrafted features do not perform well when tested in actual surveillance circumstances. They suggested a deep learning model (DeepSAR) that considers the attributes as one component and another model (DeepMAR) that jointly connects numerous attributes to solve the disadvantages associated with the localization of attributes. They have also suggested a weighted sigmoid cross entropy loss function to improve the model. Hydra Plus-Net or HP-Net was proposed, replacing CNN with multi-level attention mappings to several layers that feed multidimensional entities to identify the fine-grained personal characteristics crucial to PAR and Person Reid. Applications based on computer vision, such as disaster management systems using crowdsourced photographs, are increasingly using CNN algorithms. Reference [17] Sometimes the translation of the elements will not line up, and other times one or more variables are combined but may not be semantically related. Reference [18] presented a multitask deep model that extracts extensive feature representations using an element-wise multiplication layer to correlate the semantic relationship between a person's entire body attributes. Class Activation Maps (CAM) were introduced in [12], allowing us to see the characteristics of the picture representations that the CNN layers had captured. The pedestrian stance for person-attribute learning is explored for the first time in the PGDM [14]. The part areas are then extracted using these key points. They are both concatenated and separately extracted for attribute recognition. In addition, they create ROI recommendations for obtaining local features using Edge Boxes [19]. A straightforward visual attention technique with attribute level is an alternative method to learn the attribute feature more accurately. They have also added an attention loss function, which penalizes predictions in this visual attribute classification from attention masks with significant prediction variance to prevent destabilizing

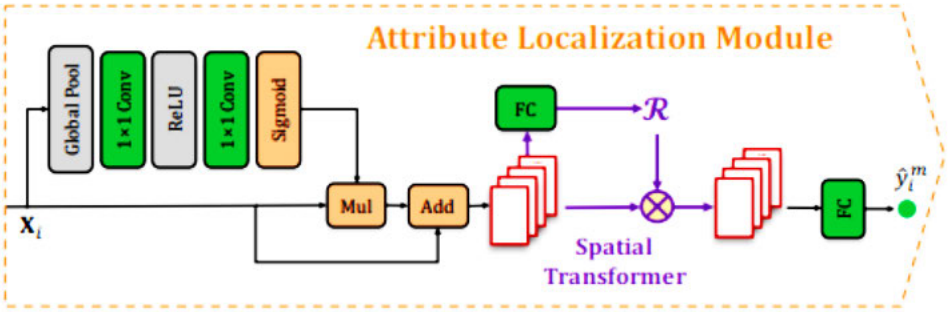

FIGURE 3. Region-based feature extraction using ALM [11].

training and reduce performance. Due to the significance of person re-identification in forensics and surveillance applications, it has received extensive research. In identification circumstances with a wide range of lighting, weather, or camera quality, gallery images are often high resolution (HR), while probe images are typically low resolution (LR). Reference [20] The visual attention regions for variety will also remain the same if the input image is spatially altered. However, such changes make CNN classifiers' visible attention areas less consistent. To solve this issue, [13] proposed a two-branch system with a novel stability loss that achieves state-of-the-art classification performance on the multi-label attribute classification, demonstrating the superiority of the proposed approach. Feature pyramid network (FPN) is the main subject [21]; with the aid of Squeeze-and Excitation (SE) blocks or (SENets), which introduce building blocks for CNNs (Convolutional Neural Networks) that increase channel interdependencies at no computational cost, Tang et al. [11] developed the Attribute Localization block. The Spatial Transformer Network (STN) [13] enhances attribute localization performance. Some methods in the person attribute field of work were concentrated on that line of optimization with a unique perspective on Attribute-based semantic relationships. A Joint Recurrent Learning (JRL) of attribute correlation and context has been developed in [13].

B. ATTENTION MODELS

The transformers were initially suggested for tasks involving natural processing language. For functions requiring detection and localization, they have since been enhanced. CNN backbones are replaced with attention layers. Both object identification and image classification tasks have used various attention modules. The aim is to use the different attention levels or transformers to enhance the current work. A multi-directional attention module was put forth by Li et al. [18] to learn multi-scale attentive elements for person analysis. With only image-level supervision, Sarafianos et al. [22] proposed a unified deep neural network that takes advantage of semantic and spatial relationships between labels. They carried the spatial regularization module [22] forward and refined it to develop efficient attention mappings at various scales. They had the spatial regularization module ahead and refined it to create efficient attention mappings at multiple scales. With the best PAR datasets currently available, such as RAPv1, PA100K, and PETA, PAR uses contemporary CNN architectures and suggested Transformer models to undertake extensive trials to improve prediction for person attribute recognition (PAR).

C. WEAKLY SUPERVISED ATTENTION LOCALIZATION

Without using region annotations, the goal of attention localization thoroughly researched in various visual tasks and pedestrian attribute recognition Jaderberg et al. [15]. Developed the renowned Spatial Transformer Network (STN), which can extract attentional regions from any spatial transformation in a trainable final manner.

D. VISUALIZATION OF ATTRIBUTE LOCALIZATION

The attribute areas are placed inside the feature maps in our approach. Figure2 depicts several instances for each of the six traits, including physical and abstract attributes. From this identification, despite the severe occlusions (a, c) or posture variations, the suggested ALMs can localize these tangible qualities, such as backpacks, Plastic Bags, into the corresponding regions (e). Figure2 illustrates a failure situation that is also offered (d).

E. ATTRIBUTE LOCALIZATION MODULES (ALM)

In Figure 3, the specifics of ALM are displayed. Only one Attribute at a single feature level is subject to attribute localization and region-based feature learning for each ALM [15]. The ALMs are trained under close supervision at various feature levels. The ALM generates an attribute-specific prediction from the input of combined features Xi. Every ALM only supports one characteristic at a time.

