POAR: Towards Open Vocabulary Pedestrian Attribute Recognition

Yue Zhang
17112065@bjtu.edu.cn
College of Computer and Information
Engineering and Key Laboratory of
Artificial Intelligence and
Personalized Learning in Education of
Henan Province, Henan Normal
University
Xinxiang, China

Suchen Wang
wang.sc@ntu.edu.sg
School of Electrical and Electronic
Engineering, Nanyang Technological
University
Singapore

Shichao Kan
kanshichao10281078@126.com
School of Computer Science and
Engineering, Central South University
Changsha, China

Zhenyu Weng
zhenyu.weng@ntu.edu.sg
School of Electrical and Electronic
Engineering, Nanyang Technological
University
Singapore

Yigang Cen
ygcen@bjtu.edu.cn
Institute of Information Science and
Beijing Key Laboratory of Advanced
Information Science and Network
Technology, Beijing Jiaotong
University
Beijing, China

Yap-peng Tan
eyptan@ntu.edu.sg
School of Electrical and Electronic
Engineering, Nanyang Technological
University
Singapore

ABSTRACT

Pedestrian attribute recognition (PAR) aims to predict the attributes of a target pedestrian. Recent methods often address the PAR problem by training a multi-label classifier with predefined attribute classes, but they can hardly exhaust all possible pedestrian attributes in the real world. To tackle this problem, we propose a novel Pedestrian Open-Attribute Recognition (POAR) approach by formulating the problem as a task of image-text search. Our approach employs a Transformer-based Encoder with a Masking Strategy (TEMS) to focus on the attributes of specific pedestrian parts (e.g., head, upper body, lower body, feet, etc.), and introduces a set of attribute tokens to encode the corresponding attributes into visual embeddings. Each attribute category is described as a natural language sentence and encoded by the text encoder. Then, we compute the similarity between the visual and text embeddings to find the best attribute descriptions for the input images. To handle multiple attributes of a single pedestrian, we propose a Many-To-Many Contrastive (MTMC) loss with masked tokens. In addition, we propose a Grouped Knowledge Distillation (GKD) method to minimize the disparity between visual embeddings and unseen attribute text embeddings. We evaluate our proposed method on three PAR datasets with an open-attribute setting. The results demonstrate

the effectiveness of our method as a strong baseline for the POAR task. Our code is available at https://github.com/IvyYZ/POAR.

CCS CONCEPTS

• **Computing methodologies → Visual content-based indexing and retrieval**.

KEYWORDS

Pedestrian attribute recognition, CLIP, Open-attribute recognition

1 INTRODUCTION

Pedestrian attribute recognition (PAR) aims to predict attributes of a target pedestrian, such as gender, age, clothing, accessories, etc. Due to the increasing importance of person search [15, 43, 44] and scene understanding [2, 38] in many applications, PAR has emerged as an active research topic in the field of computer vision. Existing methods such as global-local methods [12, 22, 35], attention methods [22, 23], textual semantic correlations methods [3, 20] address the PAR problem by training a multi-label classifier within a predetermined attribute space. Thus they cannot recognize attributes beyond the predefined classes, such as "cotton" and "long coat" shown in Figure 1. In this work, we explore how to handle new attributes in an open world and examine a new Pedestrian Open-Attribute Recognition (POAR) problem.

Figure 1: Comparison of pedestrian attribute recognition (PAR) and pedestrian open-attribute recognition (POAR). The upper part shows the current PAR methods based on multi-label classification, where the attribute categories are predefined. During the test, PAR methods cannot recognize new attributes beyond the predefined classes, such as cotton and long coat. The lower part shows our method, which encodes the images and attributes into a joint image-text feature space. The attribute labels are then determined based on the similarities between the image and attribute embeddings.

Recently, Radford *et al.* [28] proposed a Contrastive Language-Image Pre-training (CLIP) method to model the similarity relationship between images and raw text. CLIP is trained in a task-agnostic setting and can be used to recognize general objects, such as airplane, bird, ocean, and so on. However, it is not yet generalizable on more fine-grained attributes, such as "upper stride", "lower stripe", "lower jeans", in PAR task. For PAR, the challenge is that a pedestrian may have multiple attributes but there are no corresponding location and scale information in the ground truth label set. To address this challenge, part-based [48] and attention-based [35] methods have been proposed to localize attributes and learn the attribute-specific features. However, these methods ignore the regional label conflicts, e.g., "long sleeve" and "short sleeve". Zhao *et al.* [45] proposed a grouping attribute recognition method by dividing all labels into different groups and localizing regions corresponding to each group attribute based on the detection algorithm. However, the detection algorithm is attribute sensitive. Similarly, we divide the whole attribute classes into multiple groups, and each group corresponds to one visual region, as shown in Figure 1. But contrastingly, we propose a Masking the Irrelevant Patches (MIP) strategy to eliminate insignificant image patches. The appearance of pedestrians can be divided into fixed groups, with certain parts of each group requiring no attention. The proposed strategy does not compromise the ability to identify unseen classes, and it enables accurate localization of regions for improved identification.

Different from the general PAR task that recognizes only the seen attributes in the training set, our POAR task expands beyond the seen attributes to allow new pedestrian attributes based on the application's needs. To address this challenge, we formulate the POAR problem as an image-text search task, which is trained in a downstream attribute-agnostic manner under the supervision of natural language. Different from CLIP, one image or object has only one class name; our method can use multiple attributes to describe one person in the POAR task. Using Figure 1 as an example, "male

and long coat " are used to describe the pedestrian. To address this problem, we propose multiple attribute tokens ([ATT]) with a masked encoding method in the image encoding step. As shown in Figure 1, we divide the attributes into multiple groups, and each group is encoded with an attribute token. The masking mechanism is introduced to ensure the attribute tokens are independent from one another. The original image is encoded with multiple attribute tokens corresponding to multiple groups. Based on the masking strategy and the many-to-many property, we propose a Many-To-Many Contrastive (MTMC) loss with masked tokens to guide the update of the network parameter.

