Joint Implicit and Explicit Language Learning for Pedestrian Attribute Recognition

Yukang Zhang[1, 2], **Lei Tan**[3], **Yang Lu**[1, 2], **Yan Yan**[1, 2], **Hanzi Wang**[1, 2*]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, China.
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, China.
[3]Centre for Trusted Internet and Community, National University of Singapore, Singapore.
zhangyk@stu.xmu.edu.cn, lei.tan@nus.edu.sg, {luyang, yanyan, hanzi.wang}@xmu.edu.cn

Abstract

Pedestrian attribute recognition (PAR) has received increasing attention due to its wide application in video surveillance and pedestrian analysis. Some text-enhanced methods tackle this task by facilitating interactive learning between the attributes and the visual images. However, these generic languages fail to uniquely describe different pedestrian images, missing individual characteristics. In this paper, we propose a Joint Implicit and Explicit Language Guidance Enhancement Learning (JGEL) method, which converts each pedestrian image into a language description with dual language learning to effectively learn enhanced attribute information. Specifically, we first propose an Implicit Language Guidance Learning (ILGL) stream. It projects visual image features into the text embedding space to generate pseudo-word tokens, implicitly modeling image attributes and providing personalized descriptions. Moreover, we propose an Explicit Attribute Enhancement Learning (EAEL) stream to guide the generated pseudo-word tokens obtained by ILGL explicitly aligned with pedestrian attributes, which can effectively align the pseudo-word tokens with the attribute concepts in the text embedding space. Extensive experiments show that JGEL has significant advantages in improving the performance of PAR and the challenging zero-shot PAR task.

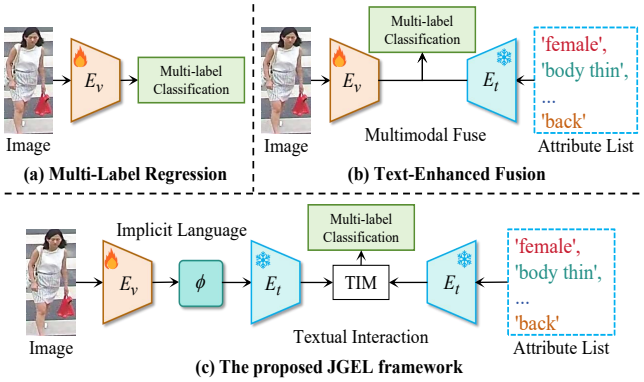

Figure 1: Comparison between existing methods and JGEL. (a) Multi-label regression-based methods treat attributes as labels to to unidirectionally optimize the features. (b) Text-enhanced methods aim to facilitate interactive learning between the attributes and visual images. (c) JGEL converts each pedestrian image into a language description to implicitly model the attributes of the images and provide personalized attribute descriptions for pedestrian images.

Introduction

Pedestrian Attribute Recognition (PAR) plays a crucial role in safety monitoring and intelligent transportation systems (Yuan et al. 2023). Given a pedestrian image, the goal of PAR is to extract a predefined set of attribute information for providing a detailed description for the given pedestrian (Han et al. 2019; Cheng et al. 2022). Thanks to the continuous development in deep learning technologies (Zhang et al. 2021; Zhang and Wang 2023; Zhang et al. 2023, 2025b), the performance of PAR has seen substantial enhancements. However, in practical applications, the PAR task still faces some challenges, such as complex backgrounds, changes in lighting and posture, which make the PAR task more challenging (Wang et al. 2017; Zhao et al. 2018; Zhang et al. 2024; Tan et al. 2024; Zheng et al. 2024).

Early PAR methods (Zeng et al. 2020; Feng et al. 2019) primarily focused on extracting visual features from images to classify each attribute (Gong, Huang, and Chen 2022; Yang, Chen, and Ye 2024; Yao et al. 2025). However, as illustrated in Fig. 1 (a), these early methods are essentially unidirectional regression-based methods from image to multi-label attribute. They regard the image as the sole source of information and simply mine the features that can be used for attribute recognition from the image, ignoring the rich semantic information contained in the attribute labels (Wang et al. 2022, 2024b; Gong et al. 2024).

Recently, the emergence of Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) has unlocked new possibilities for cross-modal interaction learning between visual and textual modalities. Some recent PAR methods (Wang et al. 2025a; Zhang et al. 2025a; Zhu et al. 2023; Wang et al. 2024b, 2025b) have explored the integration of textual information by converting attribute labels into textual descriptions. These descriptions are then fed into a text encoder to extract attribute-specific text features, as illustrated in Fig. 1 (b). These text features are subsequently combined with visual features for interactive learning. How-

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ever, these methods provide only a generic description of a category rather than a personalized representation for each pedestrian image. This lack of granularity can lead to confusion between pedestrians who share similar category-level attributes but differ in fine-grained details, ultimately compromising the accuracy of the PAR model.

In this paper, we propose a novel Joint Implicit and Explicit Language Guidance Enhancement Learning (JGEL) method, which enhances attribute information by integrating both implicit and explicit language guidance, as illustrated in Fig. 1 (c). Specifically, JGEL consists of two key components: Implicit Language Guidance Learning (ILGL) and Explicit Attribute Enhancement Learning (EAEL). ILGL generates pseudo-word tokens by projecting the visual features into the text embedding space, thereby implicitly modeling the attributes and providing personalized textual descriptions for pedestrian images. Then, a Multimodal Interactive Module (MIM) is proposed to facilitate interactive learning between implicit textual features and visual features. The proposed EAEL aims to guide the pseudo-word tokens generated by ILGL to be explicitly aligned with attributes, which can effectively ensure that the pseudo-word tokens are associated with the attributes in the text embedding space. Moreover, a Textual Interactive Module (TIM) is proposed to guide the implicit textual embeddings obtained by ILGL to attend to regions that are semantically faithful to the attributes. The main contributions of this paper can be summarized as follows:

• We propose a novel JGEL method, which innovatively combines implicit and explicit language guidance to harness the power of dual language learning, thereby effectively enhancing attribute information for the PAR task.

• We propose an ILGL stream to project the visual image features into the text embedding space to implicitly model the attributes of the images and provide personalized textual descriptions for pedestrian images.

• We propose an EAEL stream to guide the pseudo-word tokens generated by ILGL to be effectively associated with the attribute concepts in the text embedding space, which ensures that the generated tokens are properly aligned with the relevant attributes, enhancing the reliability and accuracy of the attribute recognition process.

• Extensive experiments show that JGEL has significant advantages in improving the performance of the PAR task. Furthermore, we also demonstrate the effectiveness and generalization of JGEL in the challenging zero-shot PAR task.

Related Work

Pedestrian Attribute Recognition

In recent years, various PAR methods (Welling and Kipf 2017; Nguyen et al. 2021; Xiang et al. 2019) have been proposed to learn feature representations corresponding to attributes, which can be broadly categorized into the multi-label regression methods and text-enhanced methods.

