CLIPping Imbalances: A Novel Evaluation Baseline and PEARL Dataset for Pedestrian Attribute Recognition

Kamalakar Vijay Thakare[1], Lalit Lohani[3], Kamakshya Prasad Nayak[3], Debi Prosad Dogra[3], Heeseung Choi[1,2], Hyungjoo Jung[1] and Ig-Jae Kim[1]

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

[2]Yonsei-KIST Convergence Research Institute, Yonsei University, Seoul 03722, Republic of Korea

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

Abstract

Pedestrian Attribute Recognition (PAR) serves as a fundamental task in computer vision and is crucial for upgrading security systems. It helps in precisely identifying and characterizing various attributes of pedestrians. However, current PAR datasets have certain issues in representing a wide range of attributes correctly, which makes the existing PAR methods less effective in real-world scenarios. Addressing this limitation, this paper introduces PEARL, a comprehensive dataset comprising of diverse pedestrian images annotated with 146 attributes. These samples have been sourced from surveillance videos across twelve countries. This paper also formulates an image-based PAR using language-image fusion strategy and utilizes CLIP as a new evaluation baseline. Specifically, we leverage textual information by transforming sets of attributes into meaningful sentences. Addressing the inherent data imbalance in PAR, we provide three types of prompt settings to optimize the training of the CLIP model. Our evaluation encompasses a thorough assessment of the proposed baseline model across various datasets, including PEARL dataset as well as established PAR benchmarks such as PA100K, RAP, and PETA.

1. Introduction

Pedestrian Attribute Recognition (PAR) aims to identify and characterize specific attributes such as *gender*, *accessories*, or *body posture* of pedestrians, upgrading the precision and context-awareness of security systems. In recent years, PAR has become a pivotal problem in the field of computer vision. This advancement has augmented the implementation of essential applications such as person re-identification [8,9,35], person retrieval [1,25,38], and scene analysis [34]. Despite the appealing success of recent PAR approaches [3, 11, 27, 31, 37], their effectiveness is constrained by two major problems. Firstly, the lack of expansive dataset restricts the ability of models to properly recognise a wide range of features at varying scales. Secondly, the inherent issue of data imbalance within the datasets hampers the generalization and robustness of PAR models.

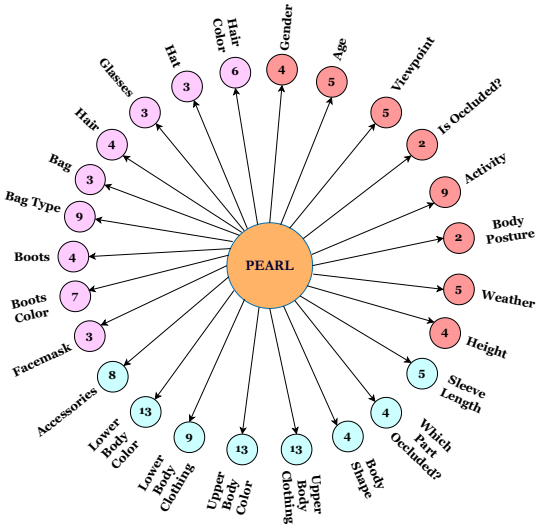

Figure 1. **PEARL: PE**destrian **A**ttribute **R**ecognition and **L**earning using **30K** images is a large-scale pedestrian attribute dataset that provides rich annotations via **25** critical attribute categories spanning over **146** sub-attributes. It has been extracted from the surveillance videos of **seven** common public places like *airports*, *stations*, and *parks* across **twelve** countries. Numbers mentioned inside the nodes represent attribute options. For instance, *Hair* has four options *Hair-short*, *Hair-long*, *Hair-Bald*, *Hair-NOB* (Not Observable). PEARL dataset available at https://github.com/draxler1/PEARL30K

The existing PAR datasets [4, 15, 18, 20] are commendable; however, they also exhibit deficiency in comprehensively representing the diversity in data. A handful of approaches attempt to unify existing datasets by adding additional annotations. Deng *et al.* [4] have accumulated 19K samples from 10 surveillance datasets and released the PETA dataset. Specker *et al.* [27] have combined

RAPv2 [15], PETA [4], PA100K [20] and Market1501 [18] and released the UPAR dataset. Despite such genuine efforts, existing datasets still pose notable limitations. Firstly, they are highly restricted to specific recording locations, resulting in the dominance of certain attributes. For example, Market1501 [18] dataset consists of 94.4% images belonging to *short sleeves*. Secondly, the presence of inaccurate guidelines during annotation introduces challenges. For instance, range of the *age* attribute can vary widely, yet PETA [4] dataset assigns *age18-60* attribute, resulting in over 73% of samples falling within this broad range. Lastly, the context-dependent nature of these datasets raises a significant concern regarding their generalization. For instance, a model trained on a dataset [4, 15, 20] recorded in sunny condition may confuse a model to recognize attributes in other weather conditions such as snow or rain. As a consequence, prevalent attributes and context-dependency leads to overfitting and hampers the performance in test scenarios.

However, a few PAR approaches have correctly identified *data imbalance* problem in PAR. Li *et al*. [13] have derived a weight factor of a specific attribute using a positive ratio and multiplied it with cross-entropy loss to effectively train the model. Yan *et al*. [37] have proposed drop loss, where attribute-based drop rate is used to subside hard samples and trained the model in easy-to-hard fashion. Tan *et al*. [28] have employed constrained loss by applying penalty coefficient to the binary cross entropy loss. In addition to this, a handful of approaches exploit correlation between attributes [11, 27, 28], utilize visual-textual features as multi-model input [1, 2, 39]. And a few of them employ GCN [17, 28], knowledge-distillation [16, 36], and recurrent networks [40]. While these approaches strive to improve recognition performance through complex architectures, the critical shortcomings still persist. The imbalance and redundancy inherent in dataset structure remains unaddressed. Moreover, these approaches fail to produce robust results as their sufferings are two-fold: (i) correlation between the attributes does not hold true always as appearance is context-dependent, (ii) attribute locations in an image may vary due to different body postures.

