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

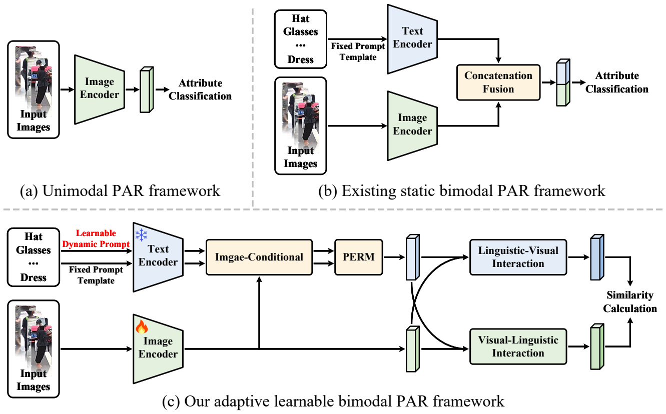
Junyi Wu[1], Yan Huang[2*], Min Gao[3], Yuzhen Niu[1], Yuzhong Chen[1*], Qiang Wu[4]
[1] College of Computer and Data Science, Fuzhou University, China
[2] Australian Artificial Intelligence Institute (AAII), University of Technology Sydney, Australia
[3] College of Physics and Information Engineering, 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

Abstract

Pedestrian attribute recognition (PAR) seeks to predict multiple semantic attributes associated with a specific pedestrian. There are two types of approaches for PAR: unimodal framework and bimodal framework. The former one is to seek a robust visual feature. However, the lack of exploiting semantic feature of linguistic modality is the main concern. The latter one utilizes prompt learning techniques to integrate linguistic data. However, static prompt templates and simple bimodal concatenation cannot to capture the extensive intra-class attribute variability and support active modalities collaboration. In this paper, we propose an Enhanced Visual-Semantic Interaction with Tailored Prompts (EVSITP) framework for PAR. We present an Image-Conditional Dual-Prompt Initialization Module (IDIM) to adaptively generate context-sensitive prompts from visual inputs. Subsequently, a Prompt Enhanced and Regularization Module (PERM) is proposed to strengthen linguistic information from IDIM. We further design a Bimodal Mutual Interaction Module (BMIM) to ensure bidirectional modalities communication. In addition, existing PAR datasets are collected over a short period in limited scenarios, which do not align with real-world scenarios. Therefore, we annotate a long-term person re-identification dataset to create a new PAR dataset, Celeb-PAR. Experiments on several challenging PAR datasets show that our method outperforms state-of-the-art approaches.

1. Introduction

Pedestrian Attribute Recognition (PAR) plays a crucial role in video surveillance systems [46, 50], facilitating the categorization of diverse semantic attributes such as age, gender, and accessories in varied scenarios, *etc.* [23, 36]. As a critical tool for advancing person re-identification [14, 16, 48, 49] and person search [6, 17] techniques, PAR has become a significant area of research within the computer vision community.



(a) Unimodal PAR framework

(b) Existing static bimodal PAR framework

(c) Our adaptive learnable bimodal PAR framework

Figure 1. Different framework for PAR. (a) Existing unimodal PAR framework, (b) existing static bimodal PAR framework, (c) our adaptive learnable bimodal PAR framework.

Despite its importance, PAR faces significant challenges, notably the substantial intra-class attribute variability among pedestrian images. Recent advancements in PAR methodologies [8, 19, 32, 34, 37], built upon unimodal framework, have propelled forward the state of the art. As shown in Fig. 1 (a), the unimodal framework refers to models that exclusively process visual information from images using CNNs [19, 33, 45], Vision Transformers (ViTs) [8, 42, 43] or their combinations [38]. However, these frameworks, while adept at analyzing image data, fail to integrate linguistic modalities, limiting their ability to in-

[*] Corresponding authors: Yan Huang, Yuzhong Chen

corporate rich semantic contexts, which is crucial for addressing complex attribute variations [40].

The emergence of large-scale vision-language models, such as Contrastive Language-Image Pre-Training (CLIP) [31], has demonstrated remarkable capabilities in reasoning across multi-modal data. Unlike the unimodal PAR approach, two innovative PAR methods, VTB [4] and PromptPAR [37], leverage prompt learning techniques [22, 30] to integrate linguistic data, enhancing semantic information processing during PAR model learning. As shown in Fig. 1 (b), these methods enrich the image encoding process by extracting and utilizing semantic information from text encoders, thereby enhancing attribute label representation learning.

Despite their advancements, VTB and PromptPAR exhibit limitations. Firstly, both methods use static prompt templates that merely embed simple attribute labels, such as generating sentences like *"this pedestrian wears a hat."*. This approach fails to capture the extensive intra-class attribute variability, providing only coarse textual insights that cannot adjust to the distinctive attribute features of each pedestrian image. Secondly, these methods do not support active collaboration between visual and textual feature during the attribute learning process. VTB and PromptPAR rely solely on merging visual and textual features using a simple concatenation operation. This method limits the potential for a more nuanced and interactive blending of modalities, failing to fully exploit the rich, contextual interplay between the modalities.

In response to these shortcomings, we propose the Enhanced Visual-Semantic Interaction with Tailored Prompts (EVSITP) framework, depicted in Fig. 1 (c), which exploits the foundational capabilities of CLIP to engage both visual and linguistic modalities. Our framework introduces the Image-Conditional Dual-Prompt Initialization Module (IDIM), which enhances standard prompt templates, such as *"this pedestrian wears a hat."* by learning to specify additional details about the hat, thus adapting the textual output to each specific image. This capability is facilitated by leveraging advanced prompt learning techniques [22, 30] that capture more nuanced contextual and attribute-specific information, thereby addressing challenging intra-class variability and enriching the descriptive accuracy of the model compared with existing fixed static template prompt solutions [4, 37]. Furthermore, we present the Prompt Enhancement and Regularization Module (PERM) that fuses features extracted from static fixed template prompts and adaptive learnable prompts to create more robust, prompt-driven text feature embeddings.