The multi-label regression methods (Wang et al. 2022; Guo et al. 2019; Li et al. 2018a; Tan et al. 2020) primarily rely on CNNs or Transformers to extract visual features to predict the probability of each attribute. For instance, Chen *et al.* (Chen et al. 2021) propose to quantify attribute contributions and visualize the discriminative attributes. Zeng *et al.* (Zeng et al. 2020) propose a co-attentive sharing module to share features across multiple tasks by extracting relevant channels and spatial regions. However, these methods treat the image as the sole source of information. This limitation may lead to suboptimal performance in complex scenarios (Yang et al. 2023b,a).

To address these limitations, some recent text-enhanced methods (Cheng et al. 2022; Wang et al. 2025a; Zhu et al. 2023; Wang et al. 2024b; Hu, Yang, and Ye 2024) have explored converting attribute labels into textual descriptions and using them as inputs to a text encoder. VTB (Cheng et al. 2022) is first proposed to introduce transformer into the PAR task for fusing visual text inputs. PromptPAR (Wang et al. 2025a) formulates PAR as a vision-language fusion problem and expands the attribute phrases into sentences to exploit the relations between pedestrian images and attribute labels. While these methods incorporate textual information, they still face challenges. The text features they generate are based on attribute categories, providing only generic descriptions rather than personalized representations for each pedestrian image. This limitation hinders the model's ability to capture the unique details of individual pedestrians.

Vision-language Pretraining

Vision-language Pre-training (VLP) aims to align the vision with its text description. Recently, VLP has significantly improved the performance of many downstream tasks by learning the semantic correspondence between vision and text through large-scale image-text datasets (Li, Sun, and Li 2023). CLIP (Radford et al. 2021) utilizes text-image pairs for training, enabling the model to understand the correspondence between text and images through contrastive learning. SimVLM (Wang et al. 2021) leverages weakly supervised data for pre-training and exhibits text-guided zero-shot generalization capabilities in a cross-modal setting. ALBEF (Li et al. 2021) aligns visual and textual features before fusing them into a multimodal transformer. In this paper, we generate pseudo-word tokens by projecting the visual image features into the text embedding space, thereby implicitly modeling the attributes of the images and providing personalized textual descriptions for pedestrian images.

Textual Inversion

Textual inversion aims to discover new pseudo-words in the text embedding space for encapsulating both the visual content and intricate visual details (Gal et al. 2022). Recently, various methods have been proposed to apply it to different tasks. For example, Pic2Word (Saito et al. 2023) transforms input images into language tokens, enabling the composition of image and text queries for zero-shot composed image retrieval tasks. KEDs (Suo et al. 2024) introduces a textual concept alignment training paradigm, ensuring semantic alignment between mapped visual features and rich semantics. In this work, we harness the power of textual inversion to learn rich visual embeddings for the PAR task and ensure that the generated pseudo-word tokens are properly aligned with relevant attributes.

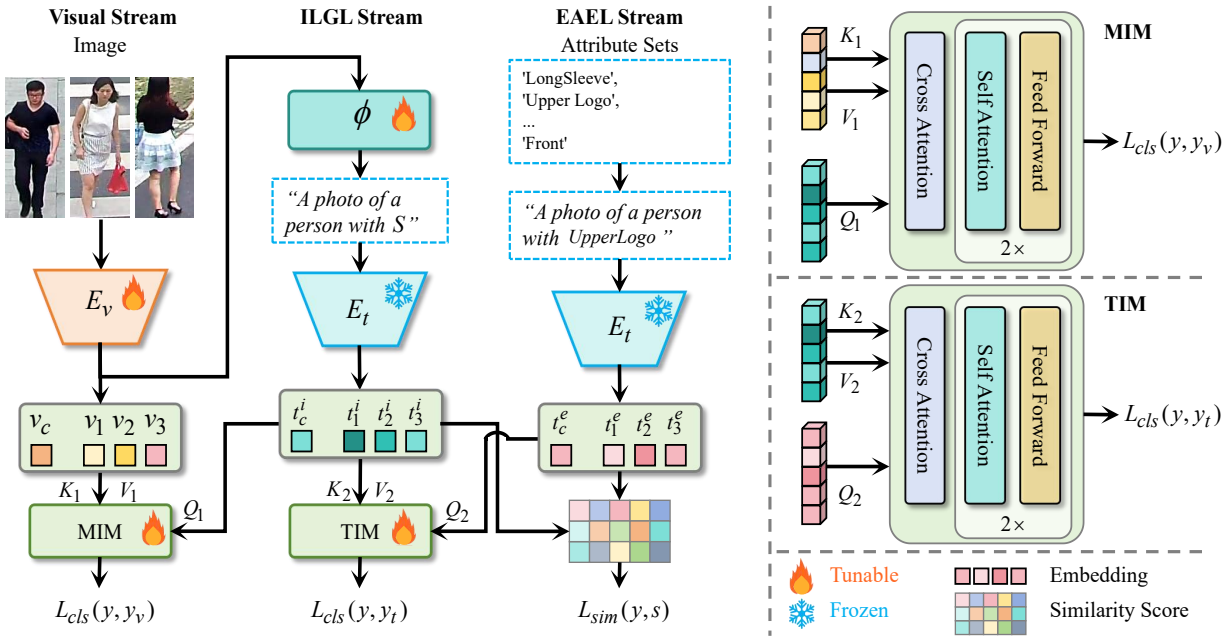

Figure 2: Overview of the proposed JGEL, which consists of a visual stream, an ILGL stream and an EAEL stream. The ILGL stream is used to learn pseudo-word tokens $\boldsymbol{S}^*$ by projecting the visual embeddings into the text embedding space. Then, a MIM is proposed to facilitate interactive learning between the implicit textual features and the visual features. The EAEL stream aims to guide the generated pseudo-word tokens by ILGL to be explicitly aligned with the attributes. Additionally, a TIM is proposed to guide the implicit textual embeddings obtained by ILGL to attend to regions that are semantically faithful to the attributes.

Methodology

Preliminaries

Contrastive Language-Image Pre-training (CLIP) is a state-of-the-art model that facilitates the retrieval and classification tasks by aligning visual and textual representations in a shared embedding space. CLIP is composed of two primary components: a visual encoder E_v and a text encoder E_t, both of which are trained on a vast dataset of image-text pairs using a contrastive learning objective. The visual encoder E_v processes an input image I to extract a visual feature vector $i = E_v(I) \in \mathbb{R}^d$, while the text encoder E_t takes a text caption T and produces a textual feature vector $t = E_t(T) \in \mathbb{R}^d$. The alignment of these features is achieved through a symmetric contrastive loss, which is calculated after normalizing the image and text features.

