

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

(Additional Document)

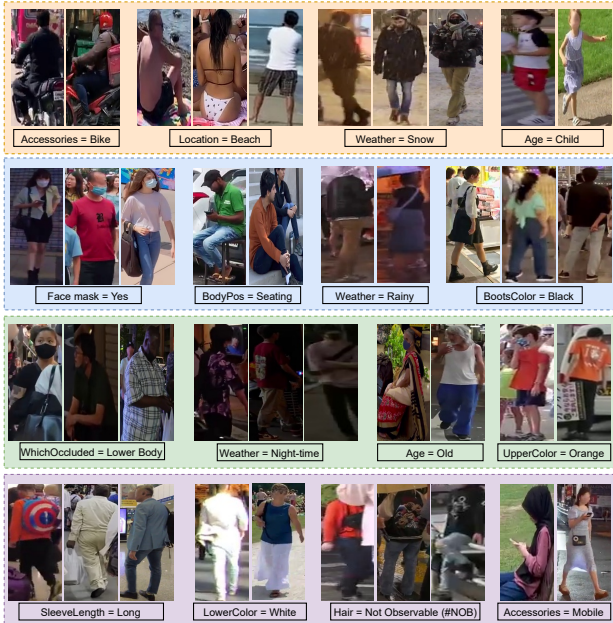


Figure 1. **Attribute-wise Data sample:** A handful of dataset samples from PEARL. Zoom over for better visibility.

1. Introduction

This supplementary document includes implementation details, and a few additional qualitative results. Sec. 2 provides additional information about PEARL and present few category-wise attribute distribution. In Sec. 3, we have highlighted implementation details of the suggested baseline. Sec. 4 explains inverse frequency loss adopted in the training CLIP. Lastly, in Sec. 5, we present a few results under zero-shot setting.

2. PEARL Statistical Details

To enhance the diversity of the PEARL, we sourced 30,000 pedestrian images from surveillance videos spanning across 12 different countries. The countries from which these images were sourced include *Japan, China, USA, Spain, Pakistan, UAE, Malaysia, UK, Germany,*

France, South Korea, and India. These images were captured in seven common public places, namely *marketplace, bus-railway stations, beach, campus, park, airport, and street.* Each image from PEARL is annotated with 25 attribute categories spanning over 146 sub-attributes. Fig. 2a and 2b summarise weather-wise and country-wise attribute distributions in the PEARL dataset, respectively and Fig. (1) depicts handful of attribute-wise PEARL data samples.

3. Implementation Details

For experimental purpose, we employ official implementation of CLIP with default parameters. This default parameter setting is shown in Table 2. Note that, the default *vocabulary size* of the CLIP is set to 77 and in our experiment three different prompts setting generates tokens less than 77.

Annotation Guidelines: Table (1) provides guidelines followed for the PEARL dataset, detailing specific criteria for various pedestrian attributes. It includes categories such as Gender, Age, Body Shape, Weather, Accessories, and Occlusion status, with specific attributes and their annotation guidelines. For gender and age, annotations use visible cues and common knowledge, while “Not Observable” is used in cases of heavy occlusion. Age annotations for children and teenagers rely on context such as school bags and relative size. Body shape is classified as Skinny, Normal, or Fat based on visible body size indicators. Weather conditions and accessories are determined from the video source and visible possession, respectively, and occlusion is noted if it prevents attribute determination.

4. Inverse Frequency Loss

It is clear that the problem of data imbalance is inherent in pedestrian attribute datasets. To tackle this challenge, we collected pedestrian images from wide-ranging countries and places. Further, we used an inverse-frequency loss to alleviate the data imbalance. Inverse-frequency loss upweights rarer-class samples by attaching a greater weight to them, enabling the model to pay more attention

Table 1. **Annotation Guidelines:** Annotation guidelines that we followed to annotate the PEARL dataset. Note that rest of the attributes have been annotated using day-to-day knowledge and visible cues.