Moreover, for a more nuanced and interactive blending of visual and linguistic modality features, we propose the Bimodal Mutual Interaction Module (BMIM). BMIM ensures bidirectional communication between visual and text features, maintaining balanced importance for both modalities during training. Specifically, BMIM includes Visual-Linguistic Information Interaction (VLII) and Linguistic-Visual Information Interaction (LVII), where VLII generates visual-guided linguistic features, and LVII produces linguistically-enhanced visual features. Given these two types of enriched features, our EVSITP framework enables more precise attribute predictions compared with the simple concatenations between visual and text features used in existing methods [4, 37].

In summary, main contributions of this paper can be summarized in four-fold:

- We introduce the Enhanced Visual-Semantic Interaction with Tailored Prompts (EVSITP) framework. Unlike traditional fixed static template prompts in existing PAR methods, our EVSITP introduces IDIM to adaptively generate context-sensitive prompts from visual inputs.
- We present PERM to aggregate prompts feature, preserving common semantic information while learning context-sensitive features. A regularization loss leverages common semantic features to prevent context-sensitive features from overfitting on the training data.
- We propose BMIM to foster a robust bidirectional interaction between visual and linguistic modalities for a more nuanced and interactive blending of visual and linguistic modality features. This integration enhances the depth and accuracy of PAR beyond previous methods.
- We conduct comprehensive experiments across four PAR datasets and our newly proposed Celeb-PAR. The results demonstrate that our EVSITP framework significantly outperforms existing unimodal and bimodal approaches.

2. Related Work

2.1. Unimodal PAR Methods

Most existing PAR methods only consider the visual modality as input and can be categorized into two types: methods that embed body information and those that apply attention mechanisms.

Unimodal PAR Methods Embedded Body Information: Li *et al.* [25] introduced a pose guided deep model (PGDM) that utilizes a pre-trained pose estimation model to identify body parts and derive relevant body features. Zhang *et al.* [47] incorporated human pose keypoints as supplementary data to guide a deep template matching network, facilitating accurate alignment of attribute-specific regions. Yang *et al.* [45] integrated keypoint estimation and PAR into a multi-task training framework, utilizing keypoints to extract prior knowledge of body parts for feature learning.

Unimodal PAR Methods Applied Attention Mechanism: Tan *et al.* [33] designed three different attention modules (*i.e.,* parsing attention, label attention and spatial attention), aimed at exploring correlated and complementary information. Guo *et al.* [9] employed the CAM network as a backbone, enhancing recognition performance through the refinement of the attention heatmap. Jia *et al.* [19] developed a spatial and semantic consistency (SSC) framework that incorporates two complementary regularizations, aiming to capture inter-image relations through both spatial and semantic lenses for each attribute.

2.2. Bimodal PAR Methods

Currently, two PAR methods, VTB [4] and PromptPAR [37], share a similar starting point with our EVSITP, integrating visual and linguistic modalities to tackle the PAR task. Cheng *et al.* [4] modeled the PAR task as a bimodal problem and introduced a textual modality to sufficiently explore the inherent textual correlations in attribute annotations. Wang *et al.* [37] employed the pre-trained vision-language models to connect the relations between pedestrian images and attribute labels.

Unlike these works, which typically utilize a fixed prompt template, our EVSITP framework recognizes that such templates can only extract common linguistic information and are insufficient for addressing intra-class attribute variability. To overcome this, we propose an innovative image-conditional prompt that generates adaptive linguistic descriptions for specific attributes. Our method distinctively considers both linguistic and visual modalities, enabling more effective integration and interaction between the two.

2.3. Vision-Language Models

Large-scale vision-language models bridge the gap between image and text embedding within a shared embedding space, providing effective across various unimodal and multimodal downstream tasks such as classification [3], action quality assessment [44], and cross-modal retrieval [10, 11]. Foundational models, like CLIP are trained on extensive image-text pairs using contrastive learning objectives.

Inspired by recent advances in NLP, prompt learning has gained attention in the vision domain. Context Optimization (CoOP) [52] explored learnable prompt optimization for few-shot classification. Conditional context optimization (CoCoOp) [51] improved CoOP's performance by learning to generate prompts conditioned on each image instance. Visual Prompt Tuning (VPT) [21] aimed to reduce the number of parameters that need fine-tuning by injecting a set of learnable prompts into the input tokens.

Although previous works have utilized these powerful vision-language models, most lack adequate exploitation of the linguistic modality. In response, we design a dual-prompt-driven strategy for the PAR task. The innovation of our approach lies not only in employing fixed prompt templates but also in introducing a learnable prompt that flexibly guides the PAR model to extract more nuanced semantic information from the linguistic modality, tailored to the unique visual features of each image.

Compared to CoCoOp, our proposed learnable prompts conditioned on visual features demonstrate greater efficiency. CoCoOp generates specific image-conditional prompts for each image-text pair using a small Meta-Net, requiring all prompts to be input into the text encoder to extract image-conditional text features for each image. This process consumes substantial GPU memory and computational resources. In contrast, we apply the same learnable prompts across all image-text pairs and utilize image-conditional cross-attention following the text encoder. This approach allows for the extraction of text features only once for all pairs, while still enabling effective visual-semantic interactions to enhance the learning of more effective prompts. More comparison can be found in the supplementary materials.

3. Methodology

3.1. Image-Conditional Dual-Prompt Initialization

Dual-Prompt. Large vision-language models pre-trained on vast image-text pairs inherently capture rich semantic knowledge. Therefore, by setting appropriate linguistic inputs for each attribute, the embedding extracted by the language model will contain the underlying semantic relations between different labels. In our method, we design a dual-prompt, which consists of three groups fixed prompt template and a learnable prompt.

For the PAR task, we take three groups of fixed prompt template (*i.e., "This pedestrian contains [attribute].", "There is a/an [attribute] in this pedestrian."*, and *"[attribute] in this pedestrian."*) as the inputs for all M attribute labels. Subsequently, the text encoder is adapted to extract three groups M label embeddings, which are denoted as $T_1 \in \mathbb{R}^{M\times d}$, $T_2 \in \mathbb{R}^{M\times d}$, and $T_3 \in \mathbb{R}^{M\times d}$, respectively.

The fixed prompt template could extract rich semantic knowledge. However, such information is static and independent of the input image. For multi-label PAR task, significant intra-class variations in attributes pose challenges, rendering fixed prompt templates inadequate for accommodating the diverse and nuanced linguistic descriptions. Therefore, except for the fixed prompt template, we adopt the learnable prompts to promote the representations of linguistic modality.