We propose an end-to-end, Transformer-based Encoder with a Masking Strategy (TEMS) model. During training, we extract visual and text embeddings using the corresponding encoders. The similarity matrix of image-text pairs in a mini-batch is calculated based on these embeddings. Then, the proposed MTMC loss is used to guide the learning of the model. During testing, attribute tokens of unseen attributes such as "cotton" and "long coat" in Figure 1 can be localized based on our MIP strategy. Then, the visual embeddings of unseen attributes are extracted from the localized regions. Meanwhile, text embeddings of these attributes can be extracted using the text encoder. Finally, the attributes of a pedestrian can be recognized based on the similarities of the visual and text embeddings. Considering the potential misalignment between image embeddings and text embeddings of unseen classes, we propose a Grouped Knowledge Distillation (GKD) method by using CLIP model as a teacher network to guide the learning of the embedding space and improve the performance of the proposed TEMS model.

Our contributions can be summarized as follows:

- We formulate the problem of pedestrian open-attribute recognition (POAR) and develop an effective TEMS method as a strong baseline to address it. To the best of our knowledge, this is the first approach to address the POAR problem.

- We propose an effective masking mechanism to address the localization and encoding of multiple attributes in the TEMS model. Furthermore, we devise a many-to-many contrastive loss with masked tokens to train the network.

- We propose a grouped knowledge distillation (GKD) method to minimize the disparity between visual embeddings and unseen attribute text embeddings, so that it can be scaled to address more unseen attributes.

- We evaluate our proposed TEMS method on benchmark datasets with an open-attribute setting, and demonstrate its effectiveness as a strong baseline.

2 RELATED WORK

2.1 Pedestrian Attribute Recognition

Pedestrian attribute recognition (PAR) has received much interest in person recognition [1, 6, 13, 16, 24, 39] and scene understanding [6, 16, 24]. The mainstream methods address this problem by building a multi-label classifier based on CNN. To improve the recognition accuracy, global methods [19, 23], local methods [30], global-local methods [32], attention-based methods [12, 25, 26], sequential prediction methods [14, 41], curriculum learning methods [46], Graphic model methods [10, 34], and group based methods

Figure 2: The proposed TEMS framework for POAR task. We evaluate the dot similarity of features from the image and text encoders and then determine the attributes of the pedestrian in the image.

[17, 45] have been proposed. Nikolaos *et al.* [30] proposed an effective method to extract and aggregate visual attention masks across different scales. Tang *et al.* [35] proposed a flexible attribute localization module to learn attribute-specific regional features. These methods focused on domain-specific model designing. To use additional domain-specific guidance, M. Kalayeh *et al.* [18] used semantic segmentation methods to learn attention maps for accurate attribute prediction. Liu *et al.* [24] learned clothing attributes with additional landmark labels. Yang *et al.* [42] proposed a cascaded split-and-aggregate learning to capture both the individuality and commonality for all attributes. Li *et al.* [20] proposed an image-conditioned masked language model to learn complex sample-level attribute correlations from the perspective of language modeling. Tang *et al.* [36] and Weng *et al.* [40] employed ViT as a feature extractor for its nature of modeling long-range relations of regions. Cheng *et al.* [3] proposed an additional textual modality to explore the textual semantic correlations from attribute annotations. These methods are trained on a predefined attribute set and used to recognize the same attributes, which limits the attribute capacity of these models. In our work, we build a TEMS model based on the CLIP [28] model to recognize pedestrian open-attributes.

2.2 Open-Attributes Recognition

In classification, open-world recognition (OWR) is first proposed by Scheirer [33], which aims to discriminate known from unknown samples as well as classify known ones. Later, prototype-based method [31, 37], knowledge distillation method [11], and out of distribution detection method [9] become popular in image classification and object detection. Definitions and solutions for open-world recognition also differ based on the specific applications. Esmaeilpour *et al.* [9] used an extended model to generate candidate unknown class names for each test sample and compute a confidence score based on both the known class names and candidate unknown class names for zero-shot out-of-distribution detection. Oza *et al.* [27] proposed the class conditional auto-encoder to tackle open-set recognition, which includes closed-set training and open-set training stages. Gu *et al.* [11] adopted an image-text pre-trained model as a teacher model to supervise student detectors. Zhao *et al.*

[47] unified the label space from the training of multiple datasets to improve the generalization ability of a model. In addition, some methods [21, 29] aligned region features and the pre-trained text embeddings in base categories to realize zero-shot detection. Dan *et al.* [4] enhanced class understanding via prompt-tuning for zero-shot text classification. Du *et al.* [8] proposed a detection prompt, to learn continuous prompt representations for open-vocabulary object detection. The CLIP model was proposed by Radford *et al.* [28], which performs task-agnostic training via natural language prompting. CLIP can realize zero-shot image recognition. However, CLIP is usually used to recognize general objects, such as airplane, bird, ocean, and so on. For fine-grained attributes recognition, CLIP will fall short in these situations. Our work addresses the pedestrian open-attribute recognition task by identifying both classes, seen and unseen classes, simultaneously.

