Supplementary Materials: Enhanced Visual-Semantic Interaction with Tailored Prompts for Pedestrian Attribute Recognition

Junyi Wu[1], Yan Huang[2*], Min Gao[3], Yuzhen Niu[1], Yuzhong Chen[1*], Qiang Wu[4]
[1] College of Computer and Data Science, Fuzhou University, China
[2] Australian Artificial Intelligence Institute (AAII), University of Technology Sydney, Australia
[3] College of Physics and Information Engineering, Fuzhou University, China
[4] School of Electrical and Data Engineering, University of Technology Sydney, Australia
{junyi.wu-1, min.gao-1}@outlook.com, {yan.huang-7, qiang.wu}@uts.edu.au,
yuzhenniu@gmail.com, yzchen@fzu.edu.cn

1. Experiments

1.1. Implementation Details

We use the image encoder and text encoder of CLIP [6] to extract the visual and linguistic features, respectively. Unless specified, the ViT-L/14 is used as the image encoder. The text encoder remains frozen in the training phase. The number of learnable prompt token L is set to 12. The hyper-parameter λ is set as 0.5. The input images are resized to 224×224 in both training and testing stages. Our EVSITP is trained for 50 epochs using AdamW optimizer with a batch size of 48. The learning rate is set as 0.0001 and decays with the cosine policy.

1.2. Datasets and Evaluation Protocols

PETA [1] is a composite collection comprising 19,000 images aggregated from 10 publicly available person re-identification datasets. Out of these, 8,705 pedestrians are annotated with 61 binary attributes and 4 multi-class attributes. The dataset is divided into training, validation, and testing subsets, containing 9,500, 1,900, and 7,600 pedestrian images, respectively.

PA100K [5] is a large-scale pedestrian dataset collected from 598 outdoor surveillance cameras. It includes 100,000 images, with each annotated for 26 binary attributes. The dataset is divided into 80,000 images for training, while the remaining 20,000 images are split equally into a validation set and a test set, each containing 10,000 images.

RAPv1 [3] comprises 41,585 pedestrian images captured from 26 indoor surveillance cameras in real-world settings. Each image is labeled with 69 binary attributes and 3 multi-class attributes. The dataset is divided into a training set containing 33,268 images and a testing set with 8,317 images.

*Corresponding authors: Yan Huang, Yuzhong Chen

Table 1. Data split of our Celeb-PAR dataset.

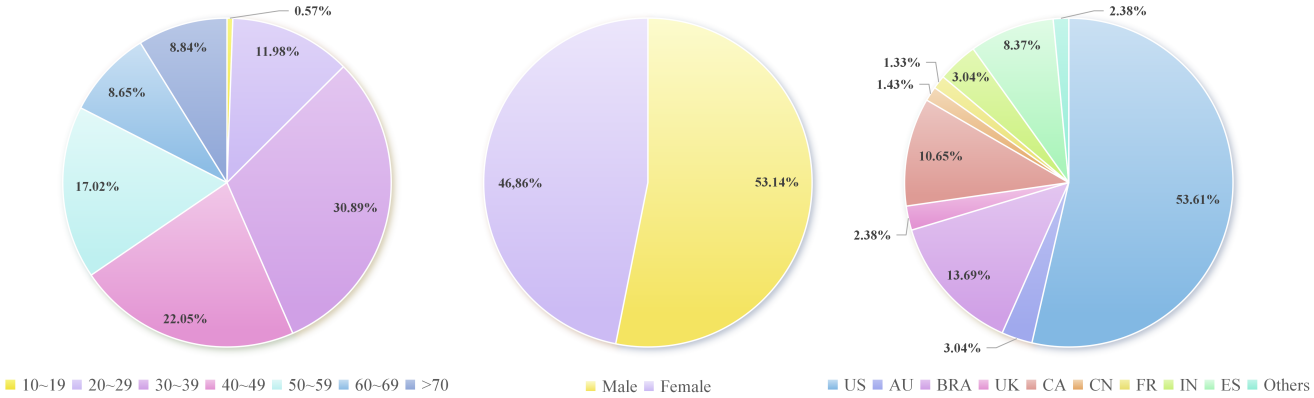
split	training	testing	total
images	23,180	11,006	34,186

RAPv2 [4] includes 84,298 pedestrian images captured from 25 indoor camera scenes in a shopping mall. Consistent with standard dataset splits, it is divided into three subsets: 50,957 images for training, 16,986 images for validation, and the remaining 16,985 images for testing. Each pedestrian image is annotated with 72 attributes.