We prepend L learnable prompt tokens to each label and yield a learnable prompt (*i.e., "A pedestrian with a* $[X_1]$ $[X_2]$ …… $[X_L]$ *attribute."*), where each $[X_l]$ ($l \in 1, \cdots, L$, $X_l \in \mathbb{R}^d$) represents a learnable text prompt tokens with dimensions consistent with word embedding. Then the sequences are fed into the text encoder to extract M label embedding $T_L \in \mathbb{R}^{M\times d}$.

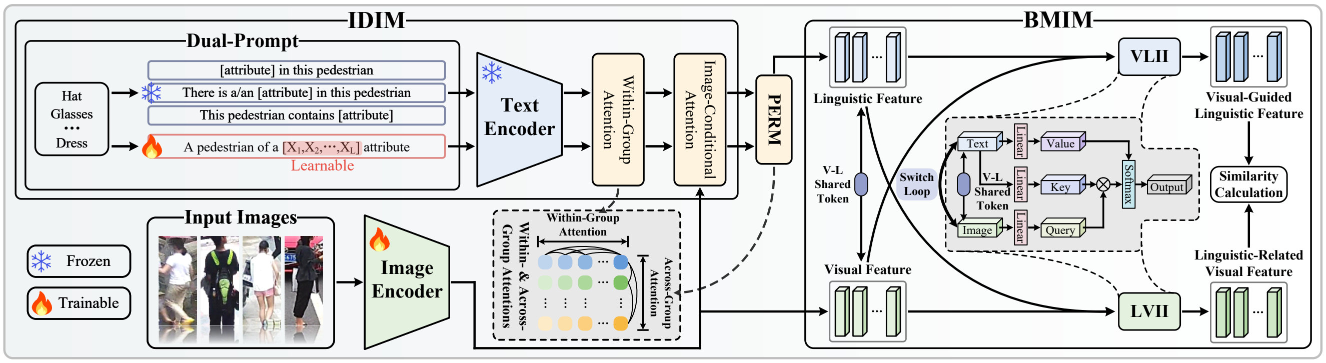


Figure 2. Our EVSITP architecture. Overall, our approach consists of CLIP, IDIM, PERM, and BMIM. IDIM includes a dual-prompt, within-group attention, and image-conditional attention. PERM includes across-group attention and prompt regularization. BMIM includes visual-linguistic information interaction (VLII), linguistic-visual information interaction (LVII), and visual-linguistic shared token (V-L shared token).

Within-Group Attention. Exploring the correlations between different attributes in multi-label PAR tasks can contribute to performance improvement [34, 39]. Guided by this consideration, we introduce a within-group attention to process the label embeddings obtained from fixed prompt (T_1, T_2, and T_3) and learnable prompt (T_L), thereby capturing the label co-occurrence relationships among individual features. It processes linguistic features of all attributes generated under a fixed prompt template. Self-attention demonstrates robust capabilities in modeling relationships within an input sequence [7]. Therefore, we utilize a self-attention to achieve our within-group attention, which can be formulated as:

$$
\begin{aligned}
M_1^{\text{out}} &= \operatorname{softmax}\left(\frac{T_1 W_Q \left(T_1 W_K\right)^{trs}}{\sqrt{d}}\right), \\
T_1^{\text{out}} &= \operatorname{softmax}\left(\frac{T_1 W_Q \left(T_1 W_K\right)^{trs}}{\sqrt{d}}\right)\left(\mathrm{T}_1 W_V\right),
\end{aligned}
\tag{1}
$$

where W_Q, W_K, and W_V are learnable weights of self-attention module. $T_1^{\text{out}} \in \mathbb{R}^{M\times d}$ represents relation-aware label embedding. $M_1^{\text{out}} \in \mathbb{R}^{M\times M}$ denotes the attention map that captures the pair-wise relations of vectors in T_1. trs is short for transposing. Since other label embeddings (T_2, T_3, and T_L) are processed using Eq. 1, the explicit formula presentations are hereby omitted for brevity (T_2^{out}, T_3^{out}, T_L^{out}, M_2^{out}, M_3^{out}, and M_L^{out}).

Image-Conditional Attention. To enable the learnable prompts to flexibly adapt to the attribute variations in each image, we design an image-conditional attention. It can utilize visual feature to guide the entire learnable prompt, enabling specific linguistic descriptions to be generated for each image. Meanwhile, the fixed prompts is a sentence formed by embedding attribute names to the fixed template. The image-conditional attention is also applied to the fixed prompt feature. Our objective is to utilize visual feature to modulate within-group attention, enhancing linguistic descriptions that match the attributes present in each image and suppressing those associated with attributes not found in the image. Cross-attention takes different sources as inputs and is good at capturing different inputs interactions. Therefore, we leverage cross-attention to condition the label embedding on visual feature, which is formulated as:

$$
T_L^{\prime ica} = \operatorname{softmax}\left(\frac{T_{\text{L}}^{\text{out}} W_Q^{\text{ica}} \left(F W_K^{\text{ica}}\right)^{trs}}{\sqrt{d}}\right)\left(F W_V^{\text{ica}}\right),
\tag{2}
$$

where W_Q^{ica}, W_K^{ica}, and W_V^{ica} are learnable weights of the cross-attention. F represents the extracted visual feature from the image encoder of CLIP. After that, we still perform a self-attention operation on $T_L^{\prime ica}$ to perceive the inter-relation between different attribute labels, which can obtain relation-enhanced linguistic representation $T_{\text{L}}^{\text{ica}}$ and relation-enhanced linguistic feature map $M_{\text{L}}^{\text{ica}}$. Since other label embeddings (T_1^{out}, T_2^{out}, and T_3^{out}) are processed using Eq. 2 and a self-attention, the explicit formula presentations are hereby omitted for brevity (T_1^{ica}, T_2^{ica}, T_3^{ica}, M_1^{ica}, M_2^{ica}, and M_3^{ica}).

3.2. Prompt Enhancement and Regularization

After passing through IDIM, we can obtain four linguistic feature (*i.e.,* T_1^{ica}, T_2^{ica}, T_3^{ica}, and $T_{\text{L}}^{\text{ica}}$). To fully leverage the strengths of different groups prompt, we propose a PERM. PERM primarily focuses on the following two aspects: (1) beyond the within-group relationships emphasized by IDIM, the across-group relationships should be considered; (2) it is necessary to prevent the introduced learnable prompt tokens from overfitting to the training data.