Categories	Attributes Options	Guidelines
Gender	Male, Female, Transgender	Use common knowledge and visible cues consider clothing, hairstyle, and any other visible indicators.
Gender, Age	Not Observable (#NOB)	Heavy occlusion (e.g., the pedestrian is partially or fully blocked by an object or another person)
Age	Child	Presence of a school bag or child holding a parent’s hand or size and height relative to surrounding pedestrians.
Age	Teenager	Size and height relative to adults and children, clothing styles typical of teenagers (e.g., trendy or casual wear).
Body Shape	Skinny (Thin)	Clearly underweight appearance, visible bone structure, or slim build.
Body Shape	Normal (Average)	Average body size, neither visibly underweight nor overweight.
Body Shape	Fat	Overweight appearance, larger body size, or visible bulk.
Weather	Sunny, Rainy, Night-time, Snow	Determined from the video source.
Accessories	Mobile, Bike, Trolley bag, Umbrella	Items in possession (e.g., cellphone in hand, seating on bike).
IsOccluded?	Yes, No	Yes, if unable to determine the presence of one or more other attributes such as hair style due to occlusion.

Table 2. **Implementation Details:** CLIP-ResNet and CLIP-ViT-B/32 hyperparameters employed for the implementation of the baseline.

Encoder →	RN50	ViT-B/32
Hyperparameter	Value	Value
Batch size	128	128
Vocabulary size	≤ 77	≤ 77
Training epochs	300	300
Weight decay	0.2	0.2
Adam β_1	0.9	0.9
Adam β_2	0.999	0.98
Adam ϵ	10^{-8}	10^{-6}
Learning Rate	5×10^{-4}	5×10^{-4}
Input Resolution	224^2	224^2

to underrepresented attributes. Algorithm (1) describes the computation of weights that can be integrated during the training process of CLIP. These weights are normalized to avoid domination of the loss function by the high-frequency classes, making it such that all attributes are learned in balance. Below is the python-like pseudo code of CLIP training implemented in the paper:

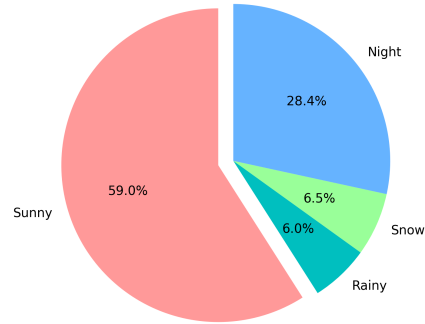
Table 3. Performance comparison of proposed CLIP with and without the addition of class-balanced loss.

Method	mA w/o L_{ar}	mA with L_{ar}
CLIP + FP	86.54	83.35
CLIP + RP	79.11	81.05
CLIP + CP	84.40	87.29

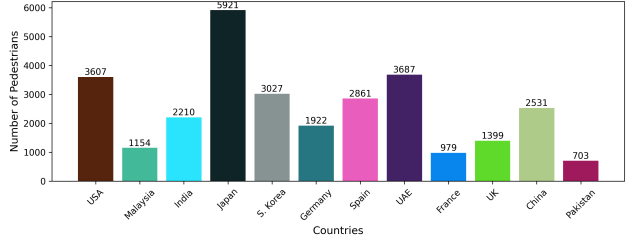
```

# Get image and text embeddings
img_embed = clip_model.encode_image(images)
txt_embed = clip_model.encode_image(texts)
# Predicts attributes
preds = img_embed * txt_embed.T
# Compute contrastive loss
Li2t = con_loss(img_embed, txt_embed)
Lt2i = con_loss(txt_embed, img_embed)

```



(a) **Weather-wise Distribution:** Pedestrian distribution over four weather conditions.



(b) **Country-wise Distribution:** Pedestrian distribution over 12 weather countries.

Figure 2. Attributes distribution in PEARL dataset.

```

# Compute class-balance loss
Lar = con_loss(preds, target, weights)
# Combine loss
loss = (Lt2i + Li2t) / 2 + Lar
# Back-Propagate
loss.backward()
optimizer.step()

```

The Table (3) compares the mean Average (mA) performance of three CLIP methods—CLIP + FP, CLIP + RP, and CLIP + CP—both with and without the additional loss component, L_{ar} . The results show that incorporating L_{ar} im-