To effectively address these challenges, an essential solution lies in the creation of a large-scale and highly diverse dataset encompassing heterogeneous set of attributes. In this paper, we release **PEARL**, **PE**destrian **A**ttribute **R**ecognition and **L**earning dataset comprises of 30K pedestrian images, each labelled with 146 critical sub-attributes. Derived from surveillance videos spanning 25 hours and recorded across twelve distinct countries, PEARL showcases a remarkable level of diversity. Fig. 1 overviews attribute categories composed in PEARL. Second, we provide a strong evaluation baseline for attribute learning by formulating this task as vision-language fusion task. Specifically, we utilize CLIP [22] (Constrastive Language-Image Pretraining) due to its capability to obtain quality feature embedding on image-text modalities.

The recent PAR approaches either attempt to learn discriminative visual features using CNNs [9, 13, 17, 33] or aims to incorporate Transformers to exploit visual-textual relationship between attribute and their appearances [2, 6, 31, 38]. However, their performance is restricted by the following reasons: CNN-based approaches fail to capture semantic relationship between attributes and Transformer-based approaches unable to efficiently handle data imbalance problem. Furthermore, the appearance of attributes is influenced by a number of contextual factors put an extra burden on the recognition capabilities of the model. Therefore, in the PAR setting, it is necessary to emphasize on learning attentive features that both exploit visual-textual relationship and handle data imbalance problem inherited by the existing datasets. To achieve this, we suggest an evaluation strategy incorporating popular CLIP [22] model with three different prompt settings. Specifically, our goal is to *"explore how effective textual descriptions are for pedestrian images and understand the relationship between the appearances of attributes through text."*

Putting all this together, the paper makes the following contributions: i) We explicitly address the inherent data imbalance problem in PAR and investigate the role of textual prompts with CLIP-based model to accurately recognize diverse attributes. ii) We also release PEARL, a novel and diverse pedestrian attribute dataset comprised of 30k samples, each annotated with 146 attributes. iii) We carried out extensive experiments to validate both usability of the PEARL dataset and suggested baseline.

2. Related Work

Pedestrian Attribute Recognition (PAR) has gained popularity in recent years for its applications in person search, person re-identification, and action recognition. A handful of notable efforts can be safely grouped into feature-centric and classifier-centric learning. In this section, we summarize them briefly.

2.1. Feature-centric Learning

Feature-centric PAR methods [14, 17, 19, 40] have been extensively explored in PAR. Cheng *et al*. [2] have proposed a visual-textual baseline for PAR by formulating it as a multi-modal problem to utilise the latent textual information in the attribute annotations. Li *et al*. [17] have utilised a continual learning strategy for multiple groups of pedestrian attributes and also incorporate a self-learning-based method to handle incomplete labels via catastrophic forgetting. Jia *et al*. [11] have exposed limitations of the one-shared-feature-for-multiple-attributes mechanism and used a disentangled attribute feature learning framework.

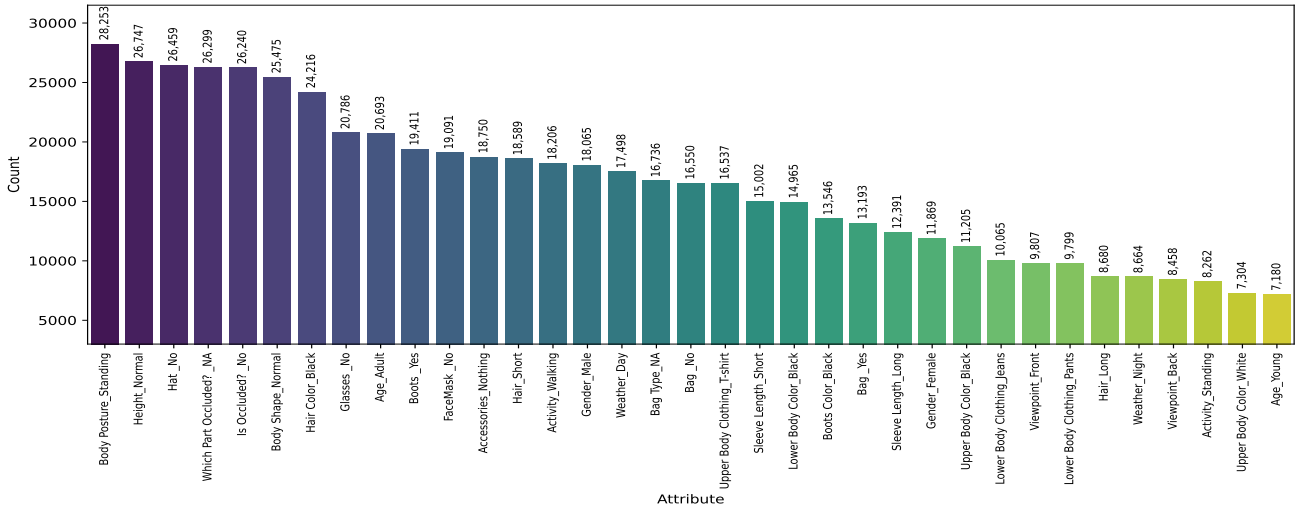

Figure 2. **Long-tail Distribution:** A few selected attribute distributions in the PEARL dataset. Distributions are sorted in decreasing order. More details about these distributions and other datasets can be found in **appendix**.