Prompt Enhancement via Across-Group Attention. In IDIM, we take into account the within-group correlations among different label within the same group of linguistic label embedding. Further, we also consider the correlations between corresponding attribute features across different groups. Therefore, we introduce an across-group attention aimed at exploring the relationships between the same attributes across different groups, thereby enhancing the representation capability of each attribute feature. It involves analyzing the relationships of the same attribute feature generated across different prompt templates. To achieve across-group attention in manner similar to within-group attention, we continue to employ the self-attention mechanism. We define the linguistic features (T_1^{ica}, T_2^{ica}, T_3^{ica}) processed through across-group attention as $T_{\text{across}}^{\text{out}}$. Both within-group and across-group attention employ the self-attention mechanism, they exhibit significant differences in processing linguistic inputs. Specifically, the linguistic input dimension for within-group attention is $1 \times M \times d$, it focus on the correlation between different attributes. The linguistic input dimension for across-group attention is $M \times 3 \times d$, which emphasizes calculating the correlations of the same attribute across different prompt templates. Finally, we utilize channel concatenation operation to fuse $T_{\text{L}}^{\text{ica}}$ and $T_{\text{across}}^{\text{out}}$, which can obtain $T_{\text{perm}}^{\text{out}}$. Then, we employ two fully-connected layers to restore the channel number of $T_{\text{perm}}^{\text{out}}$ to match the original channel number of T_L^{ica}.

Prompt Regularization. Despite the learnable prompt tokens comprising a limited number of parameters, the resulting prompts are susceptible to overfitting on the training data [1]. To address this issue, we propose an effective regularization loss that encourages the learned prompts to remain close to their linguistic counterparts in the embedding space. This loss can be naturally viewed as a regularizer that prevents the learned prompt-conditioned features from diverging too much from the hand-crafted ones. The regularization loss is formulated as:

$$\mathcal{L}_{reg} = \frac{1}{M} \sum_{i=1}^{M} \left\| t_i^{ica} - t_i \right\|_2^2, \tag{3}$$

where $T_{\text{across}}^{\text{out}} = \{t_1, \cdots t_i, \cdots, t_M\}$ and $T_L^{\text{ica}} = \{t_i^{ica}, \cdots t_i^{ica}, \cdots, t_M^{ica}\}$. $\| \cdot \|$ represents the euclidean distance. We can minimize the distance between t_i and t_i^{ica} for boosting the generalization ability.

3.3. Bimodal Mutual Interaction Module

Our EVSITP utilizes the image encoder and text encoder of CLIP to extract the visual feature F and linguistic representation $T_{\text{perm}}^{\text{out}}$. In contrast to prior works (*i.e.*, VTB [4] and PromptPAR [37]), which treat linguistic representation merely as a supplement to visual feature without considering the mutual interaction between bimodal features.

In our BMIM, we introduce three components: VLII, LVII, and visual-linguistic shared token, which are designed to transform the input sequences within cross-attention to facilitate the interaction of bimodal information. To further enhance the strengths of fixed and learnable prompts and the relation among labels, we employ a trainable parameter α that can adaptively assign different aggregation weights to $M_{\text{across}}^{\text{out}}$ and M_L^{ica}. Specifically, the weight for $M_{\text{across}}^{\text{out}}$ is α and M_L^{ica} is $1 - \alpha$, and the aggregated result T is then obtained by applying this weighted sum to $T_{\text{perm}}^{\text{out}}$. $M_{\text{across}}^{\text{out}}$ and M_L^{ica} represent the attention maps learned from $T_{\text{across}}^{\text{out}}$ and T_L^{ica}, respectively. They capture the co-occurrence relations between attribute pairs in the PAR dataset.

Specifically, we employ the cross-attention mechanism to implement two module: VILL and LVII. In VLII, we integrate visual feature into linguistic feature (label embedding) , utilizing linguistic feature T as the query for cross-attention and generating visual-guided linguistic feature T_V. The generated linguistic feature T_V is context-sensitive embeddings from visual inputs, providing tailored descriptions for each attribute. Conversely, LVII aims to integrate linguistic feature into visual feature , where visual feature F are used as the query for cross-attention and produces linguistic-related visual feature F_L. The resulted visual feature F_L is semantic-aware representations from label semantic space, supplementing the semantic information lacking in visual features.

Inspired by the learnable class token of ViT [7], our BMIM embeds the introduced V-L shared token to F and T to further facilitate interaction between visual and linguistic modalities. The V-L shared token is integrated into both the visual and linguistic pathways prior to the cross-attention operation. This placement allows the token to act as a bridge, facilitating synchronized processing of visual and linguistic information. This synchronization significantly improves the model's ability to amalgamate and leverage information from both modalities, ultimately enhancing attribute recognition.

Following the mutual interaction between visual feature and linguistic representation, the resulting output T_V and F_L serve as final attribute recognition. In traditional PAR methods, the final feature outputs obtained from CNN or Transformer are typically projected onto the label space using a linear layer for final prediction. Different from these methods, we calculate the similarity between the visual-guided linguistic feature and the linguistic-related visual feature to conduct classification directly within the feature space. The classification probability $\hat{y}_j$ for j-th attribute can be calculated as:

$$\hat{y}_j = \text{sigmoid}\left(\mathit{GAP}(F_L) \cdot T_V\right). \tag{4}$$

where GAP represents the global average pooling, $\text{GAP}(F_L) \in \mathbb{R}^{B \times 1 \times d}$, and $T_V \in \mathbb{R}^{B \times d \times M}$. Eq. 4 is followed by a squeeze operation, which can obtain $\hat{y}_j \in \mathbb{R}^{B \times M}$.

In the PAR task, we compute the dot-product similarity between F_L and T_V to determine the attribute probability. The traditional similarity leverages relative measurement between visual feature vector of each attribute and all prompts feature vectors. Our EVSITP only computes the similarity between each visual feature vector and its corresponding linguistic prompts feature, which can reduce computational redundancy and enhance computational efficiency.

3.4. Optimization

In our EVSITP, we adopt the binary cross-entropy loss (BCELoss) with sigmoid function as the optimization target. Together with the regularization loss of Eq. 3, the final objective is defined as:

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda \mathcal{L}_{reg}, \tag{5}$$

where λ is a hyper-parameter to make a trade-off between the two losses.