3 METHOD

3.1 Pedestrian Open-Attribute Recognition

Pedestrian attribute recognition aims to recognize the fine-grained attributes of a pedestrian (e.g., hairstyle, age, gender, etc.) from the given image. The conventional PAR usually predetermines a set of pedestrian attributes and follows the close-set assumption during the training and test phases. Suppose there are M predetermined pedestrian attributes $\mathcal{A} = \{A_1, A_2, \ldots, A_M\}$, e.g., $A_1 =$"long hair", $A_2 =$"short hair", etc. Then, given a labeled pedestrian dataset $\mathcal{D} = \{(I_i, \mathcal{A}_i)\}_{i=1}^N$, where each image $I_i \in \mathbb{R}^{H \times W \times 3}$ is annotated with a subset $\mathcal{A}_i \subset \mathcal{A}$ of the existing pedestrian attributes. The main objective of PAR is to learn a model to answer which pedestrian attributes from $\mathcal{A}$ appear in a given image I.

Existing approaches [3, 12, 35] usually convert this to a multi-label classification problem. However, the pedestrian attributes in the real world are potentially unlimited. It is difficult to exhaust all the attributes in a predetermined set and collect the corresponding pedestrian images. Unseen attributes are highly possible to exist in real world applications, while existing classification-based methods [3, 20] are inherently incapable of handling such cases. To address this limitation, we formulate a pedestrian open-attribute recognition problem in this work. Let $\mathcal{A}_u = \{A_{M+1}, A_{M+2}, \ldots, A_{M+M_u}\}$

Table 1: Attribute labels of the PETA dataset converted to sequence by prompt.

ATTRIBUTES	KEY	PROMPT
Long, Short	Hair	This person has {} hair.
Male, Female	Gender	This person is {}.
Less15, Less30, Less45, Less60, Larger60	Age	The age of this person is {} years old.
Backpack, MessengerBag, PlasticBags, Other, Nothing	Carry	This person is carrying {}.
Sunglasses, Hat, Muffler, Nothing	Accessory	This person is accessory {}.
LeatherShoes, Sandals, Sneaker, Shoes	Foot	This person is wearing {} in foot.
Casual, Formal, Jacket, Logo, ShortSleeve, Plaid, Stripe, Tshirt, VNeck, Other	Upperbody	This person is wearing {} in upper body.
Casual, Formal, Trousers, ShortSkirt, Shorts, Plaid, Jeans	Lowerbody	This person is wearing {} in lower body.

denote a set of extra attributes that are also of interest in the test phase but not included in the predetermined attribute set $\mathcal{A}$. In POAR, we expect that the model can recognize not only the seen attributes from $\mathcal{A}$ but also the unseen attributes from $\mathcal{A}_u$.

3.2 Framework

Overview. As illustrated in Figure 2, the proposed Transformer-based encoder with a masking strategy (TEMS) framework consists of an image encoder Φ_I and a text encoder Φ_T. The image encoder processes the input images and derives the visual representation of various pedestrian attributes. Besides, we construct a set of text descriptions (e.g., "this person has long hair.", "this person is carrying backpack.", etc, and utilize the text encoder of CLIP to encode these attribute descriptions into text embeddings. Then, we compute the vision-text similarity to find the best-matched pedestrian attributes for the input images. Different from the original CLIP, we assume one pedestrian may have more than one associated attribute, as shown in Figure 3. To train our model, we propose a many-to-many contrastive (MTMC) loss with masked tokens to handle the many-to-many relationships between the images and text descriptions. The details of each part are described below.

Image Encoding. The image encoding process is organized based on token prediction with Transformer. First, the input image I is split into a sequence of non-overlapping small patches $\{P_0, P_1, \cdots, P_{S-1}\}$, where $P_i \in \mathbb{R}^{r \times r \times 3}$ and $S = \frac{H}{r} \times \frac{W}{r}$. Then, each patch P_i ([PAT]) is projected into the embedding space as a vector $\mathbf{x}_i \in \mathbb{R}^D$, where D denotes the feature dimension. Thus, we can obtain $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{S-1}] \in \mathbb{R}^{D \times S}$ with S patch embeddings. Inspired by [7, 28], we introduce K learnable attribute tokens $\mathbf{Z} = [\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{K-1}] \in \mathbb{R}^{D \times K}$, where each attribute token $\mathbf{z}_k \in \mathbb{R}^D$ shares the same dimension as $\mathbf{x}_i$. For each patch embedding and each class token, we generate a corresponding position encoding $\mathbf{e}_j \in \mathbb{R}^D$, where $j \in \{0, 1, \cdots, S + K - 1\}$. The input of the image encoder Φ_I is the concatenation of class tokens and patch embeddings combined with their positional embeddings $\mathbf{E} = [\mathbf{e}_0, \mathbf{e}_1, \cdots, \mathbf{e}_{S+K-1}] \in \mathbb{R}^{D \times (S+K)}$, which is denoted as $\mathbf{V} = [\mathbf{Z}, \mathbf{X}] \oplus \mathbf{E}$, where $\mathbf{V} \in \mathbb{R}^{D \times (S+K)}$, $[,]$ is the concatenate operation, and $\oplus$ denotes the element-wise summation notation. The output of the image encoder Φ_I is the learned embeddings of class tokens, denoted as $\hat{\mathbf{Z}} \in \mathbb{R}^{D \times K}$. The above process can be summarized as follows:

$$\hat{\mathbf{Z}} = \Phi_I(\mathbf{V}). \tag{1}$$

Text Encoding. The text encoding is performed by using the text encoder Φ_T transferred from the CLIP model. The input of the text encoder is the prompts of pedestrian attributes which are organized

Figure 3: Diagram of image-text relationship. The right part represents the many-to-many relationship of image-text.

as follows. First, the attributes are divided into K groups based on the characteristics of pedestrians, as shown in Table 1. Then, we form a prompt template for the attributes of each group. Finally, each prompt sentence in Table 1 is tokenized into an embedding using the Byte-Pair-Encoding method[1]. Suppose that the maximum number of words in any of these sentences is L. Then, each sentence is tokenized into a vector $\mathbf{y}_i \in \mathbb{R}^L$. For the k-th group, we can obtain m sentence vectors corresponding to the m prompts in this group, denoted as $\mathbf{Y}^k = [\mathbf{y}_1^k, \mathbf{y}_2^k, \cdots, \mathbf{y}_m^k] \in \mathbb{R}^{L \times m}$. The output of the text encoder Φ_T is the learned text embeddings $\hat{\mathbf{Y}}^k$ for the prompts of the k-th group. The above process can be defined as:

$$\hat{\mathbf{Y}}^k = \Phi_T(\mathbf{Y}^k), \tag{2}$$

where $\hat{\mathbf{Y}}^k \in \mathbb{R}^{D \times m}$.