2.2. Classifier-centric Learning

Classifier-centric learning approaches often focus on inherent data imbalance problem in PAR. These can be divided into two subcategories: (i) data-centric approaches, and (ii) loss-centric approaches. In data-centric approaches, the minority class can be expanded through oversampling or majority class can be scale down for stable training [12]. Thakare *et al*. [32] incorporated multiple views of a person to bridge the recognition gap. In loss-centric approaches, the classifier is trained with novel loss considering various scenarios such as viewpoint, positive ratio of attributes, etc. One of the notable early works in PAR [13] has introduced the weighted binary cross-entropy loss function to effectively handle data imbalance. Jia *et al*. [11] have also proposed a triplet loss to assist group attention merging module to learn discriminative features. Yan *et al*. [37] have introduced dropping rate during training and enforce delay in hard samples training thus favouring easy samples. While feature-centric and classifier-centric learning methodologies have showcased promising advancements in PAR, their efficacy drops in challenging scenarios. Feature-centric approaches, relying solely on visual information or text fusion, may struggle with limited contextual understanding, hindering their performance in highly dynamic or diverse environments. Similarly, classifier-centric strategies addressing data imbalance or loss functions often lack adaptability to complex scenarios, such as dealing with occlusions, varying illuminations, or attribute dependencies, leading to reduced robustness, and generalization in real-world applications of PAR.

Table 1. **Dataset Diversity:** Quantitative comparison of PEARL with large-scale PAR datasets. The categories column indicates number of attribute categories available in datasets such as *hair style*, *bag types*, *cloth patterns* and attribute column is corresponding sub-categories such as *hair-style-long, bag-type-handbag*, etc.

Dataset	Scene	#Samples	#Categories	#Attribute	#Weather	#Countries
PETA	Mixture	19,000	12	61	1	1
PA100K	Outdoor	100K	11	26	1	1
RAP	Indoor	41,585	15	69	1	1
PEARL	**Mixture**	**30,000**	**25**	**146**	**4**	**12**

3. Why PEARL Dataset?

Recognizing pedestrian attributes faces two fundamental challenges: (i) Unavoidable parameters such as weather condition, occlusions, low illumination, camera angles, and blur, which influence attribute appearances, like labeling a *hat* as *black hair* due to low illumination; (ii) hindrances in attribute appearances due to varying viewpoints, which makes it challenging to predict attributes like *glasses* or *face masks* from a rear perspective. A comprehensive PAR dataset must tackle these issues, which are not fully addressed in the existing datasets. Thus, we propose the PEARL dataset to mitigate these challenges.

3.1. Interesting Facts About PEARL

The PEARL dataset comprises with 30K pedestrian images, each annotated with 25 attribute categories, spanning over 146 sub-attributes. We have extracted images from outdoor surveillance videos[1] that reflect practical applications and challenges. We comprehensively cover nearly all

[1]From publicly available non-copyrighted sources.

critical attributes relevant to security surveillance applications, comprising aspects such as *body posture*, *accessories*, *bag types*, *clothing styles*, *colors*, and *activities*. Fig. 3 depicts 1001 image samples taken from PEARL showing diversity in colors, body posture, weather, etc. Existing datasets are often focused on a particular context. For instance, RAP [15] was recorded in a shopping mall, and Market1501 [18] was captured inside a marketplace. Table 1 shows the quantitative comparison of PEARL against large-scale PAR datasets. To diversify, we have extracted images from twelve countries that covers *seven* distinct public locations including *streets*, *parks*, *airports*, *stations*, *college campuses*, *beaches*, and *marketplaces*. Additionally, we have incorporated four distinct weather conditions: *sunny*, *night-time*, *rainy*, and *snow*. Fig. 2 shows several attributes and their appearance counts.

3.2. Statistics of PEARL

PEARL encapsulates a wide spectrum of variations to effectively mirror practical and challenging scenarios. To achieve this, we have extracted raw pedestrians images from 25 hours of surveillance videos using Faster-RCNN [23] and filtered out extremely blurry samples, resulting in 30K images. The images exhibit diverse resolutions, ranging from 20×78 as the lowest resolution, 678×210 be the highest resolution with an average resolution of 107×274. The location-wise sample numbers within PEARL are as follows: Airport (4471), Market (5776), Beach (1706), Street (11410), Park (2408), Station (3902) and Campus (328). The weather-wise and country-wise distributions can be found in been added to the **appendix .**

Figure 3. **PEARL Mosaic**: A mosaic portrait representing sample images with different cloth patterns, colours, body postures, and genders from the PEARL dataset. *Best viewed in over 2X zoom and colour.*

3.3. PEARL vs Large-Scales

RAPv1-v2 [15] and PA100K [20] are the largest datasets available publicly. However, despite having extensive collections, they often lack diversity resulting in minimal intra-class variations. For instance, the RAP dataset solely consists of recordings from *shopping malls*. The PA100k dataset contains annotations of pedestrian images with only a limited number of attributes (26). In contrast, the proposed PEARL dataset encompasses a wide array of locations and offers rich annotations, making it a good resource for advanced research as well as suitable for practical applications.

4. Evaluation Baseline

PAR models expect to identify and characterize pedestrian attributes across a wide spectrum of vastly differing conditions and environments. However, conventional image-based methodologies often overlook the intricate relationship between attributes and their appearances. To mitigate this problem, we propose a strong evaluation baseline for recognizing attributes by formulating PAR problem in language-image fusion task. The high-level architecture of the baseline model is depicted in Fig. 4.

4.1. Problem Formulation

Following the prior works [2, 13, 37], we formulate the PAR problem as a language-image fusion problem, where the model expects to learn discriminating features that represent the presence or absence of attributes guided by textual description of an image.