4. Celeb-PAR Dataset

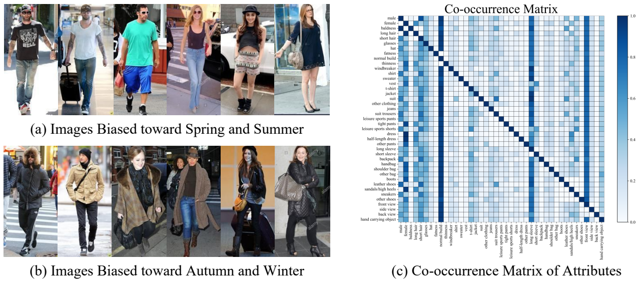
As shown in Tab. 1, all existing public PAR datasets (*i.e.,* PETA [5], PA100K [29], RAPv1 [24], and RAPv2 [26]) are derived from pedestrian images in limited scenes (*e.g.,* shopping mall) or specific season. However, the PAR algorithms in real-world scenarios should demonstrate robustness to the attributes that may appear across different scenes and seasons. For example, if the PAR dataset is collected in summer , it is unlikely to include attributes such as coats or sweaters, which may hinder the model's ability to generalize to winter attributes.

To provide a publicly available benchmark, we construct a new PAR dataset (named Celeb-PAR) based on the long-term person re-identification dataset (Celeb-reID) [12]. Celeb-reID is collected in highly diverse real-world environments/backgrounds, encompassing multiple camera views and a variety of shooting conditions, with each individual's clothing exhibiting significant dynamic changes. Celeb-reID is a benchmark dataset for evaluating long-term person re-identification algorithms [13, 15, 41]. Based on this dataset, we annotate 34,186 images (including 20,208 images in the training set and 13,978 images in the test set), with each images containing 41 attributes.

To ensure the dataset aligns more closely with real-world scenarios, we ensure that pedestrian IDs in the training set and test set are completely non-overlapping. Compared to existing non-overlapping ID PAR datasets (*i.e.,* PETA$_{zs}$ and RAP$_{zs}$), our newly proposed dataset exhibits a notable increase in image number. More importantly, these samples include clothing outfits from a variety of different scenarios and seasons (referred to Fig. 3 (a), (b)), thereby rendering the attribute information contained within our dataset more diverse and abundant.

Table 1. The statistics of our Celeb-PAR dataset and other PAR datasets.

Dataset	Year	Attributes	Images	Non-overlapping	Multi-seasons	Multi-scenarios
PETA	2014	61	19,000	×	×	×
PA100K	2017	26	100,000	×	×	×
RAP1	2016	69	41,585	×	×	×
RAPv2	2019	76	84,928	×	×	×
PETA$_{zs}$	2021	35	19,000	✓	×	×
RAP$_{zs}$	2021	53	26,638	✓	×	×
Celeb-PAR(Ours)	2024	41	34,186	✓	✓	✓



(a) Images Biased toward Spring and Summer

(b) Images Biased toward Autumn and Winter

(c) Co-occurrence Matrix of Attributes

Figure 3. The statistical properties and illustration of representative samples in our newly proposed Celeb-PAR dataset.

Fig. 3 (c) presents the co-occurrence matrix of pedestrian attributes, where each cell represents the frequency of two attributes appearing together. Darker areas indicate higher co-occurrence frequency. More information about our newly proposed dataset can be found in the supplementary materials.

In Sec. 5.1, we reproduce some state-of-the-art PAR methods on Celeb-PAR and compare their performance.

5. Experiments

Comprehensive evaluations are conducted to verify the effectiveness of the proposed EVSITP. The experiments are conducted on four PAR benchmark datasets. Details about these datasets and evaluation protocols can be found in the supplementary material. More implementation details and ablation studies are shown in the supplementary material.

5.1. Comparison with State-of-the-Art Methods

In Tab. 2, we show the performance comparison between our EVSITP and several recent SOTA methods (DAFL [20], Label2Label [28], SOFAFormer [42], AttExpIB-Net [40], VTB [4], and PromptPAR [37]) on PETA, PA100K, RAPv1, and RAPv2. It is evident that our method achieves the best performance in terms of mA on four datasets, respectively. In mA, our EVSITP outperforms the second-best performance PromptPAR [37] by 0.89%, 1.19%, 0.65%, and 0.69% on four datasets, respectively.

PA100K is the most challenging dataset among the four PAR datasets, with PromptPAR previously achieving SOTA performance. Our method surpasses all evaluation metrics, with specific improvements of 1.19% in mA, 0.76%

Table 2. Performance comparison of SOTA methods on the PETA, PA100K, RAPv1, and RAPv2 datasets. Performance in five metrics, including mean Accuracy (mA), accuracy (Accu), precision (Prec), recall, and F1, is evaluated. The **first** and second highest scores are represented by bold font and underline respectively.