III. PROPOSED METHODOLOGY

Our design uses a co-scale mechanism transformer (CoaT) to image Transformers by keeping encoder branches at different scales while focusing attention on scales that are not adjacent. Other two sorts of construction blocks are introduced here, along with Conv-Attention. This model implements cross-scale, fine-to-coarse, and coarse-to-fine visual modeling. The parallel use of all four of these serial blocks

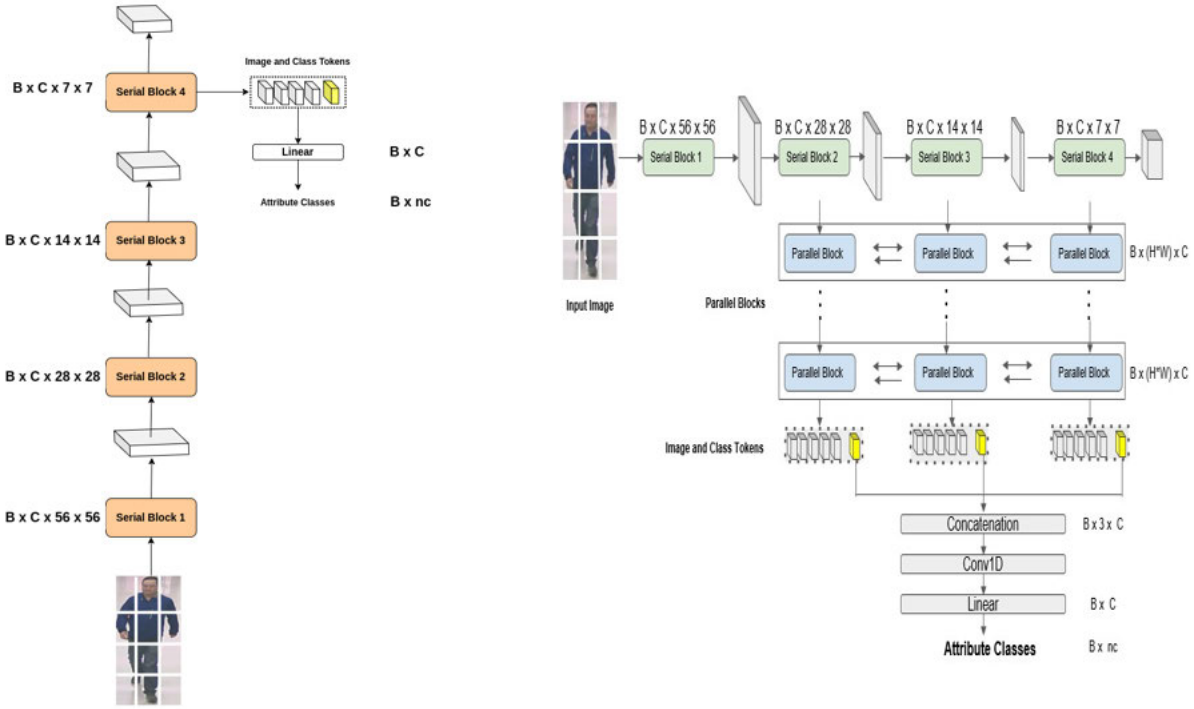

FIGURE 4. The CoaT-Lite architecture using only serial transformer (b) The complete CoaT architecture using parallel transformer.

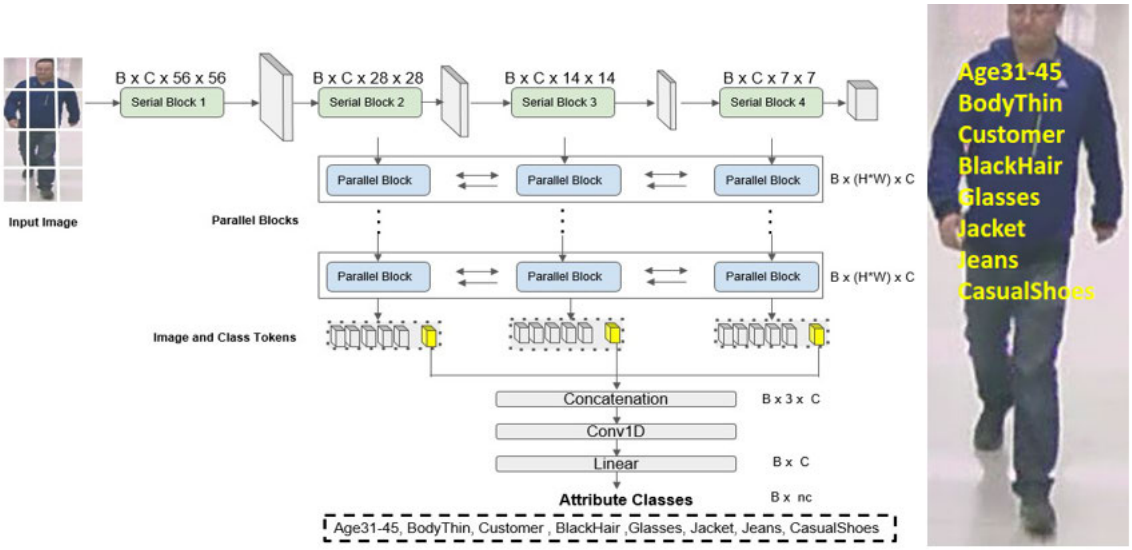

FIGURE 5. The CoaT-Lite architecture by combining serial transformer and parallel transformer.

is present here. Detach the CLS token from the image tokens to prepare them for the subsequent serial block. Within each parallel group of parallel blocks, successfully implement a co-scale mechanism. It has a standard parallel group of input feature sequences from serial blocks with different scales [23]. The parallel group, which uses two strategies, direct cross-layer attention and attention with feature interpolation, must interact between fine, coarse, and across scales. In order to support vision tasks, Vit and DieT inject absolute position embeddings, which could have issues mimicking relationships between local tokens, into the input [24].

The evaluation Protocol evaluates the performance of pedestrian attribute recognition by the widely used metric of mA, F1-Score, Accuracy, Recall, and Precision on RAP, PETA, and PA100K. The experimental results on different

CoaT levels, including Small, Mini, and Tiny, are reported unless stated otherwise.

A. PROPOSED NETWORK-ORIENTED MODELS

Deep learning algorithms are used for image processing in place of the formerly popular manual methods. The primary justification for this is that while manual approaches need user input, deep learning-based models can automatically extract features. Models based on deep neural networks (DNNs) for image processing applications could be quite complex. These DNN (Deep Neural Network) models contain numerous parameters. Image recognition can now be done in real-time due to advances in computing power, although improving outcomes is still a challenge.