Many-to-Many Contrastive Loss. For a mini-batch images $\{I_1, I_2, \cdots, I_B\}$, we first extract their token embeddings $\hat{\mathbf{Z}}_b$ and text embeddings $\hat{\mathbf{Y}}_b$ using (1) and (2), respectively. $\hat{\mathbf{Z}}_b$ and $\hat{\mathbf{Y}}_b$ are sets of token embeddings and text embeddings in all attribute groups related to the given mini-batch images, respectively. Different from the CLIP model, which has a one-to-one relationship between image and text. In the TEMS model, the image and text are involved in a many-to-many relationship, as shown in Figure 3. To effectively tackle this scenario, a loss function combined with visual-to-text and text-to-visual contrastive learning is proposed. The visual-to-text contrastive learning term is defined as follows:

$$\mathcal{L}_{v2t} = -\sum_{i=1}^{v} \sum_{j=1}^{t_i} \log \frac{exp\left(\hat{\mathbf{z}}_i^T \hat{\mathbf{y}}_j^+ / \tau\right)}{\sum_{k=1}^{t} exp\left(\hat{\mathbf{z}}_i^T \hat{\mathbf{y}}_k / \tau\right)}, \tag{3}$$

where τ is a temperature parameter, t_i is the number of positive text embeddings that have the same attribute label with $\hat{\mathbf{z}}_i$. v and t

---

[1]https://huggingface.co/transformers/v4.8.0/model_doc/clip.html

are the total numbers of token embeddings and text embeddings, respectively. $\hat{z}_i \in \hat{Z}_b$, $\hat{y}_k \in \hat{Y}_b$, $\hat{y}_j^+ \in \hat{Y}_b$. $\hat{z}_i$ and $\hat{y}_j^+$ are a positive pair share the same attribute label. Similarly, the text-to-visual contrastive learning term is defined as follows:

$$\mathcal{L}_{t2v} = -\sum_{j=1}^{t} \sum_{i=1}^{v_j} log \frac{exp\left(\hat{y}_j^T \hat{z}_i^+ / \tau\right)}{\sum_{k=1}^{v} exp\left(\hat{y}_j^T \hat{z}_k / \tau\right)}, \quad (4)$$

where v_j is the number of positive token embeddings that have the same attribute label with $\hat{y}_j$. The final loss function is the combination of (3) and (4), defined as:

$$\mathcal{L}_1 = \mathcal{L}_{v2t} + \mathcal{L}_{t2v}. \quad (5)$$

3.3 Encoder Networks and Masking Strategy

Image Encoder. The image encoder Φ_I is a stack of multiple Transformer blocks. Each Transformer block is composed of layer norm (LN) layers [7], a multi-head self-attention (MSA) layer [7], and a multi-layer perceptron (MLP) network. In the multi-head self-attention layer, the attention weights of each attribute token are automatically calculated among all input tokens. Specifically, the attention weights between the k-th attribute token and the others can be computed as

$$\text{Attn}_{\text{token}}(z_k) = \text{softmax}\left(\frac{[z_k^\top \cdot Z, \ z_k^\top \cdot X]}{\sqrt{D}}\right). \quad (6)$$

We observe that the first term $z_k^\top Z$ often dominates the attention, making the module hardly find the true regions of interest from the input image. This shortcut learning often leads to overfitting and inferior results in our experiments. To address this issue, we mask out the self-attentions between the attribute tokens. This allows useful visual information to be extracted from the input image rather than simply relying on the information that has been extracted by the others. Specifically, we calculate the attention weight of the k-th attribute token as

$$\text{Attn}_{\text{mask}}(z_k) = \text{softmax}\left(\frac{[z_k^\top \cdot X]}{\sqrt{D}}\right). \quad (7)$$

In our experiments, we observe that this technique leads to significant performance gain (Table 7).

In the PAR task, one key challenge is that a pedestrian can have multiple attributes, and there is no corresponding location and scale information in the ground truth label set. We propose to mask the irrelevant patches (MIP) to tackle this challenge. Specifically, we divide the whole attribute classes into multiple groups, and each group ([ATT] token) corresponds to one visual region. Then we mask out regions ([PAT] token) that do not need to pay attention to. For example, the "hair" class token pays more attention to the head of the pedestrian, and we block out regions on the lower part of the head region; similarly, the "upper body" class token pays more attention to the top part of the image, and we block out the bottom part of the image, etc. Thus, the final attention can thus be calculated as follows

$$\text{Attn}_{\text{mask}}(z_k^l) = \text{softmax}\left(\frac{z_k^\top \cdot X + \varpi}{\sqrt{D}}\right), \quad (8)$$

Figure 4: The grouped knowledge distillation framework.

where $\varpi \in \mathbb{R}^{1 \times S}$ is a mask vector with value $-\infty$ for the blocked image patches and 0 otherwise. Taking the "hair" class token, for example, the region of the head will be 0, and the remaining regions will be $-\infty$. The output of the self-attention unit is as follows:

$$F^l = \text{Attn}_{\text{mask}}(Z_k^{l-1}) \cdot V^{l-1}, \quad (9)$$

where l is the layer index of the self-attention, $Z^0 = Z$, $V^0 = V$. The whole process of a Transformer block can be formulated as:

$$\hat{V}^l = \text{MSA}(\text{LN}(V^{l-1})) + V^{l-1}, \quad (10)$$

$$V^l = \text{MLP}(\text{LN}(\hat{V}^l)) + \hat{V}^l. \quad (11)$$

Contrastive Learning with Masked Tokens. It should be noted that our contrastive loss is computed based on all the token embeddings and attribute embeddings. The embeddings of those masked tokens are also used to compute the loss. This is because the contrastive loss computed with the masked embeddings will lead to more robust embedding learning.