Assume the appearance-wise attribute set of a pedestrian image is denoted by $\Pi = \{\pi_1, \pi_2, \ldots, \pi_K\}$, where K is the number of attributes. Let image-attribute pair $\{(\mathcal{I}_1, \mathcal{Y}_1), (\mathcal{I}_2, \mathcal{Y}_2), \ldots (\mathcal{I}_N, \mathcal{Y}_N)\}$ be N image samples in the training set, where $\mathcal{I}_i$ is the i-th pedestrian image and $\mathcal{Y}_i \in \Pi$. More precisely, $\mathcal{Y}$ represents a human-annotated binary vector, wherein 0 and 1 denote the absence and presence of an attribute in image $\mathcal{I}$. In this context, our objective is to train a PAR model denoted as $\mathcal{H}(.)$, which computes the probability (p_i) for each attribute π_i within the set Π, expressed as $\mathcal{H}(\mathcal{I}, \Pi) = [p_1, p_2, \ldots p_M]$.

4.2. Generating Descriptive Prompts

In language-image tasks, prompt is a textual description that guides the model to associate and understand visual information. Utilizing prompts in PAR involves converting attribute sets into meaningful sentences, providing crucial context for more effective recognition as compared to relying solely on image features. However, recognizing attributes using captions can be limited by two critical factors: Firstly, as the attribute names are only added to the captions, they could lead to model confusion because due to

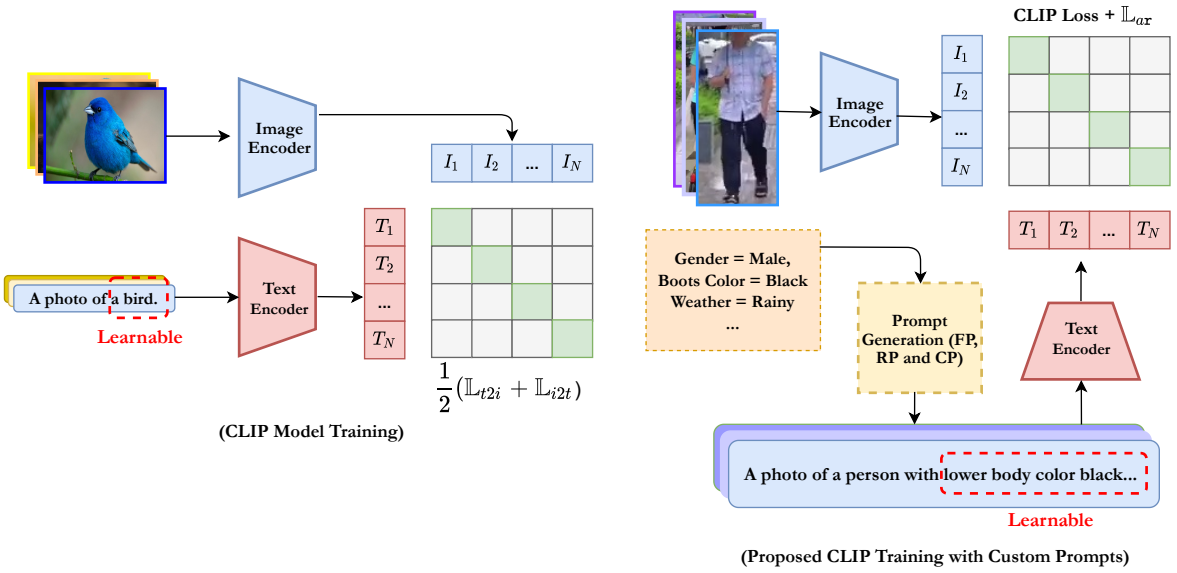

Figure 4. **Proposed CLIP-based Evaluation Baseline**: Left: An illustration of CLIP [22] with default prompts. The framework consists of visual-textual encoders and trained with both $(\mathbb{L}_{t2i})$ and $(\mathbb{L}_{i2t})$ losses. In the training stage, both classifiers and encoders train jointly by minimizing similarity loss by using joint image-text pairs. Right: The proposed CLIP-based evaluation baseline for PAR, where different prompts have been generated using three suggested settings. Moreover, both encoders have been jointly trained via constrastive and suggested inverse-frequency loss $(\mathbb{L}_{ar})$.

lack of context. Secondly, due to data imbalance, generated captions may introduce some bias for over-fitting on some of the attributes. To mitigate this challenge, it is essential to provide adequate context and put more emphasis on certain low-frequency attributes. In this paper, we introduce three prompt settings for CLIP training to enhance the attribute recognition capability. Additionally, we recommend using inverse frequency loss to address data imbalance.

Full Prompt: By design, the input format for CLIP is: *"The photo of a {CLASS}"*, where the class is the label of the object in the input image. This text format is not possible by default with the PAR setting due to multi-class classification aspect. Moreover, the caption length is set to a maximum of 77 tokens [22], therefore limiting the size of pedestrian descriptions that are more extensive. To solve this, we add a prompt setting limited to *"{Attribute} {Value}"* so that each caption in the PEARL dataset has 50 total tokens (two tokens per caption). As this prompt configuration accommodates all 25 main attributes along with their corresponding annotated subcategories, we refer to it as the **Full Prompt (FP)**. The final full-prompt textual description for each training image follows the format: *"A photo of a person with {attribute1} {value1} {attribute2} {value2} and so on"*.

Random Prompt: The CLIP model can be trained with the full prompt setting, which includes all attributes in the caption. Thus, we explore the **Random Prompt** (RP) setting, where a certain percentage of attributes are excluded *randomly*, leading to shorter captions. Training CLIP with random prompt settings serves two benefits: Firstly, the model would be trained to discriminate with increased precision attributes included due to the focused context. Secondly, this keeps the effect of over-fitting minimal, as the model is exposed to different attribute combinations. The format of the prompt is consistent with the full prompt: *"A photo of a person with {attribute1} {value1}, {attribute2} {value2}"*, here number of attributes are control by the exclusion probability specified by (ρ).