1) CNN

An essential method for drawing out properties from images is the convolutional neural network (CNN) [6]. Due to the use of various function extraction phases that may automatically gain representations from data, Deep CNN offers excellent learning capabilities. Numerous potential deep CNN designs have recently been produced, propelled by the availability of enormous amounts of data and technological advancements [25]. Because it is becoming more popular and requires less work, this issue has much study potential. It also has a variety of uses in real-world settings. This would be helpful in a drone surveillance scenario. Applications for pedestrian attribute recognition tasks include soft biometrics, suspect person identification, criminal investigation, and public safety. In this study, our main aim is to identify pedestrian characteristics in various pedestrian circumstances. PETA, PA100K, and Richly Annotated Pedestrian (RAP) V1 and V2 datasets are used to conduct the experiments. Evaluated the CNN models with different pre-trained architectures such as ResNet18, Resnet34, Resnet50, Resnet101, DeepMAR, Mobilenetv2, and Densenet121 to produce a higher-performing architecture. Several pedestrians can be seen in one photograph in a real-world surveillance scenario. Therefore, the first stage in recognizing pedestrian attributes is identifying pedestrians in a setting with several pedestrians. In the methods, numerous CNN architectures are developed. CNN architectures are created by performing the object detection task. The detection of pedestrians is also done using these frameworks. Faster, R-CNN, Faster The double-stage CNN framework that can detect objects successfully is R-CNN and Mask R-CNN. The main issue with these frameworks is that they frequently run slowly for the two stages' extensive computing needs. The single-stage CNN framework SSD, YOLO, etc., which only contains one step and oversees the object detection duty, overcomes this difficulty. The next stage is to identify each pedestrian's characteristics. Due to an object's distinctive qualities, spatial details in an image allow us to remember it. A convolutional neural network is the best option for capturing the spatial characteristics necessary for object recognition. The photos must undergo some prepossessing to get the best performance out of a CNN architecture, such as resizing, scaling, augmentation, and normalization. Then, a fully connected, tailored layer called a classifier receives the spatial features. From the many pedestrian scenarios, the classifier predicts the pedestrian qualities of each pedestrian. A set of convolution and pooling procedures make up the CNN architecture. Some kernels apply the convolution operation on an image, which results in a classifier. Then, the feature map is given to an activation function to introduce nonlinearity [26]. Convolution operations are used to minimize spatial capture characteristics and image size. Then, a pooling operation on a specified window is carried out in which the maximum or average value from that window is taken, referred to as a max pool or average pool. Pooling operations are used to significantly lower calculation costs. Transformer-based models have recently gained popularity because they can dynamically depict visual representations [27]. This is shown in Figure 4.

2) TRANSFORMERS

The use of attention processes is now necessary for many different tasks. Conv-attention models, which are more effective than just CNN-based backbone models, have recently been introduced. The transformer model focuses on using the mechanism to effectively describe sequences, enabling the modelling of dependencies without taking into account their proximity to input or output sequences. The Conv-Attention transformers model is a transformer model used in our architecture (CoaT). The coat is made up of two blocks that were combined using different computation and accuracy features. A parallel block and a serial block are the two different forms of tiny structures that make up the interior. This transformer model (CoaT) can be combined as a serial transformer, a parallel transformer, or both which is shown in Figure 5. The architecture for modelling cross-scale images, coarse-to-fine images, and fine-to-coarse images is adjusted to achieve or facilitate this. Transformers with different names, such as CoaT-Tiny, CoaT-Mini, and CoaT-Small, were used for various scales. Four serial blocks are utilised here. Conv-attention and Feed Forward Network modules make up each serial block. Linear and drop-out layers make up the Feed Forward Network module. A serial block simulates low-resolution image representations [28]. In the CoaT architecture, added attribute which predicts multi class for single person. Advantages of this is to Easy of integration [i.e. to plug and add the additional modules]. A typical serial block flattens the reduced feature maps to produce a string of picture tokens. The dividend employing a patch embedding layer predetermines the input feature map after downsampling. They then integrate the image token with the additional CLS token utilising some centralised modules, as demonstrated in section IV, a specialised vector used to categorise and earn the fundamental linkages between image tokens and CLS tokens. The image tokens are then transformed into 2D

TABLE 1. Results of comparison with different Model [6].

		HP-Net [22]	PGDM [14]	VeSPA [15]	LG-Net [5]	ALM [15]	Coat-Tiny	Coat-Mini	Coat-Small
	params	24M	87.2M	17M	-	17.1M	5.5M	10M	22M
	Flops	-	1G	3.5G	-	1.95G	4.4G	6.8G	12.6G
RAPv2	mA	-	-	-	-	78.79	78.21	**79.64**	79.22
	Accuracy	-	-	-	-	63.58	65.43	**66.45**	64.97
	Precision	-	-	-	-	72.76	76.93	**77.01**	75.82
	Recall	-	-	-	-	81.45	79.42	**81**	79.91
	F1 Score	-	-	-	-	76.86	78.16	**78.95**	77.81
RAPv1	mA	76.12	74.31	77.7	78.68	79.82	78.02	79.67	**80.71**
	Accuracy	65.39	64.57	67.35	68	64.71	**79.81**	67.91	67.43
	Precision	77.33	78.86	79.51	**80.36**	72.84	77.09	79.25	77.43
	Recall	78.79	75.9	79.67	79.82	**83.73**	78.59	80.83	82.23
	F1 Score	78.05	77.35	79.59	**80.09**	77.91	77.84	80.03	79.76
PETA	mA	81.77	82.97	83.45	-	84.22	81.8	84.1	**84.75**
	Accuracy	76.13	78.08	77.73	-	76.4	74.79	76.1	**79.56**
	Precision	84.92	86.86	86.18	-	82.94	83.62	84.43	**87.98**
	Recall	83.24	84.68	84.81	-	**86.94**	83.78	84.8	86.06
	F1 Score	84.07	85.76	85.49	-	84.89	83.7	84.62	**87.01**
PA100K	mA	74.21	74.95	76.32	76.96	80.99	81.68	83.3	**83.47**
	Accuracy	72.19	73.08	73	75.55	78.81	78.51	80.7	**81.1**
	Precision	82.97	84.36	84.99	86.99	86.34	86.23	87.75	**87.89**
	Recall	82.09	82.24	81.49	83.17	88.01	87.78	89.24	**89.44**
	F1 Score	82.53	83.29	83.2	85.04	87.17	87	88.49	**88.66**