3.4 Open-Attribute Recognition

The knowledge distillation allows for the transfer of knowledge from a complex teacher model to a simpler student model, albeit a slight decrease in performance. The previous paper underscored CLIP's impressive generalization capabilities. Nevertheless, its accuracy in certain tasks, such as pedestrian attribute recognition, leaves room for improvement. To address this, we propose the grouped knowledge distillation (GKD) method, as shown in Figure 4, which leverages the CLIP model as a teacher model and distills its extensive knowledge into our pedestrian attribute model during training. The objective function is defined as:

$$\mathcal{L}_2 = \sum_{i=1}^{K} KL(\hat{w}^l, \hat{z}_i^l) + \mathcal{L}_1. \quad (12)$$

The Kullback-Leibler (KL) divergence is computed by each visual token embedding $\hat{z}_i^l$ of the student network and the visual embedding $\hat{w}^l$ of the teacher network. These loss functions introduce a non-linear transformation to the input data, which helps to capture more complex and abstract features of the data. The CLIP model has strong generalization performance and can be used as a teacher network to guide the learning of embedding space, reducing semantic disparities between the image embedding features and the unseen text features, and thereby improving the performance of the TEMS model.

Table 2: Performance comparison of POAR experimental results. Blue indicates the model is trained and tested on the same dataset. ∗ denotes our implementation with the official code.

Method	Source Domain	Target Domain					
		PETA		PA100K		RAPv1	
		R@1	R@2	R@1	R@2	R@1	R@2
CLIP*	–	50.2	75.7	43.4	65.9	33.6	56.5
VTB*	PA100K	31.4	62.2	26.9	62.2	24.2	50.7
TEMS	PETA	87.6	96.0	45.1	73.5	42.2	68.6
TEMS+CLIP		–	–	44.7	74.7	42.1	69.7
TEMS (GKD)		86.5	95.7	**51.9**	**78.5**	**46.2**	**71.2**
TEMS	PA100K	42.3	76.2	83.3	92.6	39.4	63.6
TEMS+CLIP		50.9	77.5	–	–	39.4	64.5
TEMS (GKD)		50.8	**81.4**	81.3	92.8	44.1	63.7
TEMS	RAPv1	48.8	75.0	45.1	73.1	80.6	94.4
TEMS+CLIP		52.9	78.8	45.5	67.2	–	–
TEMS (GKD)		**55.0**	76.3	50.3	74.7	77.5	92.3

4 EXPERIMENTS

4.1 Datasets and Experimental Settings

Datasets and Evaluation Metrics. The proposed method is evaluated on three benchmark datasets, i.e., PETA [5], RAPv1 [19], and PA100K [23] with an open-attribute setting, meaning the model is trained on one dataset and evaluated on the other two datasets. The classes included in different datasets may not be identical. Recall@K and mA (mean Accuracy) are used to evaluate the performance of our open-attribute recognition model. For the PETA dataset, the "upper body" and "lower body" groups may include two attribute classes. We select R@2 texts to show for attribute groups in "upper body" and "lower body".

Implementation Details. Our experiments are conducted on the ViT-B/16 backbone networks, which are stacked with 12 Transformer blocks. The input image size is set to 224×224. The value of r is 16, and K is set to be 8, 8, and 11 in PETA, PA100K, and RAPv1, respectively. The learning rate is $5e^{-2}$ with a weight decay of 0.2. The temperature τ in contrastive loss is set as 5. Data augmentation with random horizontal flip and random erasing are used during training.

4.2 Performance Comparison of POAR

We compare the performance of our method with the CLIP [28] and VTB [3] methods based on the image-to-text K-nearest neighbor

Table 3: The number of attribute classes in the test phase.

Source Domain	Target Domain					
	PETA		PA100K		RAPv1	
	SN	UN	SN	UN	SN	UN
PETA	35	0	8	18	12	39
PA100K	8	27	26	0	8	43
RAPv1	12	23	8	18	51	0

retrieval idea, the top-1 and top-2 recall rates are shown in Table 2. In Table 3, SN and UN represent seen and unseen attributes in the different test sets, respectively.

From Table 2, we can see that the Recall@1 (R@1) scores of our method are 1.7% and 8.6% higher than the CLIP method when the model is trained on the PETA dataset and evaluated on the PA100K and RAPv1 datasets, respectively. When the model is trained on the PA100K and evaluated on the PETA dataset, the R@1 score of our method is 7.9% lower than the CLIP model. When the model is trained on the RAPv1 and evaluated on the PETA, the R@1 scores of our method are 1.4% lower than the CLIP model. This is likely due to the fact that the CLIP model is trained on 400 million image-text pairs which have a high probability of containing "age" and "hair" and other attributes; thus, those unseen attributes defined in our experiments can be considered seen in the CLIP model. However, when we fuse the features of our model and the CLIP model, a higher performance for the POAR task can be obtained. When we train the proposed TEMS model with GKD using the CLIP model as the teacher model, the model's transferred performance to other datasets improves significantly. However, the TEMS model's performance on its training dataset decreases as compared to the results without distillation. For instance, when the model is trained on the PETA dataset, its performance on the PETA test set decreases by including distillation. Nevertheless, the model exhibits significant improvements on other datasets.