Contextual Prompt: We obtained encouraging results for both the full prompt and random prompt settings. However, these methods fail to adapt to the imbalanced nature of PAR datasets. Our observations are two-fold: (i) the full prompt considers all attributes, leading to over-fitting on dominant attributes; and (ii) the random prompts often exclude some under-represented classes because of the randomness. For this purpose, we could add more context or give more importance to the minority classes. To solve this, we can add more context or emphasize on minority classes. Inspired by CLIP's success with specialized prompts such as: *"A photo of a {label}, a type of pet"* on the Oxford-IIIT Pets [21] and *"a satellite photo of a {label}"* on the OCR dataset, we suggest a new prompt setting called the **Contextual Prompt** (CP). This combines random prompts with added emphasize on minority classes. We first apply the (ρ)

exclusion to sample attributes and estimate the frequencies of all sub-categories for the attributes in the training batch, segregating dominant attributes from under-represented ones by positive-negative z-score. After segregation, we prepend token *"especially"* on each attribute with negative z-score. For instance, *"A photo of a person with gender male, especially face-mask present."* The z-score varying per the batch ensures that low-frequency sub-categories get more emphasis during the prompt construction.

Inverse Frequency Loss ($\mathbb{L}_{ar}$)**:** In addition to the different prompt settings, we have also injected **Inverse Frequency Loss** (IFL) to the constrastive loss utilized by CLIP. This basically involves computation of weights through counting of class frequencies such that minority classes are emphasized. These computed weights get integrated with a CLIP constrastive loss during training.

4.3. Training and Testing CLIP

We have followed the standard training procedure to train the CLIP model. In the PAR settings, both the visual and textual encoders have been optimized during training as they could learn the mapping of textual description of each attribute to the visual features of the pedestrian images given as input. In the inference step, we have used cosine similarity between both encoders to decide for attributes. For evaluation, we have tested with the visual encoders, e.g. CLIP ViT-B/32 [5] and ResNet50 [7].

5. Experiments

In this section, we present implementation details, datasets, evaluation metrics, proposed baseline comparisons with prior arts, benchmark PEARL, and ablation experiments.

5.1. Datasets and Evaluation Metrics

The **PA100K** [4] dataset features 100K pedestrian images captured across 598 outdoor scenes, each annotated with 26 commonly used attributes. The dataset is split into training, validation, and test sets, maintaining an 8:1:1 ratio for training. In contrast, the **PETA** [4] includes over 8.7K pedestrians captured within 19K images with varied resolutions from 17 $\times$ 39 to 169 $\times$ 365. Each pedestrian is annotated with 61 binary attributes and four multi-class attributes. However, for the present analysis, only 35 attributes with a positive label ratio exceeding 5% are considered following the established protocol. Based on the main study [4], the dataset undergoes a random division into three splits, allocating 9.5K images for training, 1.9K images for validation, and the remaining 7.6K images for testing.

The **RAP** [15] is a collection of over 41K pedestrian images. Adhering to the original protocol by Li [15], we selectively consider 51 attributes for evaluation purposes. For model evaluation, five random splits are employed, with over 33K images utilized for training and over 8K images for testing in each split. The final evaluation entails averaging the performance across all splits.

Following the prior works [11,13,16,20,32,36], we have employed instance-based **mean accuracy** (mA), along with four label-based criteria: **accuracy**, **precision**, **recall**, and **F1 score** to benchmark PEARL using proposed baseline. For experimental purpose, we employ official implementation of CLIP [22] with default parameters. Implementation details can be found in **supplementary document**.

5.2. Benchmarking PEARL

We present a comprehensive benchmarking of the dataset through evaluations against recent PAR works and the proposed baseline method. Tab. 2 shows the performance of a few important SOTA methods.

Table 2. **PEARL Benchmark Results:** Performance comparisons on PEARL. Red font highlights the highest scores, and blue font denotes the second-highest scores. FP, RP and CP being the Full, Random and Contextual prompt settings with IFL.

Method	Backbone	mA	Acc.	Prec.	Rec.	F1
CNN + SVM [4]	VGG-16	71.14	53.38	70.22	72.30	71.24
DeepSAR [13]	RN50	73.25	-	-	-	-
DeepMAR [13]	RN50	80.31	76.40	79.52	76.34	77.89
HP-Net [20]	Incep.	81.47	77.12	80.55	78.11	79.31
VTB [2]	ViT-B	83.03	73.18	82.73	81.92	82.32
DAFL [11]	ViT-B	82.11	80.68	81.62	82.34	81.97
Baseline [12]	-	81.57	77.35	80.72	79.05	79.87
SSPNet [26]	Swin-S	80.71	79.56	81.90	78.12	79.96
KD-PAR [36]	Res2Net	80.54	78.60	80.78	79.20	79.98
S-ACRM [32]	RN50	81.23	79.51	82.28	80.37	81.31
PARFormer-B [6]	ViT-B	81.55	80.19	83.60	81.22	82.39
CLIP + FP	**ViT-B**	**83.35**	**82.58**	**85.30**	**82.76**	**84.01**
CLIP + RP	**ViT-B**	**81.05**	**81.45**	**84.11**	**82.17**	**83.12**
CLIP + CP	**ViT-B**	**87.29**	**83.35**	**86.43**	**84.45**	**85.42**

In the benchmarking results using PEARL dataset, the DeepMAR [13] with ResNet50 [7] method achieves an mA of 80.31, demonstrating commendable performance. However, the suggested evaluation baseline outperforms various SOTA approaches, achieving an mA of 87.29 with contextual prompt setting. Despite promising performance, results on the PEARL dataset indicate that the dataset complexity still remains as a critical issue. It has been observed that the diverse nature of attributes and environmental conditions continue to pose challenges to PAR methods.