A simple approach to applying CLIP to the PAR task is to input attribute descriptions into the text encoder. However, since this method is based on the number of attributes given for the entire dataset, it fails to generate personalized attribute descriptions for each pedestrian image.

Overview

An overview of the proposed JGEL is depicted in Fig. 2. Specifically, starting with the visual embeddings derived from CLIP's visual encoder, we first propose an ILGL stream which employs an inversion network to project the visual image features into the text embedding space and generate pseudo-word tokens, which can implicitly model the attributes of the images and provide personalized attribute descriptions for pedestrian images. Then, a Multimodal Interactive Module (MIM) is proposed to facilitate interactive learning between the implicit textual features and the visual features. Moreover, we propose an EAEL stream to guide the generated pseudo-word tokens by ILGL to be explicitly aligned with pedestrian attributes. A Textual Interactive Module (TIM) is proposed to guide the implicit textual embeddings obtained by ILGL to attend to regions that are semantically faithful to the attributes. Note that the text encoder is frozen in the proposed JGEL.

Implicit Language Guidance Learning

Prior research has suggested that the word-embedding space has sufficient expressiveness to encapsulate basic image concepts (Cohen et al. 2022; Yang et al. 2024). However, existing methods (Cheng et al. 2022; Wang et al. 2025a) have inherent limitations due to their pre-defined prompts. These prompts are typically based on a fixed number of attributes provided by the dataset, which may fail to fully encapsulate the diverse visual context for each individual pedestrian image. Therefore, the descriptions generated by these methods may lack the necessary specificity to accurately reflect the unique characteristics of different images. In contrast, we propose to learn the pseudo-word token by textual inversion technique (Saito et al. 2023) to align with the context of the given pedestrian image. ILGL directly learns the pseudo-word token from the image itself, rather than relying on pre-defined prompts, thereby allowing for a more nuanced and context-aware representation of pedestrian attributes.

As illustrated in Fig. 2, let ϕ denote the textual inversion network, which is a lightweight random-initialized model employing a two-layered MLP of 512-dimensional hidden state. ILGL aims to invert the visual embeddings $\boldsymbol{v}$ from the visual space of CLIP into a pseudo-word tokens $\boldsymbol{S}^* \in T^*$ by $\phi(\boldsymbol{v}) = \boldsymbol{S}^*$, where T^* indicates the token embedding space. Then, the attribute prompts for the input images are structured as 'A photo of a person with $\boldsymbol{S}^*$'. The attribute prompts with a tokenization process are fed into the text encoder of CLIP to obtain implicit text embedding $\boldsymbol{t}^i$.

Considering that the textual embeddings obtained by the text encoder incorporate diverse semantics, the image patches obtained by the visual encoder corresponding to such semantics inherently possess substantial influence conducive to discrimination. Therefore, we propose a Multimodal Interactive Module (MIM) to facilitate interactive learning between the implicit text embeddings $\boldsymbol{t}^i$ and the visual embeddings $\boldsymbol{v}$, as illustrated in Fig. 2. The MIM consists of a multi-head cross attention (MCA) layer and 2-layer transformer blocks. The implicit text embedding and visual embedding obtained by the text encoder E_t and the vision encoder E_v are fed into the proposed MIM to fuse visual and implicit text embeddings. Specifically, we project the implicit text embeddings $\boldsymbol{t}^i$ into a query matrix Q_1, and project the visual embeddings into a key matrix K_1 and a value matrix V_1 with three different linear-projection layers. Then, the full interactive learning between image and implicit text embeddings can be achieved by:

$$\boldsymbol{v}_+ = Transformer(MCA(LN(Q_1, K_1, V_1))), \quad (1)$$

where $\boldsymbol{v}_+ \in \mathbb{R}^{B\times d}$, B is the batch-size in a mini batch and d is the embedding dimension. $LN(\cdot)$ is Layer Normalization, $Transformer(\cdot)$ denotes transformer blocks, and $MCA(\cdot)$ is the multi-head cross attention and can be achieved by:

$$MCA(Q_1, K_1, V_1) = Softmax\left(\frac{Q_1 K_1^T}{\sqrt{d}}\right) V_1, \quad (2)$$

where d is the embedding dimension.

Then, the obtained embeddings $\boldsymbol{v}_+$ are fed into a classifier layer to obtain predicted attribute classes $\boldsymbol{y}_v$. The binary cross entropy (BCE) loss is used to optimize the attribute classes $\boldsymbol{y}_v$ predicted by the classifier layer of the proposed ILGL, which can be mathematically expressed as follows:

$$\begin{aligned}\mathcal{L}_{cla}(\boldsymbol{y}, \boldsymbol{y}_v) = -\frac{1}{B}\sum_{i=1}^{B}\Big[\boldsymbol{y}^i \cdot \log(\sigma(\boldsymbol{y}_v^i)) \\ +(1-\boldsymbol{y}^i)\cdot\log(1-\sigma(\boldsymbol{y}_v^i))\Big],\end{aligned} \quad (3)$$

where $\boldsymbol{y}$ is the attribute label, and $\sigma(\boldsymbol{y}_v)$ is the sigmoid activation function to convert $\boldsymbol{y}_v$ into a probability value.

Explicit Attribute Enhancement Learning

Although the proposed ILGL can generate pseudo-word tokens to provide personalized attribute descriptions for each pedestrian image, it has a limitation: these pseudo-word tokens are not aligned with the textual concepts in the text embedding space, which may introduce challenges during inference. To address this issue, we propose an EAEL stream to guide the generated pseudo-word token aligned with pedestrian attributes.

Specifically, for the pedestrian attributes $A = \{A_1, A_2, \ldots, A_N\}$ in the PAR dataset, where N is the number of attributes, we first translate each attribute into a language description by following PromptPAR (Wang et al. 2025a). For example, the attribute 'upper jacket' is transformed into the language description 'A photo of a person with upper jacket'. Then, this description undergoes a tokenization process and is then input into the text encoder of CLIP to obtain the explicit text embedding $\boldsymbol{t}^e$.

Then, we propose a Textual Interactive Module (TIM) to guide the implicit text embedding $\boldsymbol{t}^i$ with the obtained explicit text embedding $\boldsymbol{t}^e$ to attend to regions that are semantically faithful to the attributes. Specifically, as illustrated in Fig. 2, similar to MIM, we project the implicit text embedding $\boldsymbol{t}^i$ into a query matrix Q_2, and project the explicit text embedding into a key matrix K_2 and a value matrix V_2 with three different linear-projection layers. Then, the full enhancement learning between implicit and explicit text embeddings can be achieved by:

$$\boldsymbol{t}_+^i = Transformer(MCA(LN(Q_2, K_2, V_2))), \quad (4)$$

where $\boldsymbol{t}_+^i \in \mathbb{R}^{B\times N\times d}$, B is the batch-size in a mini batch, N is the attribute number in dataset and d is the embedding dimension. In this way, the proposed TIM can effectively aggregates the attention map to highlight the regions of high attribute response.