Method	PETA					PA100K					RAPv1					RAPv2				
	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
PGDM [25]	82.97	78.08	86.86	84.68	85.76	74.95	73.08	84.36	82.24	83.29	74.31	64.57	78.86	75.90	77.35	-	-	-	-	-
GRL [50]	86.70	-	84.34	88.82	86.51	-	-	-	-	-	81.20	-	77.70	80.90	79.29	-	-	-	-	-
VRKD [27]	84.90	80.95	88.37	87.47	87.91	77.87	78.49	88.42	86.08	87.24	78.30	69.79	82.13	80.35	81.23	-	-	-	-	-
ALM [35]	86.30	79.52	85.65	88.09	86.85	80.68	77.08	84.21	88.84	86.46	81.87	68.17	74.71	86.48	80.16	79.79	64.79	73.93	82.03	77.77
SSC$_{hard}$ [19]	85.92	78.53	86.31	86.23	85.96	81.02	78.42	86.39	87.55	86.55	82.14	68.16	77.87	82.88	79.87	-	-	-	-	-
IAA-Caps [39]	85.27	78.04	86.08	85.80	85.64	81.94	80.31	88.36	88.01	87.80	81.72	68.47	79.56	82.06	80.37	79.99	68.03	**78.75**	81.37	79.69
FEMDAR [2]	84.73	78.45	86.79	85.69	85.90	81.02	79.65	87.99	87.45	87.32	79.71	66.88	79.11	79.24	78.76	-	-	-	-	-
EALC$_{w.ACM}$ [38]	85.94	80.58	87.49	87.38	87.44	80.52	80.13	87.18	88.59	87.88	82.09	69.30	79.63	82.77	81.17	-	-	-	-	-
SOFAFormer [42]	87.10	81.10	87.80	88.40	87.80	83.40	81.10	88.40	89.00	88.30	83.40	70.00	**80.00**	83.00	81.20	81.90	68.60	78.00	83.00	80.20
AttExpIB-Net [40]	85.90	77.58	84.88	86.36	85.32	83.23	79.42	86.70	88.60	87.23	82.46	68.81	79.67	81.63	80.25	80.60	67.31	78.66	80.38	79.15
DAFL [20]	87.07	78.88	85.78	87.03	86.40	83.54	80.13	87.01	89.19	88.09	83.72	68.18	77.41	83.39	80.29	81.04	66.70	76.39	82.07	79.13
VTB [4]	85.31	79.60	86.76	87.17	86.71	83.72	80.89	87.88	89.30	88.21	82.67	69.44	78.28	84.39	80.84	81.34	67.48	76.41	83.32	79.35
VTB+ [4]	86.34	79.59	86.66	87.82	86.97	85.30	81.76	87.87	90.67	88.86	83.69	69.78	78.09	85.21	81.10	81.36	67.58	76.19	84.00	79.52
PromptPAR [37]	88.76	82.84	89.04	89.74	89.18	87.47	83.78	89.27	91.70	90.15	85.45	71.61	79.64	86.05	82.38	83.14	**69.62**	77.42	**85.73**	81.00
EVSITP (Ours)	**89.65**	**83.93**	**89.67**	**90.73**	**90.20**	**88.66**	**84.54**	**89.90**	**92.09**	**90.98**	**86.10**	**71.64**	79.24	**86.65**	**82.78**	**83.83**	69.32	77.64	85.13	**81.21**

Table 3. Comparison with state-of-the-art methods on Celeb-PAR.

Method	mA	Accu	Prec	Recall	F1	mFive
IAA-Caps* [39]	74.95	65.39	76.28	80.01	77.78	74.88
Label2Label [28]	75.13	64.30	73.53	81.38	76.95	74.26
SSC* [19]	73.83	64.10	74.56	79.93	77.15	73.91
AttExpIB-Net* [40]	75.82	65.10	75.77	80.03	77.52	74.85
SOFAFormer* [42]	75.81	65.53	75.57	80.96	77.88	75.15
RethinkingPAR [18]	73.61	64.41	76.12	78.55	77.32	74.00
ADFL [53]	74.95	62.91	73.71	78.94	76.24	73.35
VTB [4]	75.47	64.88	74.42	81.34	77.42	74.71
PromptPAR [37]	78.36	67.88	76.48	**83.72**	79.66	77.22
EVSITP (Ours)	**78.61**	**68.57**	**78.35**	82.56	**80.40**	**77.70**

* represents the results obtained from our reproduced or unofficial code.

Table 4. Ablation study on the proposed modules. FP and LP represent fixed prompt and learnable prompt, respectively.

FP	LP	PERM	BMIM	RAPv1			RAPv2		
				mA	Recall	F1	mA	Recall	F1
×	×	×	×	82.57	77.50	80.87	79.15	74.62	79.28
✓	×	×	×	83.87	83.02	82.16	81.66	82.79	81.30
×	✓	×	×	84.25	85.34	82.33	82.35	82.73	80.61
✓	✓	✓	×	85.19	85.77	82.72	82.88	83.71	81.20
✓	✓	✓	✓	86.10	86.65	82.78	83.83	85.13	81.21

in Accu, 0.63% in Prec, 0.39% in Recall, and 0.83% in F1. Although both methods incorporate the linguistic modality as an information source, our approach further considers the intra-class attribute variability and optimizes the fusion between the two modalities compared to PromptPAR.

Compared to another bimodal method VTB , under the condition of using the same visual encoder, our method achieves significant performance improvements across four datasets, achieving mA improvement of 3.31%, 3.36%, 2.41%, and 2.47%, respectively. On the other hand, compared to the previously best-performing unimodal method DAFL [20], our approach gains significant improvements in evaluation metrics across every dataset. This notable advancement is primarily attributed to the incorporation of the linguistic modality, which enables the PAR model to access rich semantic information. By effectively integrating these semantic information with visual features, our method demonstrates superior performance.

5.2. Results on Our Newly Proposed Celeb-PAR

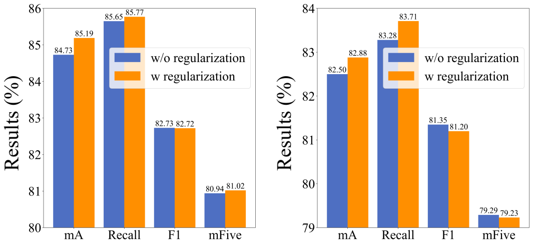
Celeb-PAR is our newly proposed dataset, which is constructed based the criteria of a highly variable clothchanging person re-identification dataset, thus exhibiting more significant intra-class attribute variability. In Tab. 3, we present the performance of selected unimodal PAR method (IAA-Caps [39], Label2Label [28], SSC [19], AttExpIB-Net [40], SOFAFormer [42], RethinkingPAR [18], and ADFL [53]) and bimodal PAR methods (PromptPAR [37] and VTB [4]) on the Celeb-PAR dataset.

Based on the result presented in Tab. 3, our method performs comparably to PromptPAR in terms of performance. Our method achieves SOTA on four evaluation metrics (*i.e.,* mA, Accu, Prec, and F1), while PromptPAR ranks first on one metrics (*i.e.,* Recall). Our EVSITP surpasses PromptPAR by 0.25%, 0.69%, 1.87%, and 0.74% across these four metrics, respectively. To more comprehensively assess the generalization ability of our method, we leverage mFive proposed in [46]. On this metric, our method demonstrates superior performance, surpassing PromptPAR by 0.48%.