5.3. SOTA Comparisons

We have compared the proposed baseline with recent SOTA approaches [2, 6, 11, 13, 14, 17, 19, 20, 24, 28–30, 33]. Tab. 3 summarises the performance comparisons on RAP [15] and PETA [4] datasets. It can be noted that the suggested baseline achieves competitive performance

Table 3. **Prior Arts Analysis:** Performance comparisons on RAP [15] and PETA [4] datasets.

		RAP [15]					PETA [4]				
Method	**Backbone**	**mA**	**Acc.**	**Prec.**	**Rec.**	**F1**	**mA**	**Acc.**	**Prec.**	**Rec.**	**F1**
CNN + SVM [4]	VGG16	72.28	31.72	35.75	71.78	47.73	76.65	45.41	51.33	75.14	61.00
DeepMAR [13]	CaffeNet	73.79	62.02	74.92	76.21	75.56	82.89	75.07	83.68	83.14	83.41
HP-Net [20]	Inception	76.12	65.39	77.33	78.79	78.05	81.77	76.13	84.92	83.24	84.07
VeSPA [24]	Inception	77.70	67.35	79.51	79.67	79.59	83.45	77.73	86.18	84.81	85.49
JRL [33]	AlexNet	77.81	–	78.11	78.98	78.58	85.67	–	86.03	85.34	85.42
PgDM [14]	CaffeNet	74.31	64.57	78.86	75.90	77.35	82.97	78.08	86.86	84.68	85.76
JLPLS-PAA [29]	-	81.25	67.91	78.56	81.45	79.98	84.88	79.46	87.42	86.33	86.87
RA [40]	Inception-V3	81.16	–	79.45	79.23	79.34	86.11	–	84.69	88.51	86.56
ALM [30]	BN-Inception	81.87	68.17	74.71	86.48	80.16	86.30	79.52	85.65	88.09	86.85
JLAC [28]	ResNet50	83.69	69.15	79.31	82.40	80.82	86.96	80.38	87.81	87.09	87.45
DAFL [11]	Inception	83.72	68.18	77.41	83.39	80.29	87.07	78.88	85.78	87.03	86.40
SSC [10]	ResNet50	82.77	68.37	75.05	87.49	80.43	86.52	78.95	86.02	87.12	86.99
SSPNet [10]	Swin-S	83.24	70.21	80.14	82.90	81.50	88.80	82.80	88.48	90.55	89.50
KD-PAR [36]	Res2Net	81.30	69.22	74.61	84.20	79.11	85.50	78.31	87.97	84.17	86.03
PARFormer-B [6]	ViT-B	**83.84**	69.70	79.24	87.81	81.16	88.65	82.34	86.89	**91.55**	88.66
CLIP + FP	**ViT-B**	**83.71**	**71.28**	**83.49**	**87.92**	**85.64**	**88.71**	**84.02**	**88.58**	**88.79**	**88.68**
CLIP + RP	**ViT-B**	**82.42**	**69.60**	**82.11**	**86.81**	**84.39**	**86.91**	**83.45**	**87.55**	**88.03**	**87.78**
CLIP + CP	**ViT-B**	**86.70**	**72.81**	**84.03**	**88.55**	**86.23**	**90.05**	**87.36**	**91.15**	**92.60**	**91.86**

through leveraging visual-textual analysis on both datasets. The experiments also reveal that ViT-based PARFormer [6] reports competitive recall values due to integration of attribute and viewpoint information. However, viewpoint may not always be a decisive feature, and completely relying on it may generate more false positives. This is evident when other metrics are used for comparisons. Fig. 5 depicts a handful of prediction results obtained through the suggested baseline.

Other notable methods such as JLAC [28], DAFL [11] show stable results on both datasets due their integration of GCN and triplet loss. Similar observations are reported on performance comparisons using the PA100K dataset. Tab. 4 shows the performance of several PAR approaches. The proposed baseline achieves SOTA performance on PA100K dataset with precision, recall, and F1 values as high as 92.67%, 92.88%, and 92.77%, respectively. It can also be observed that the overall label-based accuracy is low on PA100K [20]. It is probably due to less diverse annotations on PA100K dataset. This has led to smaller inter-class variations. More results on zero-shot setting can be found in **supplementary document**.

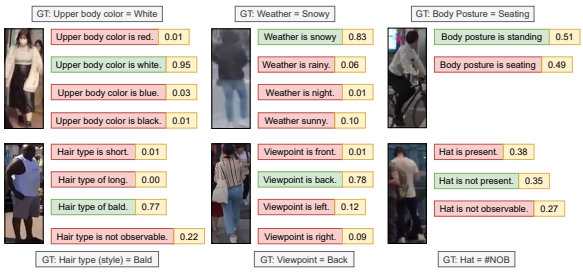

Figure 5. **Caption Score:** Few prediction results predicted by the Full Prompt (FP) CLIP-ViT-B/32 model trained on PEARL dataset. GT: Ground truth assigned by annotators. Green bars are prediction by the model with highest caption score.

Table 4. **Prior Method Analysis:** Performance comparisons on PA100K [20].