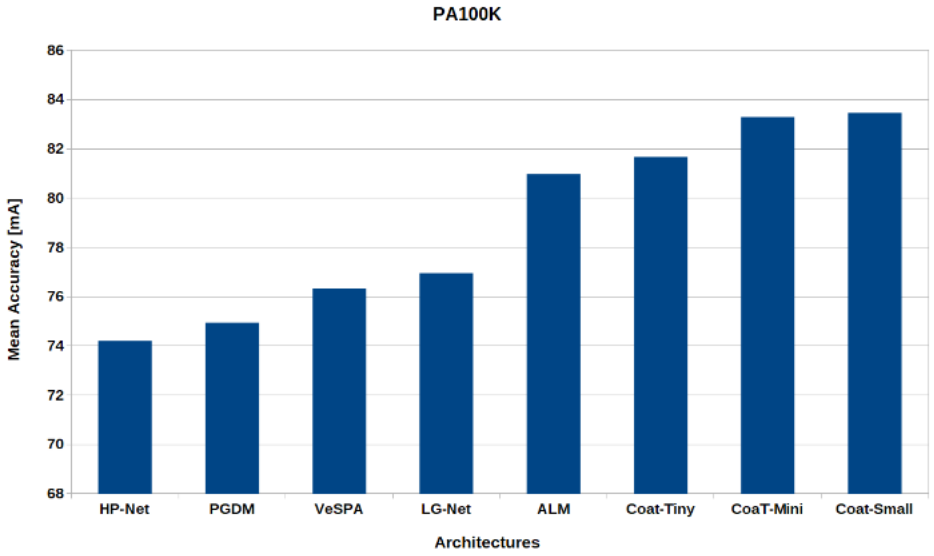

FIGURE 6. Comparison of various architecture mean accuracy concerning PA100K Dataset.

feature representations in preparation for the following serial block after the CLS token has been separated from them [29]. Each parallel group of parallel blocks' co-scale mechanism has been implemented successfully. In a typical parallel group, there are several input vectors, including the picture and CLS tokens from serial blocks with different scales. The parallel group ploys two techniques need to be able to interact between scales that are fine, coarse, and across scales. Direct cross-layer and feature interpolated attention are the two approaches to achieve this [6]. Vision Transformers and a data-efficient image transformer are used to support vision tasks by introducing absolute position embedings, which may have trouble reproducing relationships between local tokens. Instead, incorporate a relative position encoding with window size to obtain the relative attention map [6]. The binary classification is done using the loss function i.e., log Loss or binary cross entropy equation 1 shown below:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i.log(p(y_i)) + (1-y_i).log(1-(p(y_i)) \tag{1}$$

where y is the class denoted by 0 and 1(1 for attribute class present and 0 for attribute class absent), p(y) is the predicted probability of the attribute class present over all N samples. The formula evaluates each attribute class present (y=1) and

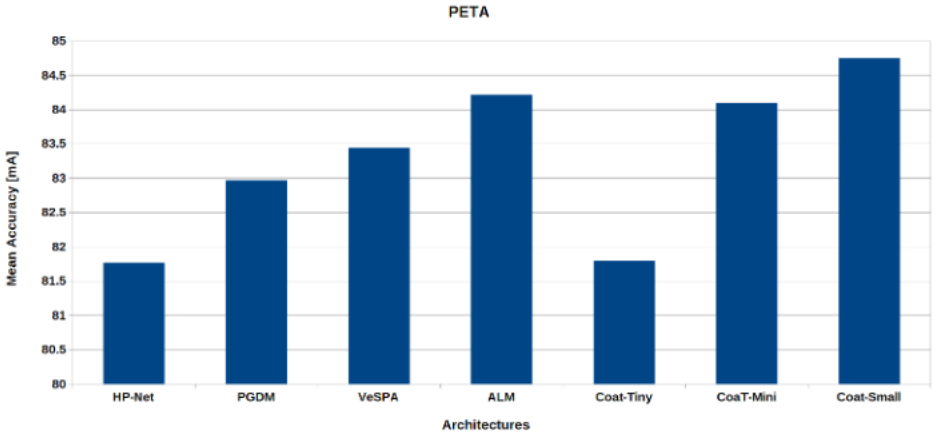

FIGURE 7. Comparison of multiple architectures mean accuracy concerning PETA Dataset.

TABLE 2. Details of existing datasets.

Dataset	Attributes	Total Number of Images
$RAPv2$	60	84,928
$RAPv1$	51	41,585
$PETA$	35	19,000
$PA-100K$	26	1,00,000

adds the loss of log probability of the attribute class present. For each attribute class absent (y=0), adds log probability of it being attribute class absent

IV. RESULTS AND EXPERIMENT

DATASETS: The proposed work has done several ablation studies on our baseline approach to show how different variables affect the datasets. The implementation flow comprises the following. First, evaluate the standard datasets with the current models. We used our baseline model (CoaT) and plotted the results. Finally, compare the results between the existing methods used for Person Attribute Detection (PAR) and the transformer models and tabulate the results on different scales. So, on a high level, we level several experiments on our baseline method to demonstrate the effect of various factors on the datasets. The experiments have been performed on the standard datasets:

- Richly Annotated Pedestrian Version 1 (RAPV1) [15]
- Richly Annotated Pedestrian Version 2 (RAPV2) [15]
- Pedestrian Attribute (PETA) [30]
- PA-100K [10]

Here the results on the four datasets are evaluated after passing them to the CoaT models across various scales. The distribution of the datasets used is given in Table 4. The implementation of the CoaT model and its results are shown in Table 4. The experimental results of the datasets used are displayed in Table 4.

Table 2 displays the benchmarking outcomes of our baseline CNN Transformer model on three sizable attribute datasets, including RAPv1, RAPv2, PETA, and PA100K datasets. On the PA100K and PETA datasets, our baseline model outperformed specially crafted pedestrian attribute identification methods in terms of performance. It's interesting to note that when dataset sizes grow, the performance gap between our primary way and cutting-edge algorithms shifts—observed that we see a comparatively significant improvement in the dataset of PA100K, which contains the most attributes.

A. DATASET PA-100

This dataset's image includes precise outdoor surveillance. There were 598 cameras used to record these. The PA100K dataset consists of one lack photos with a range of views, occlusion, and part localization, and 26 binary attributes. The resolution of the images ranges from 50 × 100 to 758 × 454. There are 80,000, 10,000, and 10,000 impressions allocated for training, testing, and validation, respectively. Comparision chart is drawn in Figure 6.

B. DATASET PETA

The PETA dataset was created using ten publically accessible datasets for Person Re-ID research. The PETA dataset consists of 19,000 photos from various angles with resolutions ranging from 17 × 39 to 169 × 365 and 61 binary attributes. 11,400 shots are in the train set, and 7,600 are in the validation set. Comparision chart is drawn in Figure 7.

C. DATASET RAP

This dataset's photos were gathered from actual surveillance situations including individuals inside. 26 cameras were set up to collect this data. It has 69 binary properties and 41,585 images with various views, occlusions, and Part Localization. The resolution of the images ranges from 36×92 to 344×554. 33,268 images are in the train set, while 8,317 images are in the validation set. Comparision chart is drawn in Figure 8.