Evaluations Protocols. Following the previous PAR method [7–9], we adopt five metrics to evaluate the performance of our EVSITP, namely: mean average precision (mA), accuracy (Accu), precision (Prec), recall (Recall), and F1 score (F1). To provide a more balanced assessment, we also report mFive, introduced in [10], which calculates the average across these five metrics. This approach ensures a more comprehensive evaluation, addressing scenarios where a model may excel in one metric but fall short in others.

2. Celeb-PAR Dataset

Our proposed Celeb-PAR possesses the following three core characteristics: (1) It originates from diversified street scenes, ensuring effective generalization of the PAR model to various real-world application scenarios; (2) Our dataset is constructed based on a long-term person re-identification dataset [2], thus encompassing pedestrian clothing changes across different seasons; (3) The pedestrian IDs in the training set are completely separated from those in the test set, a zero-shot setting that aligns more closely with real-world scenario distributions. Fig. 1 gives the statistic information

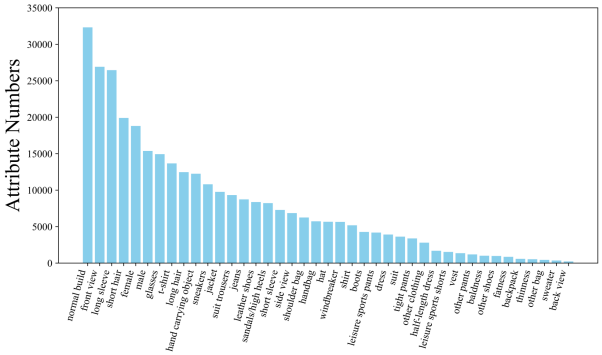


(a) (b) (c)

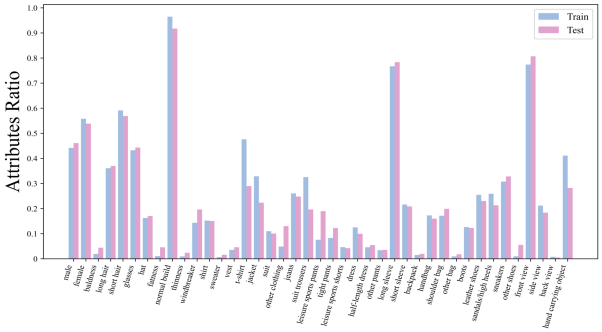
Figure 1. Statistic information of our Celeb-PAR dataset. (a), (b), and (c) show the distributions of age, gender, and nationality, respectively

Table 2. Attribute groups and details defined in our proposed Celeb-PAR.

Attribute Group	Details
Gender	male/female
Hair	baldness/long hair/short hair
Headwear	glasses/hat
Body type	fatness/normal build/thinness
Upper body	windbreaker/shirt/sweater/vest/t-shirt/ jacket/suit/other clothing
Lower body	jeans/suit trousers/leisure sports pants/tight pants/ leisure sports shorts/dress/half-length dress/other pants
Sleeve length	long sleeve/short sleeve
Bag	backpack/handbag/shoulder bag/other bag
Shoes	boots/leather shoes/sandals/high heels/ sneakers/other shoes
View angle	front view/side view/back view
Hand	hand carrying object



(a) Attributes Numbers Distribution



(b) Attributes Ratio Distribution

Figure 2. Statistic properties of our Celeb-PAR dataset. (a) and (b) show the distributions of attribute numbers and attribute ratio.

of our dataset, including the distribution of age, gender, and nationality of celebrities.

Multi-scenarios. Our newly proposed Celeb-PAR dataset is annotated with attribute labels based on the Celeb-reID dataset. Celeb-reID is compiled by collecting images from Google, Bing, and Baidu Image searches using keywords such as "celebrity name + street snapshot" (*e.g.*,"Justin Bieber street snapshot"). Consequently, the clothing worn by pedestrians in this dataset exhibits considerable diversity, and each street snapshot features a unique background. In constructing Celeb-PAR, we merge the training and query sets of Celeb-reID to form the training set of Celeb-PAR, while the gallery set of Celeb-reID is utilized as the testing set of Celeb-PAR, consisting of 23,180 and 11,006 images respectively (referred to Tab. 1).