5.3. Ablation Study

The ablation study is provided in Tab. 4:

Effects of Fixed Prompt. It is observed that with fixed prompt, the performance increases from 82.57% to 83.87% (from 79.15% to 81.66%) in term of mA on RAPv1 and RAPv2. This demonstrates that compared to merely us-



(a) Results on RAPv1

(b) Results on RAPv2

Figure 4. Ablation study on our PERM.

ing attribute names as input for the text encoder, employing fixed prompt templates can more effectively exploit the rich semantic information within the linguistic modality.

Effects of Learnable Prompt. Compared to fixed prompt templates, we introduce the learnable prompt tokens to address intra-class attribute variability. Experimental results (83.87% (81.66%) *vs.* 84.25% (82.35%)) indicate that these learnable prompt tokens can yield greater performance enhancements.

Effect of PERM. In our EVSITP, we adopt dual-prompt strategy and utilize PERM to better enhance the acquired linguistic features while prevent overfitting to the training dataset. As shown in Tab. 4, the performance of dual-prompt linguistic features enhanced by PERM surpasses that of using any single prompt feature alone. Under the enhancement of PERM, our EVSITP achieves 85.19% and 82.88% in terms of mA on RAPv1 and RAPv2, respectively.

Effect of BMIM. Our introduced BMIM approach treats visual and linguistic features equally, fully facilitating the fusion of these two modality features, and thus enabling more effective application to the final attribute recognition task. As shown in Tab. 4, our BMIM can achieve performance improvements of 0.91% and 0.95%, respectively. The performance improvement demonstrates that equal treatment of visual and linguistic features can facilitate better feature representation.

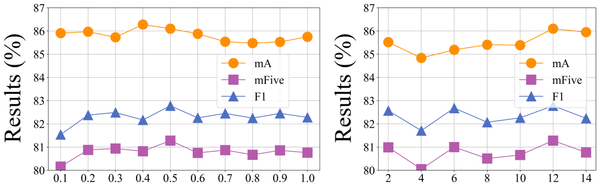
Analysis of PERM. In Fig. 4, we analyze the impact of the absence of regularization on overall performance. Despite the learnable prompt tokens comprising a limited number of parameters, the resulting prompts are susceptible to overfitting on the training data. As shown in Fig. 4, it is intuitive to observe that the proposed regularization loss can effectively mitigate the overfitting phenomenon to a certain extent, resulting in performance improvements. With the application of our regularization loss, PERM can achieve an additional performance improvement of 0.46% and 0.38% in terms of mA on RAPv1 and RAPv2, respectively.

5.4. Generalization Analysis on PETA$_{zs}$ and RAP$_{zs}$

The datasets, PETA$_{zs}$ and RAPv2$_{zs}$, have non-overlapping pedestrian IDs between the training and testing, which can provide a better measure of the model's generalization ability. The experimental results on these datasets are presented in Tab. 5. It is observed that our EVSITP achieves the best performance on two datasets. Compared with previous SOTA bimodal methods, our method achieves a significant improvement in terms of mA. Compared to previous SOTA unimodal methods, our approach significantly outperforms SOFAFormer by a large margin on both datasets.

Table 5. Performance comparison on PETA$_{zs}$ and RAP$_{zs}$.

Method	PETA$_{zs}$				RAP$_{zs}$			
	mA	Accu	Recall	F1	mA	Accu	Recall	F1
MsVAA	71.03	59.38	70.10	72.37	71.32	63.59	76.62	76.44
VAC	71.05	58.90	70.48	72.13	70.20	65.45	76.65	77.07
ALM	70.67	58.56	71.31	71.65	71.97	64.52	77.74	77.06
IAA-Caps	72.53	60.07	73.05	73.07	72.00	64.61	77.06	77.15
SOFAFormer	74.70	62.10	75.10	74.60	73.90	66.30	79.40	78.40
VTB	75.13	60.50	74.40	73.38	75.76	64.73	80.85	77.35
EVSITP	**77.83**	**65.38**	**80.75**	**78.07**	**78.82**	**68.92**	**82.93**	**81.42**



(a) Parameter λ

(b) Parameter L

Figure 5. Sensitivity of parameters λ and L.

5.5. Parameter Analysis

The parameter λ in Eq. 5 determines the trade-off between $\mathcal{L}_{bce}$ and $\mathcal{L}_{reg}$. As λ increases from 0.1 to 1.0 (Fig. 5 (a)), the mA exhibits a trend of first increasing and then decreasing, with the highest and second-highest performance results achieved at 0.4 and 0.5, respectively. After a comprehensive comparison between F1 and mFive, we set λ to 0.5.

Fig. 5 (b) demonstrates the impact of different prependable learnable prompt tokens on overall performance. There is an initial upward trend as L increases from 4 to 12, reaching a peak at L = 12, followed by a decline thereafter.

6. Conclusion

In this work, we present EVSITP, a novel visual-linguistic representation learning framework for the PAR task. To address the defects of existing bimodal PAR method, our EVSITP proposes IDIM to adaptively generate context-sensitive prompts from visual inputs. We further propose the PERM to enhance and aggregate linguistic feature. BMIM is designed to foster a robust bidirectional interaction between two modalities. Extensive experiments on the public datasets and the newly proposed Celeb-PAR demonstrate that EVSITP achieves SOTA performance.

Acknowledge

This work was supported in part by the National Natural Science Foundation of China under Grant 62471142, in part by the Natural Science Foundation of Fujian Province, China under Grants 2023J01067, in part by Industry-Academy Cooperation Project under Grant 2024H6006, in part by the Collaborative Innovation Platform Project of Fuzhou City under Grant 2023-P-002, in part by the Key Technology Innovation Project for Focused Research and Industrialization in the Software Industry of Fujian Province, and in part by the Key Research and Industrialization Project of Technological Innovation in Fujian Province under Grant 2024XQ002.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *CVPR*, pages 23232–23241, 2023. 5

[2] Yilu Cao, Yuchun Fang, Yaofang Zhang, Xiaoyu Hou, Kunlin Zhang, and Wei Huang. A novel self-boosting dual-branch model for pedestrian attribute recognition. *Signal Processing: Image Communication*, 115:116961, 2023. 7

[3] Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: adaptive reweighting via effective area for long-tailed classification. In *ICCV*, pages 19277–19287, 2023. 3

[4] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE TCSVT*, 32(10):6994–7004, 2022. 2, 3, 5, 6, 7