To analyze the details of the POAR performance, we show the performance comparison of each group of the CLIP model and the proposed TEMS model in Figure 5. The results on different datasets and attributes vary between the TEMS method and the CLIP method. However, the model trained through GKD exhibits a significant improvement in the accuracy of various attribute tokens, as evidenced by the comparison between the orange (without GKD) and yellow (with GKD) graphs in Figure 5.

Furthermore, we compute the recognition performance of seen and unseen attributes in the PA100K dataset, respectively. Results are shown in Figure 6, the proposed TEMS method has the capability

Figure 5: Performance comparison of different groups ([ATT] tokens) by different methods based on the same experimental settings. (a), (b), and (c) represent the results of PETA, PA100K, and RAPv1 datasets, respectively.

Table 4: Results of text-to-image retrieval. Blue indicates that the training and test sets belong to same dataset.

Method	Source Domain	Target Domain								
		PETA			PA100K			RAPv1		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
TEMS	PETA	90.3	100.0	100.0	38.5	57.7	61.5	35.3	52.9	62.8
TEMS	PA100K	41.9	74.2	77.4	88.5	96.2	100.0	33.3	56.9	64.7
TEMS	RAPv1	35.5	58.1	71.0	34.6	61.5	69.2	96.1	100.0	100.0

Figure 6: The R@1 scores were calculated for each attribute token in the PA100K dataset, considering both seen and unseen attributes.

Table 5: Evaluation of each component on the PETA dataset.

SAT	MAT	MAM	MIP	mA	F1	R@1	R@2
√				81.1	83.0	85.7	94.7
	√			80.6	82.2	86.2	95.7
	√	√		81.0	83.0	86.4	95.7
	√	√	√	**83.1**	**84.4**	**87.6**	**96.0**

Table 6: Our test results for different loss functions on the PETA dataset.

OTOC	MTMC	mA	F1	R@1	R@2
√		73.0	73.4	86.3	95.8
√	√	81.5	83.3	86.3	95.8
	√	**83.1**	**84.4**	**87.6**	**96.0**

Figure 7: The attention map of each token.

to recognize both seen and unseen categories. For example, when the method is trained on PETA dataset and tested on the PA100K dataset, Figure 6 shows the accuracy of the "Carry" group for seen (blue) and unseen (gray) classes. Moreover, after implementing GKD training, there is a remarkable enhancement in the model's ability to recognize unseen classes, as evidenced by the comparison of R@1 results for unseen classes in Figure 6 (a) (b) (gray and yellow). This further confirms the effectiveness of the proposed method in identifying unseen classes.

4.3 Text-to-Image Retrieval Results

To examine the generalization ability of our method, we also evaluate the text-to-image retrieval results. The results are reported in Table 4. Compared to the image-to-text retrieve performance of Table 2, we can see that the text-to-image recognition task is more challenging than the image-to-text recognition task. This is mainly due to the fact that the text embedding space is more sparse than the image embedding space.

CLIP Model:
1. This person is male.
2. This person is wearing stride in upper body.
3. This person is wearing stripe in lower body.
4. The age of this person is between eighteen and sixty years old.
5. This person is other.
6. This person is wearing other in foot.
7. This person is carrying hold objects in front.
8. This person is accessory other.
Unseen classes: not sure

TEMS Model:
1. This person is male.
2. This person is wearing short sleeve in upper body.
3. This person is wearing trousers in lower body.
4. The age of this person is over sixty.
5. This person is back.
6. This person is wearing other in foot.
7. This person is carrying other.
8. This person is accessory other.
Unseen classes: back

(i)

CLIP Model:
1. This person is female.
2. This person is wearing other in upper body.
3. This person is wearing pattern in lower body.
4. The age of this person is between eighteen and sixty years old.
5. This person is other.
6. This person is wearing other in foot.
7. This person is carrying hold objects in front.
8. This person is accessory other.
Unseen classes: not sure

TEMS Model:
1. This person is female.
2. This person is wearing short sleeve in upper body.
3. This person is wearing skirt and dress in lower body.
4. The age of this person is less eighteen years old.
5. This person is back.
6. This person is wearing other in foot.
7. This person is carrying shoulder bag.
8. This person is accessory other.
Unseen classes: skirt and dress , less eighteen, back, shoulder bag

(ii)

(a) Examples of image and text matching on the PA100K dataset.

CLIP Model:
1. This person is male.
2. This person is wearing short sleeve in upper body.
3. This person is wearing long trousers in lower body.
4. The age of this person is between thirty-one and forty-five years old.
5. This person is wearing casual in foot.
6. This person is carrying paper bag.
7. This person is accessory glasses .
8. This person is clerk.
9. This person is fat.
10. This person has bold head.
11. This person is carry arm.
Unseen classes: not sure

TEMS Model:
1. This person is male.
2. This person is wearing cotton in upper body.
3. This person is wearing jeans in lower body.
4. The age of this person is between seventeen and thirty years old.
5. This person is wearing casual in foot.
6. This person is carrying nothing.
7. This person is accessory other .
8. This person is clerk.
9. This person is normal.
10. This person has long hair.
11. This person is nothing special.
Unseen classes: between seventeen and thirty, clerk, normal, nothing special

(i)

CLIP Model:
1. This person is female.
2. This person is wearing tight in upper body.
3. This person is wearing short skirt in lower body.
4. The age of this person is between thirty-one and forty-five years old.
5. This person is wearing casual in foot.
6. This person is carrying paper bag.
7. This person is accessory other.
8. This person is clerk.
9. This person is fat.
10. This person has black hair.
11. This person is carry arm.
Unseen classes: not sure

TEMS Model:
1. This person is female.
2. This person is wearing short sleeve in upper body.
3. This person is wearing tight trousers in lower body.
4. The age of this person is between seventeen and thirty years old.
5. This person is wearing other in foot.
6. This person is carrying shoulder bag.
7. This person is accessory other.
8. This person is clerk.
9. This person is normal.
10. This person has long hair.
11. This person is nothing special.
Unseen classes: between seventeen and thirty, shoulder bag, clerk, normal, nothing special

(ii)

(b) Examples of image and text matching on the RAPv1 dataset.