Then, the obtained embeddings $\boldsymbol{t}_+^i$ are fed into N classifier layer to obtain predicted attribute classes $\boldsymbol{y}_t$. The BCE loss is used to optimize the attribute classes $\boldsymbol{y}_t$ predicted by the classifier layer of the proposed EAEL, which can be mathematically expressed as follows:

$$\begin{aligned}\mathcal{L}_{cla}(\boldsymbol{y}, \boldsymbol{y}_t) = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{N}\sum_{j=1}^{N}\Big[\boldsymbol{y}^i \cdot \log(\sigma(\boldsymbol{y}_t^{i,j})) \\ +(1-\boldsymbol{y}^i)\cdot\log(1-\sigma(\boldsymbol{y}_t^{i,j}))\Big].\end{aligned} \quad (5)$$

In addition, to fully align the implicit text embedding $\boldsymbol{t}^i$ obtained by the pseudo-word tokens in ILGL with explicit text embedding $\boldsymbol{t}^e$, we calculate the similarity scores $\boldsymbol{s} \in \mathbb{R}^{B\times N}$ between those two embedding vectors. Then, the BCE loss is used to optimize similarity scores $\boldsymbol{s}$, which can be mathematically expressed as follows:

$$\begin{aligned}\mathcal{L}_{sim}(\boldsymbol{y}, \boldsymbol{s}) = -\frac{1}{B}\sum_{j=1}^{B}\big[\boldsymbol{y}_j \cdot \log(\sigma(\boldsymbol{s}_j)) \\ +(1-\boldsymbol{s}_j)\cdot\log(1-\sigma(\boldsymbol{s}_j))\big],\end{aligned} \quad (6)$$

where $\boldsymbol{s}_j = \frac{\boldsymbol{t}_j^i \cdot \boldsymbol{t}_j^e}{\|\boldsymbol{t}_j^i\|_2 \|\boldsymbol{t}_j^e\|_2}$ denotes the cosine similarity between two vectors $\boldsymbol{t}_j^i$ and $\boldsymbol{t}_j^e$.

Optimization

Training optimization. In summary, the overall objective function for the proposed JGEL method can be mathematically expressed as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{cla}(\boldsymbol{y}, \boldsymbol{y}_v) + \mathcal{L}_{sim}(\boldsymbol{y}, \boldsymbol{y}_t) + \mathcal{L}_{sim}(\boldsymbol{y}, \boldsymbol{s}). \quad (7)$$

Methods	PETA				PA100K				RAPv1				RAPv2			
	mA	Accu	Recall	F1	mA	Accu	Recall	F1	mA	Accu	Recall	F1	mA	Accu	Recall	F1
AttExpIB-Net(Wu et al. 2023)	85.90	77.58	86.36	85.32	83.23	79.42	88.60	87.23	82.46	68.81	81.63	80.25	80.60	67.31	80.38	79.15
EALC (Weng et al. 2023)	85.94	80.58	87.38	87.44	80.52	80.13	88.59	87.88	82.09	69.30	82.77	81.17	-	-	-	-
CAS (Yang et al. 2021)	86.40	79.93	87.33	87.18	82.86	79.64	87.79	85.18	84.18	68.59	83.81	80.56	-	-	-	-
VAC (Guo, Fan, and Wang 2022)	-	-	-	-	82.19	80.66	88.10	88.41	81.30	70.12	81.51	**81.54**	79.23	64.51	79.43	77.10
DAFL (Jia et al. 2022)	87.07	78.88	87.03	86.40	83.54	80.13	89.19	88.09	83.72	68.18	83.39	80.29	81.04	66.70	82.07	79.13
CGCN (Fan et al. 2020)	87.08	79.30	89.38	86.59	-	-	-	-	84.70	54.40	83.68	70.49	-	-	-	-
SOFAFormer (Wu et al. 2024)	87.10	81.10	88.40	87.80	83.40	81.10	89.00	88.30	83.40	70.00	83.00	81.20	81.90	**68.60**	83.10	80.20
PARFormer (Fan et al. 2023)	88.65	82.34	91.55	88.66	83.95	80.26	91.07	87.69	83.84	69.70	87.81	81.16	-	-	-	-
SSPNet (Shen et al. 2024)	88.37	82.80	90.55	**89.50**	83.58	80.63	89.32	88.55	83.24	70.21	82.90	81.50	-	-	-	-
DRFormer (Tang and Huang 2022)	89.96	81.30	91.08	88.30	82.47	80.27	88.49	88.04	81.81	**70.60**	80.12	81.42	-	-	-	-
VTB* (Cheng et al. 2022)	85.31	79.60	87.17	86.71	83.72	80.89	89.30	88.21	82.67	69.44	84.39	80.84	81.34	67.48	83.32	79.35
PromptPAR* (Wang et al. 2025a)	87.09	80.19	88.66	86.89	85.47	81.33	90.53	88.55	84.48	69.97	85.87	81.23	81.67	68.00	85.02	**79.85**
JGEL (Ours)	**90.27**	**83.23**	**93.64**	89.22	**87.61**	**82.03**	**94.91**	**88.92**	**86.09**	70.23	**90.64**	81.32	**85.54**	68.24	**91.73**	79.79

Table 1: Comparisons with state-of-the-art methods on four PAR datasets. The best and the second best results are represented by bold font and underline, respectively. $*$ indicates that the method also employs language to assist in the PAR task.

Experiments

Datasets and Evaluation Metrics

We evaluate JGEL on four widely used datasets (including PETA (Deng et al. 2014), PA100K (Liu et al. 2017), RAPv1 (Li et al. 2016) and RAPv2 (Li et al. 2018b)) for the PAR task, and two zero-shot datasets (including PETA-ZS (Jia et al. 2021) and RAP-ZS (Jia et al. 2021)) for the zero-shot PAR task, and adopt the standard data settings as (Jia, Chen, and Huang 2021; Wu et al. 2024). Details about these datasets and the implementation details of the proposed method can be found in the supplementary material.

Comparison with State-of-the-art Methods

To verify the superiority of JGEL, we compare it with several state-of-the-art methods. The quantitative results on four PAR datasets are shown in Tab. 1.