Multi-seasons. In Tab. 2, we meticulously list 41 attributes annotated in the dataset. By observing the Tab. 2, it becomes evident that our dataset encompasses a variety of clothing attributes suitable for different seasons, such as attire for both summer and winter, which were previously unrepresented in existing datasets. As shown in Fig. 3,

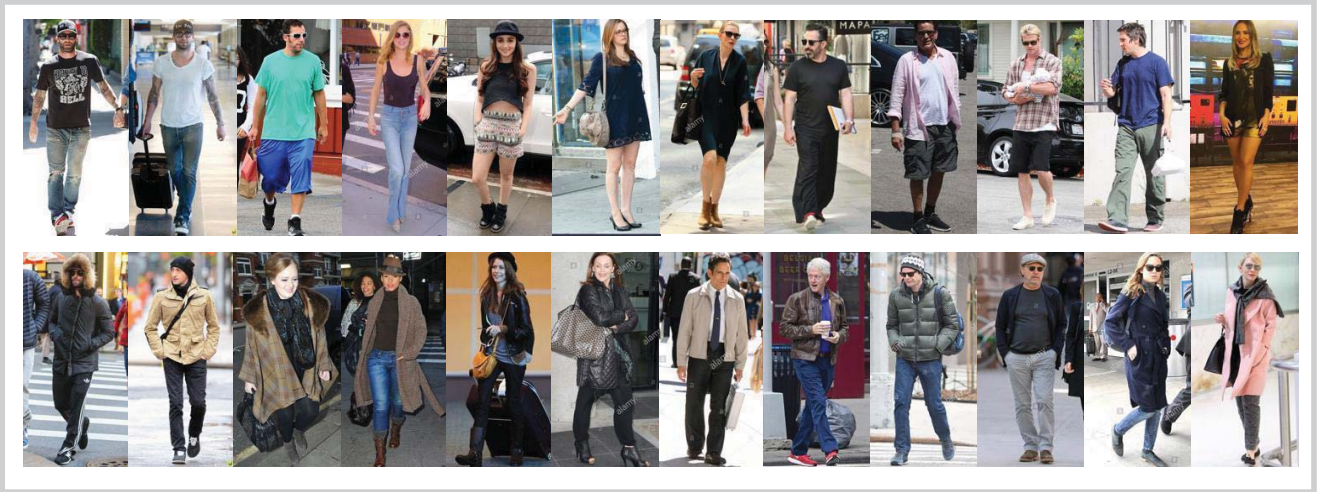


Figure 3. An illustration of representative samples in our newly proposed Celeb-PAR. The first row shows clothing suitable for spring and summer, while the second row displays attire more suited to autumn and winter.

our Celeb-PAR includes clothing styles from various seasons, which aligns more closely with the distribution of real-world scenarios, enabling the trained model to better meet the application requirements in real-life situations. Our Celeb-PAR also exhibits a long-tail effect, similar to existing PAR datasets, depicted in Fig. 2 (a) and (b) and reflects real-world attribute distributions.

Non-overlapping Pedestrian IDs. Our Celeb-PAR, along with datasets PETA$_{zs}$ and RAP$_{zs}$, adheres to the zero-shot dataset partition principle, ensuring that pedestrian IDs in the training and test sets are completely non-overlapping to reflect real-world scenarios. However, PETA$_{zs}$ and RAP$_{zs}$, which were collected from PETA and RAPv2 respectively, do not possess the multi-scene and multi-season the above-mentioned characteristics. In comparison, our Celeb-PAR surpasses PETA$_{zs}$ and RAP$_{zs}$ in terms of the number of images it contains.

Table 3. Performance comparison with varying numbers of FP.

Number of FP	RAPv1		
	mA	Recall	F1
one (FP1)	85.57	87.14	81.99
one (FP2)	85.59	86.29	82.58
one (FP3)	85.48	84.27	82.58
average (one FP)	85.55	85.90	82.38
two (FP1 + FP2)	85.85	84.93	82.37
two (FP1 + FP3)	85.73	86.31	82.41
two (FP2 + FP3)	85.92	87.34	82.35
average (two FPs)	85.83	86.19	82.38
three (FP1 + FP2 + FP3)	86.10	86.65	82.78

3. Ablation Study

3.1. Analysis of Using Three Fixed Prompts

We conducted experiments to assess the impact of using different fixed prompt (FP) templates, with the results displayed in Tab. 3. These results indicate that combining three templates introduced in our paper yields the best performance, suggesting a more effective capture of semantic information within the linguistic modality compared to using one or two FPs alone.