These methods are implemented using the Py Torch framework using pre-trained ImageNet models. Person Images are resized to 224 × 224. Here we used 'Adam' as our opti-

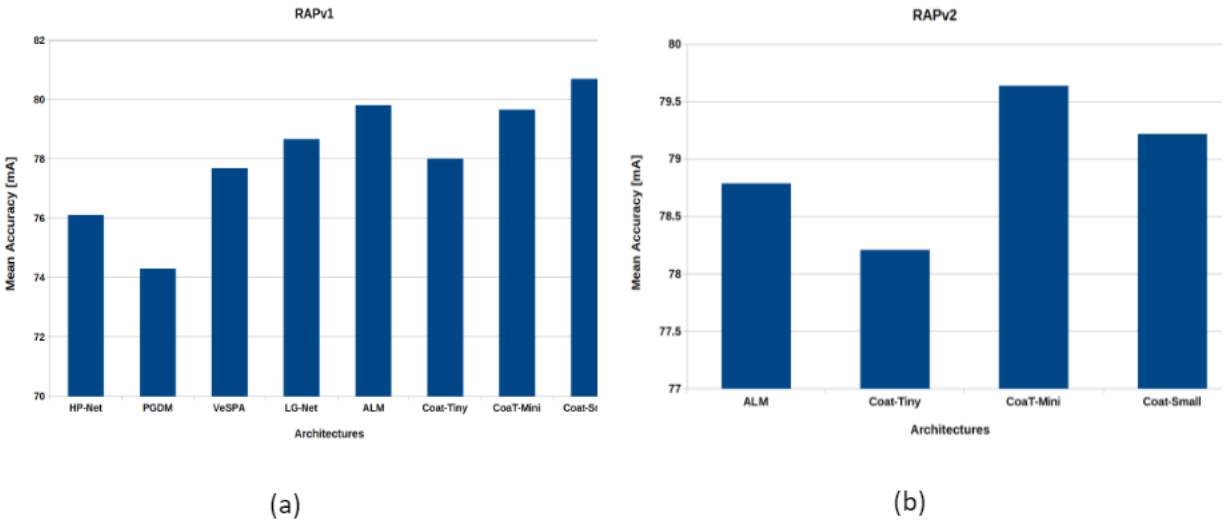

FIGURE 8. (a) Comparison of different architectures means accuracy concerning RAPv1 Dataset (b) Comparison of various architectures mean accuracy for RAPv2.

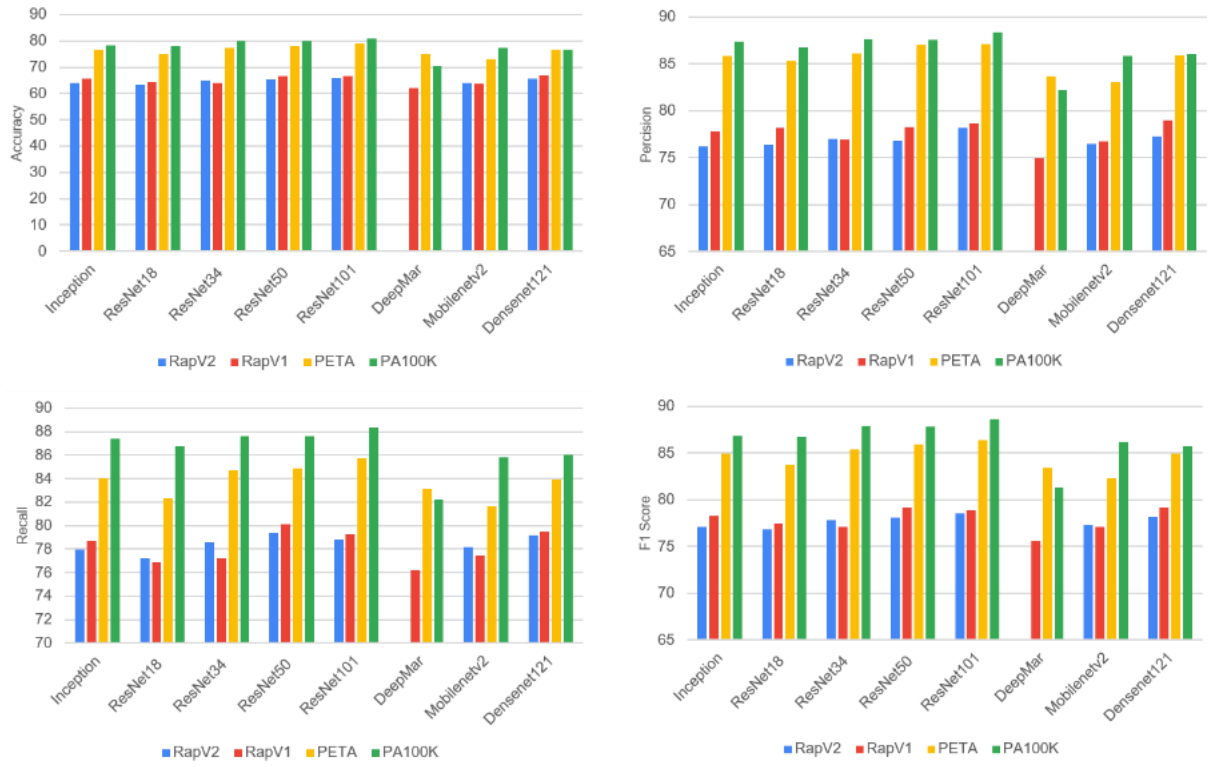

FIGURE 9. (a) Comparison of various architecture mean accuracy for Dataset. (b) Comparison of different architecture precision for Dataset. (c) Comparison of diverse architecture Recalls for Dataset (d) Comparison of various architecture F1 Score for Dataset.

mizer because it performs faster than SGD. The parameter momentum is set to 90×10^{-1}, and the weight decay is initialized as 5×10^{-4}. The learning rate is 1 x $10^{-4}$with plateau learning. The rate scheduler and the batch size are given as 64. For the RAP, PETA, and PA-100K datasets, the model is trained for 30 epochs. For training used the Tesla P100 GPU system [6].

We experimented with numerous CNN backbones and different transformer designs. The four separate datasets were used to analyze the results using various models. CNN experimented with various CNN models in this architecture, including BnInception, ResNet variations, HP-Net, PGDM, MobileNetv2, DeepMar, VESPA, LG-Net, and ALM. Each of these models has a unique set of parameters, ranging in

TABLE 3. Results of comparison with different model.