It can be observed that JGEL achieves the best performance in most evaluation metrics on the four datasets. Specifically, in mA, JGEL outperforms the second-best performance by 0.31% (DRFormer (Tang and Huang 2022)) on PETA, 2.14% (PromptPAR* (Wang et al. 2025a)) on PA100k, 1.61% (PromptPAR* (Wang et al. 2025a)) on RAPv1 and 3.64% on RAPv2, respectively. It can be noticed that the PromptPAR* also utilizes language to assist in the PAR task. Under a fair comparison (using the same backbone ViT-B/16), JGEL surpasses PromptPAR* in all four evaluation metrics on four datasets. For example, JGEL respectively surpasses PromptPAR by 2.18%, 3.03%, 4.08%, and 0.95% in the four evaluation metrics on the PETA dataset even though it uses a larger image input scale (224 $\times$ 224). The consistent improvements across various PAR datasets and evaluation metrics demonstrate the effectiveness of JGEL for the PAR task. Moreover, JGEL can maintain good performance on four different datasets, indicating it has a certain degree of universality and stability.

Generalization Analysis on Zero-Shot PAR

To evaluate the generalization performance of JGEL, we conduct experiments on the PETA-ZS and RAP-ZS datasets.

Data.	Methods	mA	Accu	Recall	F1
PETA-ZS	VAC (Guo, Fan, and Wang 2022)	71.91	57.72	70.64	70.90
	MCFL (Chen et al. 2022)	72.91	57.04	74.35	71.29
	ALM (Tang et al. 2019)	73.01	57.78	73.69	71.53
	JLAC (Tan et al. 2020)	73.60	58.66	72.41	72.05
	SOFAFormer (Wu et al. 2024)	74.70	62.10	75.10	74.60
	VTB* (Cheng et al. 2022)	75.13	60.50	74.40	73.38
	PromptPAR* (Wang et al. 2025a)	78.37	63.77	77.72	76.05
	JGEL	**80.56**	**64.98**	**86.39**	**76.75**
RAP-ZS	VAC (Guo, Fan, and Wang 2022)	73.70	63.25	76.97	76.12
	MCFL (Chen et al. 2022)	74.37	63.37	83.86	77.02
	ALM (Tang et al. 2019)	74.28	63.22	80.73	76.65
	JLAC (Tan et al. 2020)	76.38	62.58	79.20	76.05
	SOFAFormer (Wu et al. 2024)	73.90	66.30	79.40	78.40
	VTB* (Cheng et al. 2022)	75.76	64.73	80.85	77.35
	PromptPAR* (Wang et al. 2025a)	78.35	**68.08**	84.08	78.74
	JGEL (Ours)	**80.10**	67.96	**90.15**	**79.73**

Table 2: Comparison with several state-of-the-art methods on the PETA-ZS and RAP-ZS datasets.

As we can see from Tab. 2, in both the PETA-ZS and RAP-ZS datasets, JGEL demonstrates remarkable superiority across most evaluation metrics. For instance, on the PETA-ZS dataset, JGEL achieves the best values in mA (80.56%), Accu (64.98%), Recall (86.39%), and F1 (76.75) compared to existing methods. Similarly, on the RAP-ZS dataset, JGEL also shows excellent performance with best results in mA and Recall, and competitive results in other metrics as well. These results indicate that JGEL has a strong ability to generalize and adapt to new data in the zero-shot scenario, effectively extracting relevant information to recognize pedestrian attributes accurately.

Ablation Studies

Ablation studies for the effectiveness of different components. To demonstrate each component's contribution in JGEL, we conduct ablation studies by evaluating results of different components on the PETA and RAPv1 datasets. As we can see in Tab. 3, the results show that: (1) Comparing

	Components				PETA		RAPv1	
	ILGL	MIM	EAEL	TIM	mA	Recall	mA	Recall
a)	✗	✗	✗	✗	86.84	88.72	81.25	81.31
b)	✓	✗	✗	✗	88.94	91.67	84.88	87.65
c)	✓	✓	✗	✗	89.56	92.36	85.44	88.38
d)	✗	✗	✓	✗	87.14	89.06	84.57	85.44
e)	✗	✓	✓	✗	88.31	90.04	84.83	86.90
f)	✓	✓	✓	✗	89.90	93.20	85.88	89.96
i)	✓	✓	✓	✓	90.27	93.64	86.09	90.64

Table 3: Ablation studies for the proposed JGEL method on the PETA and RAPv1 datasets.

Components	PETA		RAPv1	
	mA	Recall	mA	Recall
ILGL & w/o MIM	88.94	91.67	84.88	87.65
+1 Cross Attention	89.26	91.97	85.23	88.12
+1 Cross Attention & 1 Self Attention	89.39	92.12	85.37	88.29
+1 Cross Attention & 2 Self Attention	89.56	92.36	85.44	88.38
+1 Cross Attention & 4 Self Attention	89.68	92.20	85.55	87.64
ILGL + & EAEL & w/o TIM	89.90	93.20	85.88	89.96
+1 Cross Attention	90.05	93.43	85.96	90.22
+1 Cross Attention & 1 Self Attention	90.18	93.55	86.05	90.53
+1 Cross Attention & 2 Self Attention	90.27	93.64	86.09	90.64
+1 Cross Attention & 4 Self Attention	90.33	93.19	86.12	90.02

Table 4: Ablation studies for the proposed MIM and TIM.

case (a) (baseline) with case (b) and case (c), in the PETA dataset, the mA increases from 86.84% to 89.56%, and the Recall increases from 88.72% to 92.36%, which shows that the ILGL plays a significant role in improving the mA and Recall of the model by generating pseudo-word tokens to provide personalized attribute descriptions. (2) Comparing case (a) with case (d) and case (e) (in which the text embeddings are fused with visual image embeddings), the mA increases from 86.84% to 88.31%, and the Recall increases from 88.72% to 90.04%, which indicates that the EAEL can also improve the performance of the model. (3) Comparing case (c) with case (f), the mA increases from 89.56% to 89.90%, and the Recall increases from 92.36% to 93.20%, which shows that EAEL can effectively align the generated pseudo-word tokens aligned with pedestrian attributes in the text embedding space. (4) By integrating the above components into an end-to-end joint implicit and explicit language guidance enhancement learning framework, JGEL can effectively improve the performance of the PAR task, achieving competitive results.

The effectiveness of MIM and TIM. To validate the effectiveness of MIM and TIM in JGEL, we conduct comprehensive ablation studies by evaluating the experimental results corresponding to varying numbers of cross-attention and self-attention layers on the PETA and RAPv1 datasets, as shown in Tab. 4. With the number of Self Attention layers increases, the results initially show an upward trend, but subsequently, they begin to fluctuate and decline. This observation implies that while the addition of Self Attention layers can enhance the model's performance, an excessive number of such layers may lead to overfitting, thereby adversely affecting the Recall. Concurrently, it is important to note that an increase in the number of Self Attention layers also demands more memory space for parameter storage. Consequently, to strike an optimal balance between performance and computational resource consumption, we have determined that employing 1 cross-attention layer and 2 self-attention layers for both MIM and TIM within the proposed JGEL represents the most favorable configuration.