Method	**mA**	**Acc.**	**Prec.**	**Rec.**	**F1**
DeepMAR [13]	72.70	70.39	82.24	80.42	81.32
HP-Net [20]	74.21	72.19	82.97	82.09	82.53
JLPLS-PAA [29]	81.61	78.89	86.83	87.73	87.27
ALM [30]	80.65	77.08	84.21	88.84	86.46
JLAC [28]	82.31	79.47	87.45	87.77	87.61
Baseline [12]	81.61	79.45	87.66	87.59	87.62
DAFL [11]	83.54	80.13	87.01	89.19	88.09
SSPNet [11]	83.58	80.63	87.79	89.32	88.55
KD-PAR [36]	81.56	78.45	87.90	86.05	86.96
S-ACRM [32]	82.26	77.19	86.36	87.92	87.32
PARFormer-B [6]	**83.95**	80.26	87.51	**91.07**	87.69
CLIP + FP	**83.76**	**82.33**	**89.50**	**91.01**	**90.24**
CLIP + RP	**82.24**	**81.78**	**88.72**	**90.33**	**89.51**
CLIP + CP	**86.45**	**86.12**	**92.67**	**92.88**	**92.77**

5.4. Cross-Dataset Validation

Cross-data validation experiments are vital to guarantee generalization and model robustness between datasets. In order to validate the usability of PEARL dataset, we have conducted these cross-validation experiments on suggested CLIP with Contextual Prompt and latest transformer-based method PARFormer-B [6]. From the Tab. 5, it can be observed that both methods experience similar gain when tested on PETA and PA100K after training using PEARL. Precisely, the CLIP + CP method, when trained on PEARL, has the mA 93.45% and overall accuracy of 89.5% on PETA with a gain of +2.65%, while it attains the mA of 90.11% and overall accuracy of 89.57% on PA-100K, with a gain of +3.55%. This basically shows that the training on PEARL is able to infuse generalized learning over other datasets. The positive gain and notably higher accuracy of PEARL

as compared to training on PETA or PA100K dataset are encouraging. These datasets show slightly negative gains. The same can be observed through Fig. 7. This indicates that PEARL offers a more comprehensive and balanced training set that can be used for producing better results and offer more generalization as compared to other testing datasets.

Table 5. **Dataset Cross-Validation:** Performance comparison of PARFormer-B [6] and the proposed CLIP + CP method across PEARL, PETA, and PA-100K datasets. (△) represents the difference of average values of mA and Accuracy obtained during training and testing on the same dataset.

Training Dataset	Method	Testing on PETA			Testing on PA-100K		
		mA	Acc.	△	mA	Acc.	△
PEARL	PARFormer-B	92.35	87.17	**+4.26**	88.20	85.51	**+4.75**
	CLIP + CP	93.45	89.5	**+2.65**	90.11	89.57	**+3.55**
		Testing on PEARL			Testing on PA-100K		
		mA	Acc.	△	mA	Acc.	△
PETA	PARFormer-B	73.37	71.29	**-8.89**	79.23	77.51	**-3.73**
	CLIP + CP	83.03	76.59	**-5.51**	83.72	82.88	**-2.98**
		Testing on PEARL			Testing on PETA		
		mA	Acc.	△	mA	Acc.	△
PA-100K	PARFormer-B	71.32	69.73	**-10.34**	83.30	78.86	**-4.41**
	CLIP + CP	82.30	79.51	**-4.45**	87.23	84.12	**-3.03**

5.5. Ablation Study

The primary components of the proposed baseline include visual encoders for visual feature extraction and three prompt settings. To comprehend individual impact on the overall performance, we have conducted an ablation study. We have shown the performance difference against both ViT-B/32 and RN50 as visual encoders under three prompt settings in Fig. 6. It can be observed from the bar chart of the image that the CLIP with ViT-B/32 as visual encoder performs better than RN50 under all settings. It is due to the inclusion of self-attention and patch embedding in ViT. This helps the model to learn diverse appearances. On the other hand, the graph depicts the variation in mean average with different values of (ρ) used for random prompts. Note that, for Full Prompt (FP) setting, $(\rho = 0)$ suggests that no attribute is excluded during the prompt generation. The results mentioned in the paper for CLIP + RP and CP in Tables 2-4 for random prompt are obtained by setting $\rho = 0.2$.

6. Conclusion and Future Work

In this work, we introduce PEARL, a comprehensive PAR dataset comprising of 30k images, each annotated with 146 critical sub-attributes. Extracted from surveillance videos across twelve countries, PEARL dataset exhibits remarkable diversity, spanning various public places and illuminations. To tackle the challenges posed by existing PAR datasets, we establish a robust baseline using CLIP (Constrastive Language-Image Pretrainig) model with three different prompt settings, i.e., full prompt, random prompt, and contextual prompt. The experiments reveal that both

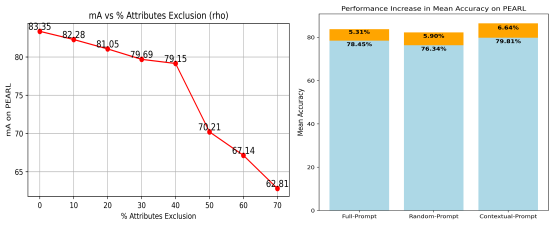

Figure 6. **Ablation and Effect of** (ρ)**:** Left: The variation in mean average for different values of (ρ) employed during Random Prompt setting. Right: Bars showing a significant gain with two visual encoders, i.e., ViT-B/32 (Orange) and ResNet50 (sky-blue).