		BnInception	ResNet18	ResNet34	ResNet50	ResNet101	DeepMap	Mobilenetv2	Densenet121
	params	11.3M	11.174M	21.282M	25.6M	44.5M	58.5M	2.4M	8M
	Flops	1.78G	2G	4G	4G	8G	0.72G	0.3G	3G
RAPv2	mA	77.25	74.61	76.9	77.29	77.57	-	73.54	**78.29**
	Accuracy	63.92	63.54	64.9	65.26	**66.09**	-	64.18	65.51
	Precision	76.21	76.43	77.06	76.81	**78.23**	-	76.49	77.26
	Recall	77.94	77.26	78.57	**79.39**	78.85	-	78.15	79.21
	F1 Score	77.07	76.84	77.81	78.08	**78.54**	-	77.31	78.23
RAPv1	mA	78.66	75.34	77.44	79.64	78.49	73.79	75.83	78.27
	Accuracy	65.55	64.42	64.09	66.77	66.54	62.02	63.82	**66.82**
	Precision	77.87	78.15	76.89	78.25	78.62	74.92	76.73	**78.98**
	Recall	78.71	76.84	77.26	80.13	79.26	76.21	77.42	**79.46**
	F1 Score	78.28	77.49	77.08	79.18	78.94	75.56	77.07	**79.22**
PETA	mA	82.63	80.99	83.43	83.71	**84.79**	82.89	80.23	82.59
	Accuracy	76.68	75.09	77.42	78.07	**78.82**	75.07	73.04	76.79
	Precision	85.84	85.3	86.1	87.02	**87.08**	83.68	83.02	85.96
	Recall	84.03	82.36	84.75	84.85	**85.73**	83.14	81.66	83.91
	F1 Score	84.92	83.8	85.42	**85.92**	86.4	83.41	82.33	84.92
PA100K	mA	79.13	79.82	81.81	81.99	**82.69**	72.7	79.85	78.53
	Accuracy	78.19	78.12	79.98	79.93	**81.04**	70.39	77.3	76.59
	Precision	87.42	86.77	87.64	87.59	**88.34**	82.24	85.85	86.04
	Recall	86.21	86.65	88.24	88.11	**88.9**	80.42	86.59	85.41
	F1 Score	86.81	86.71	87.94	87.85	**88.62**	81.32	86.22	85.73

FIGURE 10. Identify the objects using Class Activation Maps.

size from 11.3M to 87.2M. The ResNet101 model has the most flops, and MobileNetv2 has the fewest losses. Table 4 displays the benchmarking outcomes of our baseline CNN Transformer model on three sizable attribute datasets, including RAPv1, RAPv2, PETA, and PA100K datasets. The following figure gives the Comparison of various architecture mean accuracy for PA100K Dataset. Table 4 indicates that when all architectures are considered, the PA100K architectures have a robust F1 Score, Accuracy, Precision and Recall are 86.81, 78.19, 87.42, 86.21 for the BnInception model—similarly, calculated for other models. ResNet101 Accuracy and precision is high that is 66.09 and 78.23 for RAPv2 Dataset. Similarly All the calculated values and Highest value(Represented in Bold)are represented in the Table 4. Another advantage of the Transformers is the ability to learn attributes based on semantic relationships and observed that the serial transformers are obtaining better metrics when compared to the PA100K dataset. The transformers can attain improved accuracy and F1 Score even with fewer parameters. This provides an additional benefit when training with other models as well. This Observation is observed in Figure9. For the proposed Architechture We calculated the speed such that 26ms,28ms and 29ms for Coat-Tiny,CoaT-Mini and Coat-Small respectively.

- Added Serial CoaT-Architecture variants [Coat-Lite-Tiny, Coat-Lite-mini, Coat-Lite-Small, Coat-Lite-Medium] in addition to the existing Parallel Variants [Coat-Tiny, Coat-Mini, Coat-Small] which has better performance and accuracy compared to the parallel variants that is represented in Figure 11, Figure 12, Figure 13 and Figure 14 respectively.
- Extensive training and evaluation of CoaT architecture with various CNN architectures like Resnet Variants [R18, R34, R50, R101], Bn-Inception, DeepMar, Mobilenetv2 and DenseNet121 and proving that our Coat Architecture is better than the existing CNN architectures.
- Class Activation Maps (CAM) - A class activation map for a particular category indicates the discriminative image regions used by the CNN to identify that category. CAM is a weighted activation map generated for each

TABLE 4. Results of comparison with different model by adding serial CoaT-Architecture variants.

		Coat-Tiny	Coat-Mini	Coat-Small	Coat-Lite-Tiny	Coat-Lite-mini	Coat-Lite-Small	Coat-Lite-Medium
	params	5.5M	10M	22M	5.7M	11M	20M	45M
	Flops	4.4G	6.8G	12.6G	1.6G	2G	4G	9.8G
RAPv2	mA	78.21	**79.64**	79.22	77.21	76.79	76.79	79.33
	Accuracy	65.43	66.45	64.94	65.59	65.5	65.53	**66.75**
	Precision	65.43	66.45	64.94	77.39	**77.46**	77.42	76.75
	Recall	79.42	81	79.91	79.22	79.08	79.08	**81.79**
	F1 Score	78.16	78.95	77.81	78.3	78.26	78.26	**79.19**
RAPv1	mA	78.02	79.67	80.71	78.09	77.23	79.57	**81.14**
	Accuracy	**79.81**	67.91	67.43	65.59	65.07	66.71	68.12
	Precision	77.09	**79.25**	77.43	77.4	78.02	78.22	78.89
	Recall	78.59	80.83	**82.23**	79.11	77.71	79.84	81.57
	F1 Score	77.84	80.03	79.76	78.25	77.87	79.02	**80.2**
PETA	mA	81.8	84.1	84.75	81.29	82.24	83.69	**86.34**
	Accuracy	74.79	76.1	79.56	75.49	75.28	78.14	**80.91**
	Precision	83.62	84.43	87.98	85.02	84.13	86.87	**88.19**
	Recall	83.78	84.8	86.06	83.23	83.88	85.13	87.59
	F1 Score	83.7	84.62	87.01	84.12	84	85.99	**87.89**
PA100K	mA	81.68	83.3	83.47	80.03	80.3	83.08	**84.29**
	Accuracy	78.51	80.7	81.1	78.45	77.88	81.24	**81.7**
	Precision	86.23	87.75	87.89	86.85	86.44	**88.03**	87.93
	Recall	87.78	89.24	89.44	87.02	86.78	89.45	**90.26**
	F1 Score	87	88.49	88.66	86.93	86.61	88.73	**89.08**