In fact, we also attempted to introduce a fourth fixed prompt template: "(_) can be seen in this pedestrian." However, we found that the inclusion of a fourth template did not improve the performance (mA 85.95 vs. 86.10, Recall 86.57% vs. 86.65%, F1 82.55% vs. 82.78%). This may be due to the introduction of redundant features for the same attribute when more templates are used, slightly affecting the final experimental outcomes.

3.2. Different Conditional Prompts Methods

The previously most classic method for leveraging visual features to refined learnable prompts is CoCoOp [11], but it differs significantly from our approach. In CoCoOp, as illustrated in Fig. 4 (a), the visual feature is fed into a lightweight Meta-Net (*i.e.,* two linear layers and a ReLU) to generate a specific image-conditional token for each image-text pair. This token is then summed with learnable vectors form the prompt. However, CoCoOp has notable drawbacks: it exhibits low computational efficiency and fails to

Table 4. Comparison with different conditional prompts methods.

Method	RAPv1						RAPv2					
	mA	Accu	Prec	Recall	F1	mFive	mA	Accu	Prec	Recall	F1	mFive
CoCoOp	85.02	71.49	80.65	84.77	82.66	80.92	83.06	69.38	78.77	83.61	81.12	79.19
EVSITP	86.10	71.64	79.24	86.65	82.78	81.28	83.83	69.32	77.64	85.13	81.21	79.43

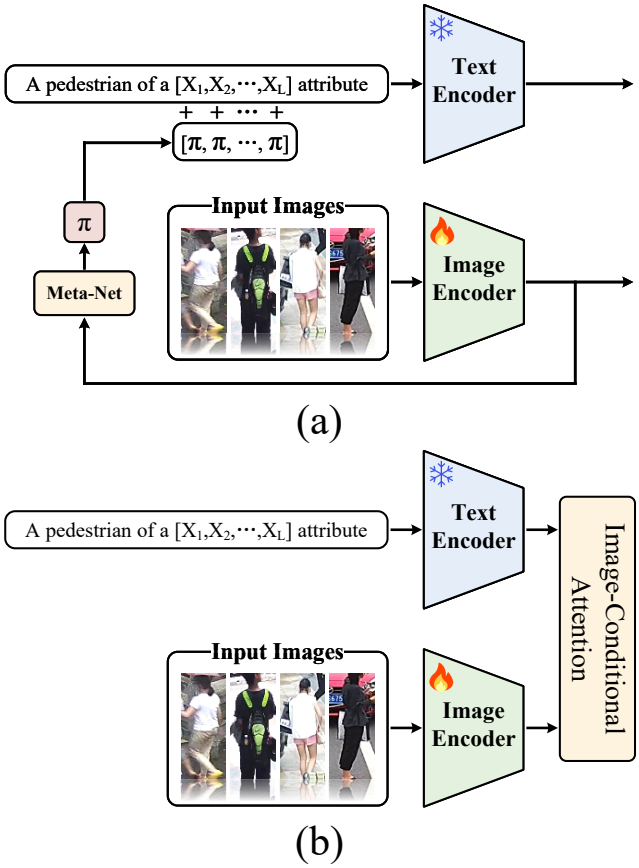


Figure 4. Different conditional prompts feature via visual feature. (a) is the image-conditional method from CoCoOp, (b) is our method.

fully exploits the relationship between visual features and learnable prompts.

Due to the fact that CoCoOp generates unique learnable prompts for each image instance, separate forward propagation calculations are required for each instance. For every input image, a conditional token is produced by the Meta-Net, which is then integrated with the prompt vector and fed into the text encoder. This process is resource-intensive, consuming significant GPU memory and computational power.

To demonstrate this, we conduct a comparative evaluation using CoCoOp's conditional prompts method within our EVSITP framework, measuring a batch forward propagation time against our method. Under identical hardware and parameter settings, our experimental results show that CoCoOp requires 13.09 ms to process one batch, whereas our EVSITP only needs 1.91 ms. Additionally, with the same GPU memory resources, the batch size is limited to 8 when using CoCoOp's conditional prompts, whereas our method allows for a substantial increase in batch size to 48, thereby significantly enhancing computational efficiency.