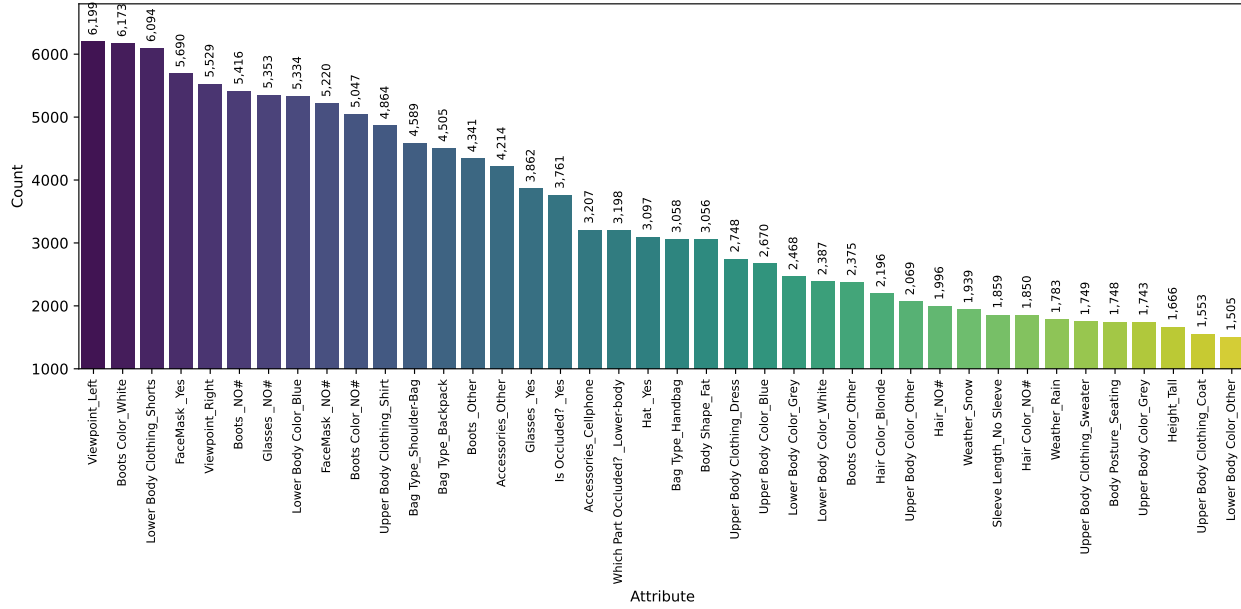


Figure 3. **Long-tail Distribution:** A few selected attribute distributions in the PEARL dataset. Distributions are sorted in decreasing order. First part is included in the manuscript.

Algorithm 1 Inverse Frequency Calculation

```

1: function COMPUTEFREQUENCIES( $\mathcal{D}, C$ )
2:    $counts \leftarrow \text{zeros}(C)$ 
3:   for each  $d \in \mathcal{D}$  do
4:     for each  $a \in d.\text{attributes}$  do
5:        $counts[a] \leftarrow counts[a] + 1$ 
6:     end for
7:   end for
8:   return  $counts$ 
9: end function
10:
11: function COMPUTEINVERSE( $counts$ )
12:    $total \leftarrow \sum counts$ 
13:    $inv\_freqs \leftarrow \frac{total}{C \cdot counts}$ 
14:   return  $inv\_freqs$ 
15: end function
16:
17: function NORMALIZEWEIGHTS( $inv\_freqs$ )
18:   return  $\frac{inv\_freqs}{\sum inv\_freqs} \cdot \text{len}(inv\_freqs)$ 
19: end function
20:
21:  $\mathcal{D}, C \leftarrow$  Training Dataset, Number of Classes
22:  $counts \leftarrow$  COMPUTEFREQUENCIES( $\mathcal{D}, C$ )
23:  $inv\_freqs \leftarrow$  COMPUTEINVERSE( $counts$ )
24:  $weights \leftarrow$  NORMALIZEWEIGHTS( $inv\_freqs$ )

```

proves the mA across all methods. CLIP + CP achieves the

highest mA, both without (84.40) and with (87.29) L_{ar} , indicating its superior performance among the tested methods. The conclusion is that the addition of L_{ar} consistently enhances the performance of CLIP-based methods via adding advantage for minority classes, with CLIP + CP being the most effective.

5. Zero-shot Recognition

Zero-shot CLIP testing on pedestrian attribute recognition involves evaluating the model’s ability to identify and describe attributes without explicit training on those specific attributes. It is possible because CLIP is trained on a large and diverse dataset of images and text from one of the PAR dataset, allowing it to understand and generalize across a wide range of visual and linguistic descriptions derived from the other PAR datasets. To test this capability of the proposed CLIP-based approach, we trained it on training set of Market-1501 [3] *age*, *backpack*, *bag* and tested on attributes from PEARL dataset. Fig. (5) depicts a handful of caption score generated the proposed CLIP + FP baseline on unseen attributes from PEARL dataset. Fig. (4) depicts a few prompt examples used for training the CLIP model. In Fig. (6) a bar graph illustrates a substantial increase in accuracy when Visual-Textual Baseline [1] was specifically trained on attributes unique to the PEARL dataset and then tested on the same attributes within the test set of a target dataset.