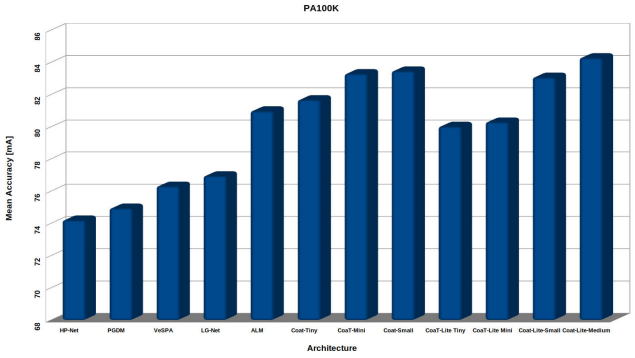

FIGURE 11. Accuracy concerning PA100K Dataset by added features.

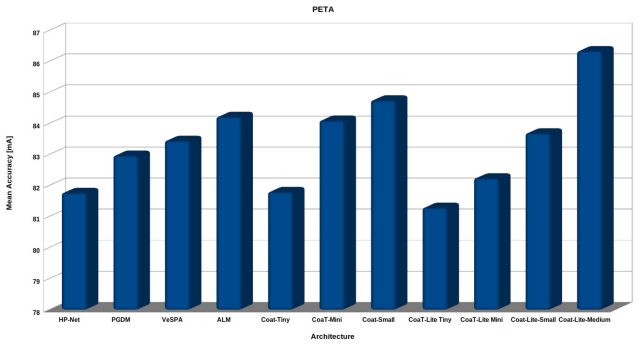

FIGURE 12. Accuracy concerning PETA Dataset by added features.

image. It helps to identify the region a CNN is looking at while classifying an image. In our paper we provide the visual evaluation of the model I.e., model predictions are performed by looking at the right position in an image for example to predict the upper wear it is looking in the upper portion of the image and to predict footwear it is looking in the lower portion of the image as shown in Figure 10.

From this we can observe that with little time we got better Accuracy for different Dataset.

Further came forward to perform some ablation study on CoAT-Architecture. the used mechanism is Model Interpretability (MI) approaches such as Eigen CAM, Grad CAM, Score CAM, Eigen CAM, etc. To learn and understand the underlying operations performed or decisions taken by the model such as the focusing region to predict a particular class by using MI methods. This helps to get more information into the learning of the network and also helps in debugging because the user gets the localization of the predicted attributes without having to explicitly label the object. They

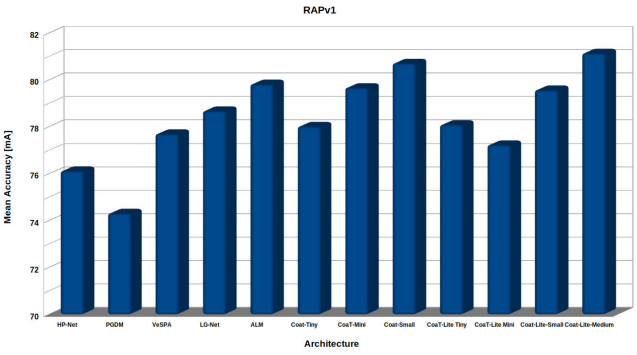

FIGURE 13. Accuracy concerning RAPv1 Dataset by added features.

have the ability to analyze and visualize each layer's features and results in the focusing regions. There are many MI methods available that help us to visualize the features or labels using Class Activation Maps(CAM) are shown in Figure 10. It is used for different multi-purpose scenarios depending on the visualize different class labels. We demonstrate the Eigen

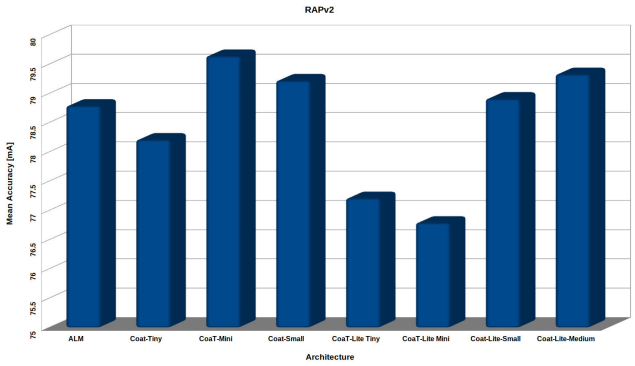

FIGURE 14. Accuracy concerning RAPv2 Dataset by added features.

CAM capability in detecting CAM features. For better capability in localizing the discriminated regions of the images and their respective attributes Eigen- CAM gives better consistency in detecting single and multi- attributes mainly in crowded and other pedestrian scenarios. Eigen CAM used to evaluate the architecture. Eigen CAM was applied to the test set. This Eigen CAM approach has a better perspective of the model learning and its validations.

V. CONCLUSION

In this paper, research is focused on Person attribute recognition. The most discriminative region was found using a hybrid method that included a convolutional neural network with transformers. It introduced serial and parallel blocks for person attribute recognition which consists of conv-attention, feed-forward network, attention with feature interpolation, and direct cross-layer attention. The hybrid model compared the generalizability of convolutional neural networks with that of more recent transformative networks on four different datasets. Our baseline model outperformed specially designed pedestrian attribute identification algorithms on the PA100K and PETA datasets. A direct dataset examination demonstrated that the transformers perform better than CNN. The experiment results are observed to have outperformed when compared to existing methods. The transformers can attain improved accuracy and F1 Score even with fewer parameters. We observed that the serial transformers are obtaining better metrics when compared to the PA100K dataset. This provides an additional benefit when training with other models as well. The extensive analysis suggests the most informative region and accurate results. In the future, we plan to analyze the hybrid methods with standard state-of-art architecture. Better application opportunities for the upgraded YOLOv4 snd YOLOv5 in the multi-object recognition challenge [8].

REFERENCES

[1] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.

[2] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Jan. 2019.

[3] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti, "Attribute-based people search: Lessons learnt from a practical surveillance system," in *Proc. Int. Conf. Multimedia Retr.*, Apr. 2014, pp. 153–160.

[4] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *Proc. CVPR*, Jun. 2011, pp. 801–808.

[5] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019.