Moreover, CoCoOp directly utilizes the Meta-Net to convert visual features into tokens and appends them to the prompt vectors. However, this approach has limitations as it fails to capture the individual importance of each prompt token and overlooks the potential relationships between visual features and prompts. In contrast, as shown in Fig. 4 (b), our method incorporates a cross-attention mechanism to achieve image-conditional learnable prompts. This enables a more nuanced capture of the associations between learnable prompt features and visual features. Specifically, cross-attention computes the similarity between learnable prompts and visual features, assigning corresponding weights to each feature, thereby achieving precise feature alignment. This approach enables the model to obtain more accurate and detailed linguistic descriptions, facilitating more effective visual-semantic interactions. As evident from Tab. 4, our method outperforms the image-conditional approach employed by CoCoOp. Specifically, on the RAPv1 and RAPv2 datasets, our method achieves an improvement of 1.08% and 0.77% in terms of mA, respectively .

3.3. Ablation Study of BMIM

Our BMIM incorporates two interaction mechanisms: VLII (integrating visual information into linguistic modal information) and LVII (incorporating semantic information from the linguistic modal into visual features). In Tab. 5, we conduct a thorough ablation study analysis of BMIM. The results from Tab. 5 clearly demonstrate that adopting either VLII or LVII alone can lead to a certain degree of performance improvement. When both are used in combination, the performance enhancement is even more pronounced. This outcome strongly supports the necessity and effectiveness of establishing interaction mechanisms between the two modalities.

Table 5. Ablation study on BMIM.

Method		RAPv1					RAPv2				
VLII	LVII	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
×	×	85.19	71.55	79.88	85.77	82.72	82.88	69.54	78.84	83.71	81.20
✓	×	85.29	71.23	80.39	84.63	82.46	83.18	69.29	77.72	81.02	81.02
×	✓	85.34	72.07	81.43	84.57	82.97	83.31	69.29	77.72	84.60	79.21
✓	✓	86.10	71.64	79.24	86.65	82.78	83.83	69.32	77.64	85.13	81.21

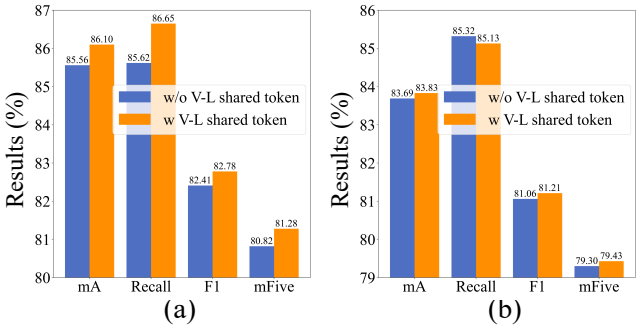


Figure 5. Ablation study on our BMIM. Results are reported on RAPv1 (a) and RAPv2 (b).

3.4. Analysis of V-L Shared Token

BMIM serves as the core of our approach, integrating the two pivotal components fo VLII and LVII. Unlike previous bimodal PAR approaches, we accord equal significance to both visual and linguistic features to facilitate effective interaction between these two modalities. Notably, we introduce an innovative V-L shared token, inspired by the additional class token in ViT, which captures the most critical feature information across all patches. Consequently, we aim to leverage this V-L shared token within BMIM to absorb and integrate information from both modalities, further enhancing the interaction of model information in VLII and LVII.

As illustrated in Fig. 5, we analyze the impact of V-L shared token on the performance of BMIM. It is evident that the introduced V-L shared token results in improvements of 0.54% and 0.14% in terms of mA on RAPv1 and RAPv2, respectively. The performance enhancement demonstrates the effectiveness of the introduced V-L shared token in facilitating modality information interaction.

References

[1] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, pages 789–792, 2014. 1

[2] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE TCSVT*, 30(10):3459–3471, 2019. 1

[3] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 1

[4] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE TIP*, 28(4):1575–1590, 2018. 1

[5] 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 *ICCV*, pages 350–359, 2017. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[7] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z Li. Attention-based pedestrian attribute analysis. *IEEE TIP*, 28(12):6126–6140, 2019. 1

[8] Xiao Wang, Jiandong Jin, Chenglong Li, Jin Tang, Cheng Zhang, and Wei Wang. Pedestrian attribute recognition via clip based prompt vision-language fusion. *IEEE TCSVT*, 2024.

[9] Junyi Wu, Yan Huang, Min Gao, Yuzhen Niu, Mingjing Yang, Zhipeng Gao, and Jianqiang Zhao. Selective and orthogonal feature activation for pedestrian attribute recognition. In *AAAI*, pages 6039–6047, 2024. 1

[10] Yang Yang, Zichang Tan, Prayag Tiwari, Hari Mohan Pandey, Jun Wan, Zhen Lei, Guodong Guo, and Stan Z Li. Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. *IJCV*, pages 1–14, 2021. 1

[11] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 3