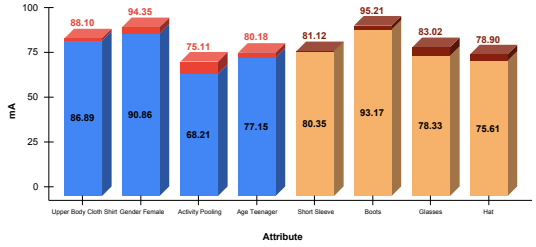

Figure 7. **Attribute-wise Testing:** Bars showing a significant gain in mA when the VTB [2] model has been explicitly trained on PEARL specific attributes and tested on the attributes from PA100K [20] (Blue) and RAP [15] (yellow) datasets.

visual and textual modalities are helpful in recognizing attribute in adverse conditions. Moreover, it can also be concluded that a diverse PAR dataset can help models learn more discriminating features. The proposed PEARL dataset is sufficiently large and diverse to achieve SOTA performance. While the CLIP-based baseline demonstrates promising attribute recognition capabilities, its performance may vary depending on the complexity of environmental factors and the diversity of attributes in real-world scenarios. Future directions encompass exploring diversity offered by the PEARL to learn more compact visual-textual relationships and advance feature extraction process to explore better embedding are avenues for future investigation.

Acknowledgement

This research was supported by Development of Proactive Crowd Density Management Platform Based on Spatiotemporal Multimodal Data Analysis and High-Precision Digital Twin Simulation Program through the Korea Institute of Police Technology (KIPoT) funded by the Korean National Police Agency & Ministry of the Interior and Safety (RS-2024-00405100), and Korea Institute of Science and Technology (KIST) Institutional Program (Project No.2E33001). Data collection and initial work were done at IIT Bhubaneswar under the Project code CP438.

References

[1] Surbhi Aggarwal, R. Venkatesh Babu, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *Wint. Conf. on Appli. of Comp. Vis.*, 2020. 1, 2

[2] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Trans. Circuit Syst. Video Technol.*, 2022. 2, 4, 6, 8

[3] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1

[4] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM Int. Conf. Multimedia*, 2014. 1, 2, 6, 7

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6

[6] Xinwen Fan, Yukang Zhang, Yang Lu, and Hanzi Wang. Parformer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Trans. Circuit Syst. Video Technol.*, 2023. 2, 6, 7, 8

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[8] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *Int. Conf. Comput. Vis.*, pages 11895–11904, 2021. 1

[9] Yan Huang, Qiang Wu, Jingsong Xu, Yi Zhong, and Zhaoxiang Zhang. Unsupervised domain adaptation with background shift mitigating for person re-identification. *Int. J. Comput. Vis.*, 129, 2021. 1, 2

[10] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Int. Conf. Comput. Vis.*, 2021. 7

[11] Jian Jia, Naiyu Gao, Fei He, Xiaotang Chen, and Kaiqi Huang. Learning disentangled attribute representations for robust pedestrian attribute recognition. In *Proceedings of the AAAI Conf. on Arti. Intel.*, 2022. 1, 2, 3, 6, 7

[12] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021. 3, 6, 7

[13] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *IAPR Asian Conf. on Pat. Recog.*, 2015. 2, 3, 4, 6, 7

[14] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *IEEE Int. Conf. on Multi. and Expo*, 2018. 2, 6, 7

[15] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Trans. Image Process.*, 28, 2018. 1, 2, 4, 6, 7, 8

[16] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation. In *Proc. of Int. Joint Conf. on Arti. Intel*, 2019. 2, 6

[17] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Visual-semantic graph reasoning for pedestrian attribute recognition. In *Proceedings of the AAAI Conf. on Arti. Intel.*, 2019. 2, 6

[18] 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*, 2019. 1, 2, 4

[19] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. In *Brit. Mach. Vis. Conf.*, 2018. 2, 6

[20] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Int. Conf. Comput. Vis.*, 2017. 1, 2, 4, 6, 7, 8

[21] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 5

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. on Machine Lear.*, 2021. 2, 5, 6

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 4

[24] M Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017. 6, 7

[25] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *ACM Int. Conf. Multimedia*, pages 5566–5574, 2022. 1

[26] Jifeng Shen, Teng Guo, Xin Zuo, Heng Fan, and Wankou Yang. Sspnet: Scale and spatial priors guided generalizable and interpretable pedestrian attribute recognition. *Pattern Recognition*, 148, 2024. 6

[27] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. In *Wint. Conf. on Appli. of Comp. Vis.*, 2023. 1, 2

[28] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the AAAI Conf. on Arti. Intel.*, 2020. 2, 6, 7

[29] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z Li. Attention-based pedestrian attribute analysis. *IEEE Trans. Image Process.*, 2019. 6, 7

[30] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Int. Conf. Comput. Vis.*, 2019. 6, 7

[31] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 2

[32] Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. Let's observe them over time: An improved pedestrian attribute recognition approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 708–717, 2024. 3, 6, 7

[33] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *Int. Conf. Comput. Vis.*, 2017. 2, 6, 7

[34] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[35] Guile Wu, Xiatian Zhu, and Shaogang Gong. Learning hybrid ranking representation for person re-identification. *Pattern Recognition*, 121, 2022. 1

[36] Peishu Wu, Zidong Wang, Han Li, and Nianyin Zeng. Kd-par: A knowledge distillation-based pedestrian attribute recognition model with multi-label mixed feature learning network. *Expert Systems with Applications*, 2023. 2, 6, 7

[37] Yan Yan, Youze Xu, Jing-Hao Xue, Yang Lu, Hanzi Wang, and Wentao Zhu. Drop loss for person attribute recognition with imbalanced noisy-labeled samples. *IEEE Transactions on Cybernetics*, 2023. 1, 2, 3, 4

[38] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *ACM Int. Conf. Multimedia*, 2023. 1, 2

[39] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *ACM Int. Conf. Multimedia*, 2023. 2

[40] Xin Zhao, Liufang Sang, Guiguang Ding, Jungong Han, Na Di, and Chenggang Yan. Recurrent attention model for pedestrian attribute recognition. In *Proceedings of the AAAI Conf. on Arti. Intel.*, 2019. 2, 7