[6] S. Abhilash and V. M. Nookala, "Person attribute recognition using hybrid transformers for surveillance scenarios," in *Proc. Int. Conf. Distrib. Comput., VLSI, Electr. Circuits Robot.*, Oct. 2022, pp. 186–191.

[7] X. Jia, X.-Y. Jing, X. Zhu, S. Chen, B. Du, Z. Cai, Z. He, and D. Yue, "Semi-supervised multi-view deep discriminant representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2496–2509, Jul. 2021.

[8] X. Huang, S. Hu, and Q. Guo, "Multi-object recognition based on improved YOLOv4," in *Proc. CAA Symp. Fault Detection, Supervision, Saf. Tech. Processes*, Dec. 2021, pp. 1–4.

[9] K. Ding, X. Li, W. Guo, and L. Wu, "Improved object detection algorithm for drone-captured dataset based on YOLOv5," in *Proc. 2nd Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2022, pp. 895–899.

[10] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.

[11] C. Tang, L. Sheng, Z.-X. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4997–5006.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012.

[13] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 87–95.

[14] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.

[16] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 789–792.

[17] P. Vallimeena, U. Gopalakrishnan, B. B. Nair, and S. N. Rao, "CNN algorithms for detection of human face attributes—A survey," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 576–581.

[18] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 111–115.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1150–1157.

[20] X.-Y. Jing, X. Zhu, F. Wu, R. Hu, X. You, Y. Wang, H. Feng, and J.-Y. Yang, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1363–1378, Jan. 2017.

[21] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[22] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 680–697.

[23] J. Wu, H. Liu, J. Jiang, M. Qi, B. Ren, X. Li, and Y. Wang, "Person attribute recognition by sequence contextual relation learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3398–3412, Oct. 2020.

[24] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, "Conditional positional encodings for vision transformers," 2021, *arXiv:2102.10882*.

[25] H. Li, Y. Chen, D. Tao, Z. Yu, and G. Qi, "Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1480–1494, 2021.

[26] X. Chen, J. Weng, W. Lu, and J. Xu, "Multi-gait recognition based on attribute discovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1697–1710, Jul. 2018.
[27] X. Zheng, H. Sun, X. Tian, Y. Li, G. He, and F. Fan, "Attribute memory transfer network for unsupervised cross-domain person re-identification," *IEEE Access*, vol. 8, pp. 186951–186962, 2020.
[28] R. Abbaszadi and N. Ikizler-Cinbis, "Merging super resolution and attribute learning for low-resolution person attribute recognition," *IEEE Access*, vol. 10, pp. 30436–30444, 2022.
[29] V. Mirjalili, S. Raschka, and A. Ross, "PrivacyNet: Semi-adversarial networks for multi-attribute face privacy," *IEEE Trans. Image Process.*, vol. 29, pp. 9400–9412, 2020.
[30] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, vol. 38. PMLR, 2015, pp. 562–570. [Online]. Available: https://proceedings.mlr.press/v38/lee15a.html

S. RAGHAVENDRA received the bachelor's degree in computer science and engineering from the BMS Institute of Technology, Bengaluru, the master's degree from the R. V. College of Engineering, Bengaluru, and the Ph.D. degree from the University Visvesvaraya College of Engineering, Bengaluru. He is currently working as an Assistant Professor with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal. He has 11 years of teaching and research experience in various institutes. He has authored more than 40 research publications. His research interests include cloud computing, machine learning, and the Internet of Things. He is currently serving as an Editorial Board Member, a Reviewer, and a Guest Editor for a number of prestigious journals of publishers, such as IEEE, Elsevier, Springer, Wiley, Taylor & Frances, and KJIP. He is serving as the Publication Chair for Discover-21. He was an Organizing Committee Member for conferences such as ICCN-14, ICCN 15, ICCN-16, ICInPro-18, DISCOVER-19, ICInPro-2019, and Discover-20. He also served as a Joint Secretary of the IEEE Mangalore Sub-Section, in 2020, and worked as the Website Co-Chair of IEEE MSS, in 2019.

RAMYASHREE received the B.E. degree in computer science and engineering from SMVITM, Bantakal, in 2015, and the M.Tech. degree in computer science and engineering from NMAMIT, Nitte, in 2019. She worked as an Assistant Professor at the Department of Computer Science and Engineering Department, SMVITM, from 2015 to 2017 and from 2019 to 2021. She also worked as an Assistant Lecturer at NITK, Surathkal. She is currently working as an Assistant Professor with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal. She has published more than ten papers in various conferences and journals and one book chapter. Her research interests include image and video processing, artificial intelligence, and machine learning. She can contribute toward engineering education and research in the area of computer science and professional service activities.

S. K. ABHILASH received the M.Tech. degree in electronics and communications from the University Visveswaraya College of Engineering (UVCE), Bengaluru. He is currently a Technical Lead with KPIT Technologies, Bengaluru. He has published journals and conference papers in preferred international conferences. He has also obtained a patent to his credits. His areas of research interests include automotive driver assistance systems (ADAS), surveillance analytics systems, and building unified agnostic-based computer vision frameworks. His contribution to the architecture and algorithm design at KPIT has been vital.

VENU MADHAV NOOKALA received the B.Tech. degree in electronics and communications from the Amrita Vishwa Vidyapeetham, Coimbatore. He is currently a Software Engineer with KPIT Technologies, Bengaluru. He has published conference papers at reputed international conferences. His areas of research interests include automotive and surveillance analytics systems, developing AI-based tools based on graphical user interfaces, deployment of computer vision models on various frameworks, and developing the architectures.

S. KALIRAJ received the B.E., M.E. (Hons.), and Ph.D. degrees from Anna University, Chennai, Tamil Nadu, India. He is currently a Senior Assistant Professor with the Department of Information and Communication Technology, MIT Manipal, Manipal Academy of Higher Education (Institution of Eminence), India. He has completed two industry certifications, MCTS (Microsoft Certified Technology Specialist) and the EMC Academic Associate, Data Science and Big Data Analytics. He has published four patents and more than 25 research papers covering all major areas of software engineering, machine learning, and data science in top journals and conferences. His area of research interests include verification of machine learning systems, fault prediction and localization, data science, machine learning applications in society, NLP, and software testing. He has guided more than 35 students in their master's and undergraduate research. He has served as the session chair and a member of the Advisory Committee and Technical Committees of various international conferences. He has acted as a resource person for the faculty development programs, workshops, guest lectures, and conferences organized by various institutions and universities. He was a reviewer of Scopus and WOS-indexed international journals in his area of research.