Figure 8: Image-to-text retrieval examples. The model is trained on the PETA dataset. Prompts with blue color indicate that the predicted text is inconsistent with the ground truth. Text with an underline denotes unseen attributes during training.

4.4 Ablation Study

Our ablation study is conducted on the PETA dataset using the close-set and open-set evaluation mechanisms. Table 7 shows the image-to-text retrieval performance of different components in our proposed method, SAT represents a single attribute token. MAT represents multiple attribute tokens. MAM represents multiple attribute tokens with the masking. Table 6 shows the image-to-text retrieval performance for different loss functions, OTOC represents one-to-one contrastive loss. MTMC represents many-to-many contrastive loss. The one-to-one loss function is performed by defining all attributes of one image in a paragraph description as text input. From Table 7, we can see that each component of the proposed method can contribute positively to the final performance gain. From Table 6, we can see that the many-to-many loss function significantly improves the final performance.

4.5 Visualization of the Attention Maps

To illustrate the effectiveness of the proposed masking method, we show the attention map of each attribute token on the PETA dataset in Figure 7. In our method, we proposed a masking strategy to block out distract regions and mask the corresponding attribute tokens. From the attention map, we can see that each attribute token is responsible for a specific part of the body, which shows the effectiveness of the proposed masking method.

4.6 Image-to-Text Retrieval Examples

In Figure 8, we visualize some image-to-text retrieval examples based on embeddings obtained using the CLIP model and our proposed TEMS model. The TEMS model was trained on the PETA dataset. Figure 8 (a) and Figure 8 (b) show examples tested on the PA100K and RAPv1 datasets, respectively. We can see that there

are a lot of unseen attributes in the test set, and our model can effectively recognize the unseen attributes. From Figure 8, the CLIP model exhibits a higher rate of errors in recognizing pedestrian attribute categories. In contrast, as illustrated in Figure 8 (a) (i), our proposed model achieves perfect recognition of the pedestrian attributes, while the CLIP model only correctly identifies three attributes for the same pedestrian. In the RAPv1 dataset, there are 39 unseen classes, the recognition accuracy for these classes is much higher by the TEMS model than that of the CLIP model.

5 CONCLUSIONS

In this work, we propose a novel method to tackle the problem of pedestrian open-attribute recognition. Our key idea is to formulate the POAR problem as an image-text search problem. Specifically, we propose a TEMS method to encode image patches with attribute tokens. Then, a many-to-many contrastive loss function with masked tokens is developed to train the model. Finally, the knowledge distillation method is employed to improve the recognition of unseen attribution classes. Experimental results on benchmark PAR datasets with an open-attribute setting show the effectiveness of the proposed method. Our proposed POAR task is also promising, as its performance can be further improved by leveraging on the advances in more sophisticated multimodality technologies, like ChatGPT / T5-11B.

Limitations: One limitation of the work is that the text encoder is directly transferred from the CLIP. It is beneficial to examine whether a more effective attribute encoder can be developed in our future work. Another limitation is that the input of the framework is the detected pedestrian; future work could be focused on integrating pedestrian detection and POAR into a unified framework to improve the efficiency of pedestrian open-attribute identification.

ACKNOWLEDGMENTS

REFERENCES

[1] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. 2018. Partially Shared Multi-Task Convolutional Neural Network With Local Constraint for Face Attribute Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alexander G Hauptmann. 2021. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[3] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. 2022. A Simple Visual-Textual Baseline for Pedestrian Attribute Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–1.

[4] Yuhao Dan, Jie Zhou, Qin Chen, Qingchun Bai, and Liang He. 2022. Enhancing Class Understanding Via Prompt-Tuning For Zero-Shot Text Classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4303–4307.

[5] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*. 789–792.

[6] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. *CoRR* abs/2107.12666 (2021). arXiv:2107.12666 https://arxiv.org/abs/2107.12666

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14084–14093.

[9] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. 2022. Zero-Shot Out-of-Distribution Detection Based on the Pretrained Model CLIP. In *Proceedings of the AAAI conference on artificial intelligence*.

[10] Haonan Fan, Hai-Miao Hu, Shuailing Liu, Weiqing Lu, and Shiliang Pu. 2022. Correlation Graph Convolutional Network for Pedestrian Attribute Recognition. *IEEE Transactions on Multimedia* 24 (2022), 49–60.

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921* (2021).

[12] Hao Guo, Xiaochuan Fan, and Song Wang. 2022. Visual Attention Consistency for Human Attribute Recognition. *International Journal of Computer Vision* 130, 4 (2022), 1088–1106.

[13] Emily M. Hand and Rama Chellappa. 2017. Attributes for Improved Attributes: A Multi-Task Network Utilizing Implicit and Explicit Relationships for Facial Attribute Classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder Singh and Shaul Markovitch (Eds.). AAAI Press, 4068–4074.

[14] Keke He, Zhanxiong Wang, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. 2017. Adaptively weighted multi-task deep network for person attribute classification. In *Proceedings of the 25th ACM international conference on Multimedia*. 1636–1644.

[15] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. TransReID: Transformer-Based Object Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 15013–15022.

[16] Jian Jia, Xiaotang Chen, and Kaiqi Huang. 2021. Spatial and Semantic Consistency Regularizations for Pedestrian Attribute Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 962–971.