Methods	RAPv1			
	TT	Params	mA	Recall
SNN-PAR (Wang et al. 2024a)	45.92 h	583.40 M	75.4	78.3
PARFormer-L (Fan et al. 2023)	63.33 h	755.60 M	84.1	88.2
VTB (Cheng et al. 2022)	0.86 h	157.54 M	82.67	84.39
PromptPAR (Wang et al. 2025a)	15.88 h	2.30 M	84.48	85.87
JGEL	2.83 h	95.83 M	86.1	90.6

Table 5: Comparison of training times and the number of parameters under the same backbone (ViT-B/16) on the RAPv1 dataset. TT means the total training time, and Params denotes the number of learnable parameters in the whole model. All models are evaluated on a single 2080Ti GPU.

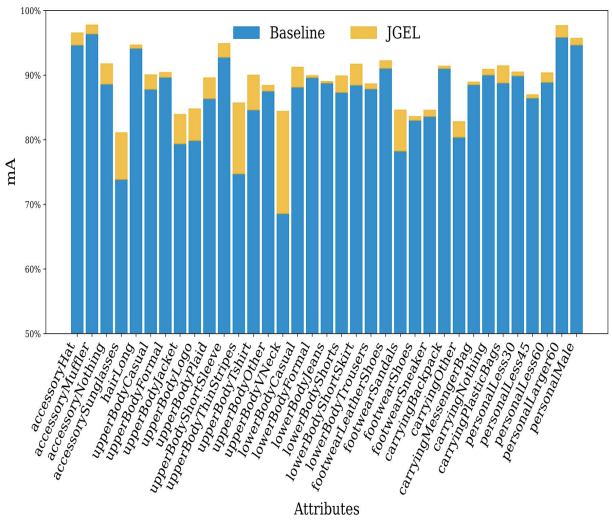

Figure 3: Attribute-wise mA comparison between the baseline (blue box) and JGEL (yellow box) on the PETA dataset.

Comparison of training efficiency. In order to evaluate the efficiency of JGEL, we carry out a comparative analysis between JGEL and some other methods, as shown in Tab. 5. Overall, the proposed JGEL exhibits outstanding comprehensive advantages in training time, the number of model parameters, and performance metrics. Compared with VTB, JGEL not only ensures relatively high accuracy and recall rates, but the training time is not overly long, and the number of parameters is significantly reduced. Compared with PromptPAR, JGEL has a significant improvement in performance (both mA and Recall are higher), while the training time is greatly shortened, and it also makes more rational use of resources in terms of parameter scale to achieve high performance. These advantages enable JGEL to stand out in

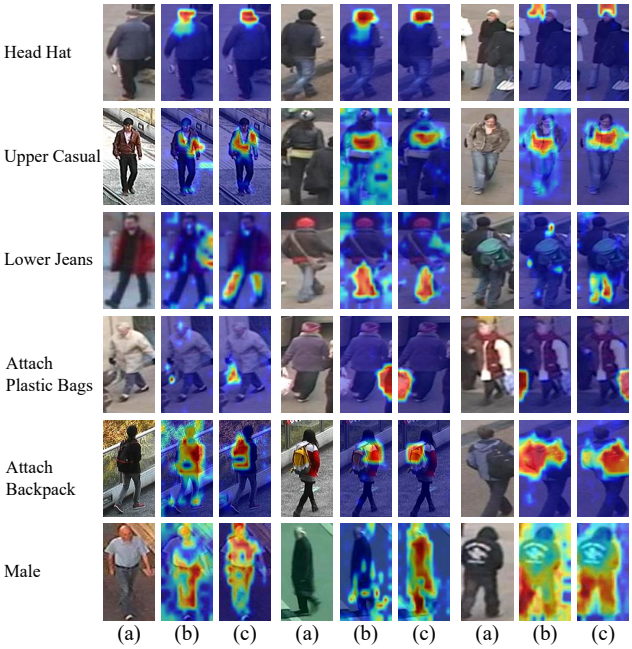

Figure 4: Visualization of specific attribute response maps using only the image encoder and with the guidance of ILGL on the PETA dataset. (a) are the input images, (b) are the results obtained using only the image encoder, and (c) are the results obtained with the guidance of ILGL.

the PAR task and are more suitable for a variety of practical application scenarios.

Visualization Analysis

Attribute-wise mA comparison. To evaluate the effectiveness of the JGEL, we compare the attribute-wise mA value obtained by the baseline and JGEL on the PETA dataset, which is shown in Fig. 3. Generally speaking, the proposed JGEL performs better than the baseline in all the attributes, which indicates that JGEL method has certain advantages in the PAR task. For example, in accessory attributes, as well as various upper body clothing, and lower body clothing, the recognition accuracy of JGEL is higher than that of the baseline, indicating that JGEL can more effectively recognize these attributes. This may be because JGEL has a better ability to capture and understand the texture and style of the upper body clothing in the model structure or feature learning process. Whether it is the detailed features of clothing or the clues related to personal characteristics, they can be better learned and utilized, thus improving the recognition performance in various attributes.

Specific-attribute attention heat maps. To further demonstrate the effectiveness of the ILGL, we compared the differences between the heatmaps obtained by using only the visual encoder and those obtained with the guidance of ILGL. The results are shown in Fig. 4. It can be seen that the heatmaps obtained by the visual encoder show that the attention areas are relatively scattered and not accurate enough. The highly activated areas may not accurately focus on the

Figure 5: Some failure cases on the PETA dataset. The attributes highlighted in red are the incorrectly recognized results, while those in blue are the unrecognized results.

parts of the image that are closely related to the target attributes. Some of the areas covered by the heatmaps contain a large amount of irrelevant background information. In contrast, the activated areas in the heatmaps guided by ILGL are more concentrated and accurately cover the image areas related to the target attributes, showing that the model can precisely capture the visual clues related to the attribute. This is due to the pseudo-word tokens generated by ILGL can implicitly model the image attributes and provide more targeted guidance for the model, making the model more directional when paying attention to the image content.

Analysis of failure cases. To further analyze JGEL, we visualize some failed cases, as shown in Fig. 5. In terms of incorrectly recognized attributes, for clothing attributes, there are cases such as misrecognizing "upper other" as "upper jacket" and "shoes other" as "shoes sneaker", because the model has insufficient ability to distinguish and classify the styles of tops and shoes. In terms of unrecognized attributes, age attributes such as "age less 30" and "age 30 - 45" are not correctly recognized, because the judgment requires comprehensive information and the model has difficulty in integrating it. Therefore, JGEL still needs to be improved. For clothing attributes, it is necessary to improve the ability to distinguish different styles; for attributes such as age and clothing style, it is necessary to optimize the ability to process comprehensive information.