[5] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM MM*, pages 789–792, 2014. 6

[6] Qi Dong, Shaogang Gong, and Xiatian Zhu. Person search by text attribute query as zero-shot learning. In *ICCV*, pages 3652–3661, 2019. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4, 5

[8] Xinwen Fan, Yukang Zhang, Lu Yang, and Wang Hanzi. Parformer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE TCSVT*, 34(1):411–423, 2023. 1

[9] Hao Guo, Xiaochuan Fan, and Song Wang. Human attribute recognition by refining attention heat map. *PRL*, 94:38–45, 2017. 3

[10] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *CVPR*, pages 18962–18972, 2023. 3

[11] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *CVPR*, pages 6565–6574, 2023. 3

[12] 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. 6

[13] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *ICCV*, pages 11895–11904, 2021. 6

[14] Yan Huang, Qiang Wu, Jingsong Xu, Yi Zhong, and Zhaoxiang Zhang. Unsupervised domain adaptation with background shift mitigating for person re-identification. *IJCV*, 129(7):2244–2263, 2021. 1

[15] Yan Huang, Zhang Zhang, Qiang Wu, Yi Zhong, and Liang Wang. Enhancing person re-identification performance through in vivo learning. *IEEE TIP*, 33:639–654, 2023. 6

[16] Yan Huang, Zhang Zhang, Qiang Wu, Yi Zhong, and Liang Wang. Attribute-guided pedestrian retrieval: Bridging person re-id with internal attribute variability. In *CVPR*, pages 17689–17699, 2024. 1

[17] Boseung Jeong, Jicheol Park, and Suha Kwak. Asmr: Learning attribute-based person search with adaptive semantic margin regularizer. In *ICCV*, pages 12016–12025, 2021. 1

[18] Jian Jia, Houjie Huang, Wenjie Yang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition:realistic datasets and a strong baseline. *arXiv preprint arXiv:2005.11909*, 2020. 7

[19] Jian Jia, Xiaotang Chen, and kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *ICCV*, pages 962–971, 2021. 1, 3, 7

[20] Jian Jia, Naiyu Gao, Fei He, Xiaotang Chen, and Kaiqi Huang. Learning disentangled attribute representations for robust pedestrian attribute recognition. In *AAAI*, pages 1069–1077, 2022. 6, 7

[21] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 3

[22] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *TACL*, 8:423–438, 2020. 2

[23] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, pages 111–115. IEEE, 2015. 1

[24] 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. 6

[25] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *ICME*, pages 1–6, 2018. 2, 7

[26] 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. 6

[27] Qiaozhe Li, Xin Zhao, Ran He, and Kaiqi Huang. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation. In *IJCAI*, pages 833–839, 2019. 7

[28] Wanhua Li, Zhexuan Cao, Jianjiang Feng, and Jiwen Lu. Label2label: A language modeling framework for multi-attribute learning. In *ECCV*, pages 562–579, 2022. 6, 7

[29] 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. 6

[30] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1, 2020. 2

[31] 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. 2

[32] Hao Tan, Zichang Tan, Dunfang Weng, Ajian Liu, Jun Wan, Zhen Lei, and Stan Z Li. Vision transformer with relation exploration for pedestrian attribute recognition. *IEEE TMM*, 27:198–208, 2024. 1

[33] 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, 2

[34] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z Li. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *AAAI*, pages 12055–12062, 2020. 1, 4

[35] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *ICCV*, pages 4997–5006, 2019. 7

[36] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. In *ICCV*, pages 531–540, 2017. 1

[37] 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. 1, 2, 3, 5, 6, 7

[38] Dunfang Weng, Zichang Tan, Liwei Fang, and Guodong Guo. Exploring attribute localization and correlation for pedestrian attribute recognition. *Neurocomputing*, 531:140–150, 2023. 1, 7

[39] Junyi Wu, Yan Huang, Zhipeng Gao, Yating Hong, Jianqiang Zhao, and Xinsheng Du. Inter-attribute awareness for pedestrian attribute recognition. *PR*, 131:108865, 2022. 4, 7

[40] Junyi Wu, Yan Huang, Min Gao, Zhipeng Gao, Jianqiang Zhao, Jieming Shi, and Anguo Zhang. Exponential information bottleneck theory against intra-attribute variations for pedestrian attribute recognition. *IEEE TIFS*, pages 5623–5635, 2023. 2, 6, 7

[41] Junyi Wu, Yan Huang, Min Gao, Zhipeng Gao, Jianqiang Zhao, Huiji Zhang, and Anguo Zhang. A two-stream hybrid convolution-transformer network architecture for clothing-change person re-identification. *IEEE TMM*, 2023. 6

[42] 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, 6, 7

[43] Junyi Wu, Yan Huang, Min Gao, Yuzhen Niu, Yuzhong Chen, Qiang Wu, and Jianqiang Zhao. Learning comprehensive representation via selective activation and dual-level orthogonality for pedestrian attribute recognition. *IEEE TCSVT*, 2025. 1

[44] Huangbiao Xu, Xiao Ke, Yuezhou Li, Rui Xu, Huanqi Wu, Xiaofeng Lin, and Wenzhong Guo. Vision-language action knowledge learning for semantic-aware action quality assessment. In *ECCV*, pages 423–440, 2024. 3

[45] Luwei Yang, Ligen Zhu, Yichen Wei, Shuang Liang, and Ping Tan. Attribute recognition from adaptive parts. In *BMCV*, pages 81.1–81.11, 2016. 1, 2

[46] 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, 7

[47] Jiajun Zhang, Pengyuan Ren, and Jianmin Li. Deep template matching for pedestrian attribute recognition with the auxiliary supervision of attribute-wise keypoints. *arXiv preprint arXiv:2011.06798*, 2020. 2

[48] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, pages 2153–2162, 2023. 1

[49] Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *ACM MM*, pages 788–796, 2021. 1

[50] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In *IJCAI*, pages 3177–3183, 2018. 1, 7

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

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3

[53] Yibo Zhou, Hai-Miao Hu, Jinzuo Yu, Zhenbo Xu, Weiqing Lu, and Yuran Cao. A solution to co-occurrence bias: Attributes disentanglement via mutual information minimization for pedestrian attribute recognition. In *IJCAI*, pages 1831–1839, 2023. 7