In Fig. 7, a comparative analysis was conducted between


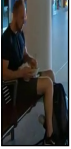
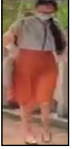
	Full Prompt A photo of a person with gender male age adult hair shot hair color black sleeve length long viewpoint back isOccluded no activity walking pos standing hat no glasses nob bagtype shoulder-bag body shape normal facemask nob weather sunny height normal accessories nothing uppercloth tshirt uppercolor blue lowercloth shorts lowercolor brown.
	Random Prompt A photo of a person with viewpoint right, bagtype backpack, glasses yes, lowercloth shorts, bodypos seating, face mask no, hat no, upper color black, age adult.
	Context Prompt A photo of a person with gender is female, especially a facemask is present, hair long, bodyshape normal, especially lower color is orange, weather sunny, activity walking, viewpoint is front, especially upper color is gray.

Figure 4. **Training Prompts:** Depiction of three suggested prompts used to train the proposed CLIP baseline. Examples are taken from PEARL.

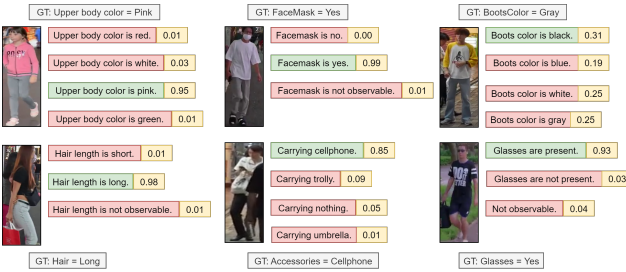


Figure 5. **Zero-shot Caption Score:** Few prediction results predicted by the full prompt CLIP-ViT-B/32 model trained on Market-1501 dataset. Attribute are adopted from PEARL. GT:Ground truth assigned by annotators. Green bars are prediction by the model with highest caption score.

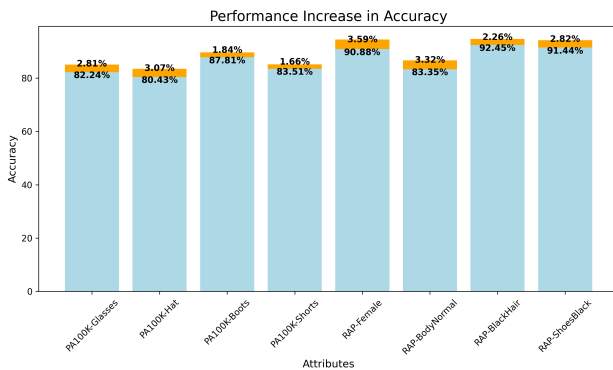


Figure 6. **Attribute-wise Testing:** Bars showing a significant gain in Accuracy when the VTB [1] was explicitly trained on PEARL specific attribute and tested on the same attribute in test set of a target dataset. Herewith,

DeepMar [2], VTB [1], and the proposed baseline using the PEARL30K dataset samples. The bars in the graph represent the prediction probabilities, ranging from 0 to 1, corresponding to each method’s performance.

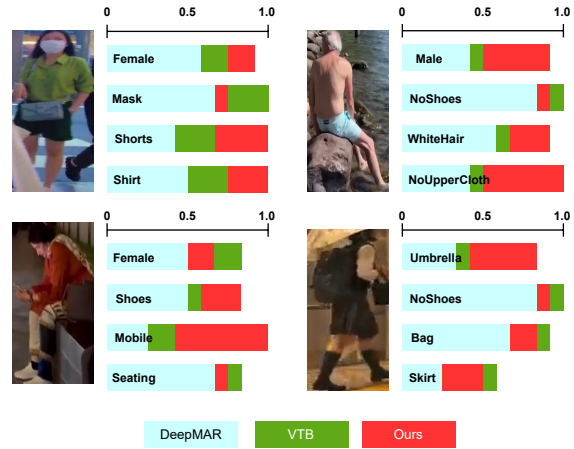


Figure 7. **Prediction Probabilities:** Comparison outcomes among DeepMar [2], VTB [1], and proposed baseline on PEARL30K dataset samples. The bars denote the prediction probabilities between 0 to 1 and are plotted accordingly for each method.

References

- [1] 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. 3, 4
- [2] 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. 4
- [3] 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. 3