Conclusion

In this paper, we present a novel JGEL method for the PAR task, which consists of two key components: ILGL and EAEL. ILGL generates pseudo-word tokens by projecting visual features into the text embedding space, enabling implicit modeling of image attributes and providing personalized descriptions for pedestrians. EAEL refines these tokens by explicitly aligning them with pedestrian attributes, ensuring accurate representation of attribute concepts in the text embedding space. Extensive experiments show that the proposed JGEL has significant advantages in improving the performance of the PAR task. Furthermore, we also demonstrate the effectiveness and generalization of JGEL in the challenging zero-shot PAR task.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U25A20531, U21A20514, 62502401, 62376233, 62372388; in part by the China Postdoctoral Science Foundation under Grant 2025M771584 and 2025T180439; in part by the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City under Grant 3502Z20241027 and 3502Z20241029, and in part by the Xiaomi Young Talents Program.

References

Chen, L.; Song, J.; Zhang, X.; and Shang, M. 2022. MCFL: Multi-label Contrastive Focal Loss For Deep Imbalanced Pedestrian Attribute Recognition. *Neural Computing and Applications*, 34(19): 16701–16715.

Chen, X.; Liu, X.; Liu, W.; Zhang, X.-P.; Zhang, Y.; and Mei, T. 2021. Explainable Person Re-identification with Attribute-guided Metric Distillation. In *IEEE ICCV*, 11813–11822.

Cheng, X.; Jia, M.; Wang, Q.; and Zhang, J. 2022. A Simple Visual-Textual Baseline for Pedestrian Attribute Recognition. *IEEE TCSVT*, 32(10): 6994–7004.

Cohen, N.; Gal, R.; Meirom, E. A.; Chechik, G.; and Atzmon, Y. 2022. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *ECCV*, 558–577.

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian Attribute Recognition at Far Distance. In *ACM MM*, 789–792.

Fan, H.; Hu, H.-M.; Liu, S.; Lu, W.; and Pu, S. 2020. Correlation Graph Convolutional Network for Pedestrian Attribute Recognition. *IEEE TMM*, 24: 49–60.

Fan, X.; Zhang, Y.; Lu, Y.; and Wang, H. 2023. Parformer: Transformer-based Multi-task Network for Pedestrian Attribute Recognition. *IEEE TCSVT*, 34(1): 411–423.

Feng, X.; Li, Y.; Du, H.; and Wang, H. 2019. Research on Pedestrian Attribute Recognition Based on Semantic Segmentation in Natural Scene. In *International Conference on Artificial Intelligence and Security*, 498–509.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-image Generation Using Textual Inversion. *Arxiv*.

Gong, Y.; Huang, L.; and Chen, L. 2022. Person Re-identification Method Based on Color Attack and Joint Defence. In *CVPR*, 4313–4322.

Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024. Cross-modality Perturbation Synergy Attack for Person Re-identification. *NeurIPS*, 37: 23352–23377.

Guo, H.; Fan, X.; and Wang, S. 2022. Visual Attention Consistency for Human Attribute Recognition. *IJCV*, 130(4): 1088–1106.

Guo, H.; Zheng, K.; Fan, X.; Yu, H.; and Wang, S. 2019. Visual Attention Consistency Under Image Transforms for Multi-label Image Classification. In *IEEE CVPR*, 729–739.

Han, K.; Wang, Y.; Shu, H.; Liu, C.; Xu, C.; and Xu, C. 2019. Attribute Aware Pooling for Pedestrian Attribute Recognition. In *IJCAI*, 2456–2462.

Hu, Z.; Yang, B.; and Ye, M. 2024. Empowering Visible-Infrared Person Re-Identification with Large Foundation Models. In *NeurIPS*, volume 37, 117363–117387.

Jia, J.; Chen, X.; and Huang, K. 2021. Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recognition. In *IEEE ICCV*, 962–971.

Jia, J.; Gao, N.; He, F.; Chen, X.; and Huang, K. 2022. Learning Disentangled Attribute Representations for Robust Pedestrian Attribute Recognition. In *AAAI*, volume 36, 1069–1077.

Jia, J.; Huang, H.; Chen, X.; and Huang, K. 2021. Rethinking of Pedestrian Attribute Recognition: A Reliable Evaluation under Zero-shot Pedestrian Identity Setting. *Arxiv*.

Li, D.; Chen, X.; Zhang, Z.; and Huang, K. 2018a. Pose Guided Deep Model for Pedestrian Attribute Recognition in Surveillance Scenarios. In *IEEE ICME*, 1–6.

Li, D.; Zhang, Z.; Chen, X.; and Huang, K. 2018b. A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios. *IEEE TIP*, 28(4): 1575–1590.

Li, D.; Zhang, Z.; Chen, X.; Ling, H.; and Huang, K. 2016. A Richly Annotated Dataset for Pedestrian Attribute Recognition. *Arxiv*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. *NeurIPS*, 34: 9694–9705.

Li, S.; Sun, L.; and Li, Q. 2023. CLIP-ReID: Exploiting Vision-Language Model for Image Re-identification Without Concrete Text Labels. In *AAAI*, volume 37, 1405–1413.

Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; and Wang, X. 2017. Hydraplus-net: Attentive Deep Features for Pedestrian Analysis. In *IEEE ICCV*, 350–359.

Nguyen, B. X.; Nguyen, B. D.; Do, T.; Tjiputra, E.; Tran, Q. D.; and Nguyen, A. 2021. Graph-based Person Signature for Person Re-identifications. In *IEEE CVPR*, 3492–3501.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 8748–8763.

Saito, K.; Sohn, K.; Zhang, X.; Li, C.-L.; Lee, C.-Y.; Saenko, K.; and Pfister, T. 2023. Pic2word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval. In *IEEE CVPR*, 19305–19314.

Shen, J.; Guo, T.; Zuo, X.; Fan, H.; and Yang, W. 2024. SSP-Net: Scale and Spatial Priors Guided Generalizable and Interpretable Pedestrian Attribute Recognition. *Pattern Recognition*, 148: 110194.

Suo, Y.; Ma, F.; Zhu, L.; and Yang, Y. 2024. Knowledge-enhanced Dual-stream Zero-shot Composed Image Retrieval. In *IEEE CVPR*, 26951–26962.

Tan, L.; Zhang, Y.; Han, K.; Dai, P.; Zhang, Y.; Wu, Y.; and Ji, R. 2024. RLE: A Unified Perspective of Data Augmentation for Cross-spectral Re-identification. *NeurIPS*, 37: 126977–126996.

Tan, Z.; Yang, Y.; Wan, J.; Guo, G.; and Li, S. Z. 2020. Relation-aware Pedestrian Attribute Recognition with Graph Convolutional Networks. In *AAAI*, volume 34, 12055–12062.

Tang, C.; Sheng, L.; Zhang, Z.; and Hu, X. 2019. Improving Pedestrian Attribute Recognition with Weakly-supervised Multi-scale Attribute-specific Localization. In *IEEE ICCV*, 4997–5006.

Tang, Z.; and Huang, J. 2022. DRFormer: Learning Dual Relations Using Transformer for Pedestrian Attribute Recognition. *Neurocomputing*, 497: 159–169.

Wang, H.; Zhu, Q.; She, M.; Li, Y.; Song, H.; Xu, M.; and Wang, X. 2024a. SNN-PAR: Energy Efficient Pedestrian Attribute Recognition via Spiking Neural Networks. *Arxiv*.

Wang, J.; Zhu, X.; Gong, S.; and Li, W. 2017. Attribute Recognition by Joint Recurrent Learning of Context and Correlation. In *IEEE ICCV*, 531–540.

Wang, X.; Jin, J.; Li, C.; Tang, J.; Zhang, C.; and Wang, W. 2025a. Pedestrian Attribute Recognition via Clip Based Prompt Vision-Language Fusion. *IEEE TCSVT*, 35(1): 148–161.

Wang, X.; Liu, L.; Yang, B.; Ye, M.; Wang, Z.; and Xu, X. 2025b. TokenMatcher: Diverse Tokens Matching for Unsupervised Visible-Infrared Person Re-Identification. In *AAAI*, volume 39, 7934–7942.

Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; and Luo, B. 2022. Pedestrian Attribute Recognition: A Survey. *Pattern Recognition*, 108220.

Wang, X.; Zhu, Q.; Jin, J.; Zhu, J.; Wang, F.; Jiang, B.; Wang, Y.; and Tian, Y. 2024b. Spatio-Temporal Side Tuning Pre-trained Foundation Models for Video-based Pedestrian Attribute Recognition. *Arxiv*.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. Simvlm: Simple Visual Language Model Pretraining with Weak Supervision. *Arxiv*.

Welling, M.; and Kipf, T. N. 2017. Semi-supervised Classification with Graph Convolutional Networks. In *ICLR*, 1–14.

Weng, D.; Tan, Z.; Fang, L.; and Guo, G. 2023. Exploring Attribute Localization and Correlation for Pedestrian Attribute Recognition. *Neurocomputing*, 140–150.

Wu, J.; Huang, Y.; Gao, M.; Gao, Z.; Zhao, J.; Shi, J.; and Zhang, A. 2023. Exponential Information Bottleneck Theory Against Intra-attribute Variations for Pedestrian Attribute Recognition. *IEEE TIFS*, 18: 5623–5635.

Wu, J.; Huang, Y.; Gao, M.; Niu, Y.; Yang, M.; Gao, Z.; and Zhao, J. 2024. Selective and Orthogonal Feature Activation for Pedestrian Attribute Recognition. In *AAAI*, volume 38, 6039–6047.

Xiang, L.; Jin, X.; Ding, G.; Han, J.; and Li, L. 2019. Incremental Few-shot Learning for Pedestrian Attribute Recognition. In *IJCAI*, 3912–3918.

Yang, B.; Chen, J.; Chen, C.; and Ye, M. 2023a. Dual Consistency-Constrained Learning for Unsupervised Visible-Infrared Person Re-Identification. *IEEE TIFS*, 19: 1767–1779.

Yang, B.; Chen, J.; Ma, X.; and Ye, M. 2023b. Translation, Association and Augmentation: Learning Cross-modality Re-identification from Single-modality Annotation. *IEEE TIP*, 32: 5099–5113.

Yang, B.; Chen, J.; and Ye, M. 2024. Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification. In *CVPR*, 16870–16879.

Yang, Y.; Tan, Z.; Tiwari, P.; Pandey, H. M.; Wan, J.; Lei, Z.; Guo, G.; and Li, S. Z. 2021. Cascaded Split-and-Aggregate Learning with Feature Recombination for Pedestrian Attribute Recognition. *IJCV*, 129: 2731–2744.

Yang, Z.; Wu, D.; Wu, C.; Lin, Z.; Gu, J.; and Wang, W. 2024. A Pedestrian is Worth One Prompt: Towards Language Guidance Person Re-Identification. In *IEEE CVPR*, 17343–17353.

Yao, H.; Yang, B.; Huang, W.; Du, B.; and Ye, M. 2025. Unsupervised Visible-Infrared Person Re-identification under Unpaired Settings. In *ICCV*, 11916–11926.

Yuan, J.; Zhang, X.; Zhou, H.; Wang, J.; Qiu, Z.; Shao, Z.; Zhang, S.; Long, S.; Kuang, K.; Yao, K.; et al. 2023. HAP: Structure-Aware Masked Image Modeling for Human-Centric Perception. *NeurIPS*, 50597–50616.

Zeng, H.; Ai, H.; Zhuang, Z.; and Chen, L. 2020. Multi-task Learning via Co-attentive Sharing for Pedestrian Attribute Recognition. In *IEEE ICME*, 1–6.

Zhang, Y.; Fan, X.; Yang, Y.; Lu, Y.; and Wang, H. 2025a. Image-Attribute and Frequency-Spatial Dual Collaborative Learning for Pedestrian Attribute Recognition. *IEEE TIFS*, 20: 11715–11727.

Zhang, Y.; Lu, Y.; Yan, Y.; Wang, H.; and Li, X. 2025b. Frequency Domain Nuances Mining for Visible-Infrared Person Re-Identification. *IEEE TIFS*, 20: 5411–5424.

Zhang, Y.; and Wang, H. 2023. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification. In *CVPR*, 2153–2162.

Zhang, Y.; Yan, Y.; Li, J.; and Wang, H. 2023. MRCN: A Novel Modality Restitution and Compensation Network for Visible-Infrared Person Re-identification. In *AAAI*, volume 37, 3498–3506.

Zhang, Y.; Yan, Y.; Lu, Y.; and Wang, H. 2021. Towards a Unified Middle Modality Learning for Visible-infrared Person Re-identification. In *ACM MM*, 788–796.

Zhang, Y.; Yan, Y.; Lu, Y.; and Wang, H. 2024. Adaptive Middle Modality Alignment Learning for Visible-Infrared Person Re-identification. *IJCV*, 1–21.

Zhao, X.; Sang, L.; Ding, G.; Guo, Y.; and Jin, X. 2018. Grouping Attribute Recognition for Pedestrian with Joint Recurrent Learning. In *IJCAI*, 3177–3183.

Zheng, X.; Zhang, Y.; Lu, Y.; and Wang, H. 2024. Semi-supervised Visible-Infrared Person Re-identification via Modality Unification and Confidence Guidance. In *ACM MM*, 5761–5770.

Zhu, J.; Jin, J.; Yang, Z.; Wu, X.; and Wang, X. 2023. Learning Clip Guided Visual-text Fusion Transformer for Video-based Pedestrian Attribute Recognition. In *IEEE CVPR*, 2626–2629.