[17] Jian Jia, Naiyu Gao, Fei He, Xiaotang Chen, and Kaiqi Huang. 2022. Learning disentangled attribute representations for robust pedestrian attribute recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1069–1077.

[18] Mahdi M. Kalayeh, Boqing Gong, and Mubarak Shah. 2017. Improving Facial Attribute Prediction Using Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. 2015. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 111–115.

[20] Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. 2022. Label2Label: A Language Modeling Framework for Multi-Attribute Learning. *arXiv e-prints*, Article arXiv:2207.08677 (July 2022), arXiv:2207.08677 pages. arXiv:2207.08677 [cs.CV]

[21] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. 2019. Zero-shot object detection with textual descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8690–8697.

[22] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. 2018. Localization Guided Learning for Pedestrian Attribute Recognition. In *British Machine Vision Conference 2018, BMVC 2018*.

[23] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. 2017. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*. 350–359.

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Zhenyu Liu, Zhang Zhang, Da Li, Peng Zhang, and Caifeng Shan. 2022. Dual-branch self-attention network for pedestrian attribute recognition. *Pattern Recognition Letters* 163 (2022), 112–120.

[26] Wei-Qing Lu, Hai-Miao Hu, Jinzuo Yu, Yibo Zhou, Hanzi Wang, and Bo Li. 2023. Orientation-Aware Pedestrian Attribute Recognition based on Graph Convolution Network. *IEEE Transactions on Multimedia* (2023).

[27] Poojan Oza and Vishal M Patel. 2019. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2307–2316.

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

[29] Shafin Rahman, Salman Khan, and Nick Barnes. 2020. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11932–11939.

[30] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. 2018. Deep Imbalanced Attribute Classification using Visual Attention Aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[31] Piyapat Saranrittichai, Chaithanya Kumar Mummadi, Claudia Blaiotta, Mauricio Munoz, and Volker Fischer. 2022. Multi-attribute Open Set Recognition. In *DAGM German Conference on Pattern Recognition*. Springer, 101–115.

[32] M. Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen. 2017. Deep View-Sensitive Pedestrian Attribute Inference in an end-to-end Model. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press.

[33] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. 2013. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 7 (2013), 1757–1772.

[34] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z Li. 2020. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 12055–12062.

[35] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. 2019. Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[36] Zengming Tang and Jun Huang. 2022. DRFormer: Learning dual relations using Transformer for pedestrian attribute recognition. *Neurocomputing* 497 (2022), 159–169.

[37] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2022. Open-Set Recognition: a Good Closed-Set Classifier is All You Need?. In *International Conference on Learning Representations*.

[38] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. 2022. Learning Transferable Human-Object Interaction Detector With Natural Language Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 939–948.

[39] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. 2022. Pedestrian attribute recognition: A survey. *Pattern Recognition* 121 (2022), 108220.

[40] Dunfang Weng, Zichang Tan, Liwei Fang, and Guodong Guo. 2023. Exploring attribute localization and correlation for pedestrian attribute recognition. *Neurocomputing* 531 (2023), 140–150.

[41] Jie Yang, Jiarou Fan, Yiru Wang, Yige Wang, Weihao Gan, Lin Liu, and Wei Wu. 2020. Hierarchical Feature Embedding for Attribute Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] Yang Yang, Zichang Tan, Prayag Tiwari, Hari Mohan Pandey, Jun Wan, Zhen Lei, Guodong Guo, and Stan Z Li. 2021. Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. *International Journal of Computer Vision* 129 (2021), 2731–2744.

[43] Yue Zhang, Yi Jin, Jianqiang Chen, Shichao Kan, Yigang Cen, and Qi Cao. 2020. PGAN: Part-Based Nondirect Coupling Embedded GAN for Person Reidentification. *IEEE MultiMedia* 27, 3 (2020), 23–33.

[44] Yue Zhang, Fanghui Zhang, Yi Jin, Yigang Cen, Viacheslav Voronin, and Shaohua Wan. 2022. Local Correlation Ensemble with GCN based on Attention Features for Cross-domain Person Re-ID. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[45] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin. 2018. Grouping attribute recognition for pedestrian with joint recurrent learning.. In *IJCAI*, Vol. 2018. 27th.

[46] Xin Zhao, Liufang Sang, Guiguang Ding, Jungong Han, Na Di, and Chenggang Yan. 2019. Recurrent attention model for pedestrian attribute recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9275–9282.

[47] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. 2020. Object detection with a unified label space from multiple datasets. In *European Conference on Computer Vision*. Springer, 178–193.

[48] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z. Li. 2015. Multi-label CNN based pedestrian attribute learning for soft biometrics. In *2015 International Conference on Biometrics (ICB)*. 535–540.

A   SUPPLEMENTARY MATERIAL

A.1   Datasets and Metrics

We evaluate the proposed TEMS baseline on three large-scale benchmark pedestrian attribute datasets.

PETA dataset. Following the standard settings, 9,500 images are used for training, 1,900 images are used for verification, and 7,600 images are used for testing. The model is evaluated on 35 attributes.

RAPv1 dataset. It is collected from 26 real indoor surveillance cameras and contains 41,585 pedestrians, where 33,268 images are used for training, and 8,317 images are used for testing. The model is evaluated on 51 attributes.

PA100K dataset. It includes 100,000 images, of which 80,000 images are used for training, 10,000 images are used for validation, and 10,000 images are used for testing. Each image is annotated with 26 commonly used attributes.

Metrics. Our model is trained on one dataset and evaluated on the other two datasets. The top-k results are used to evaluate the performance of our open-attribute recognition model.

A.2   Ablation Experiment

Our ablation study is conducted on the RAPv1 dataset using the open-set evaluation mechanisms. Table 7 shows the image-to-text retrieval performance of different components in our proposed method. From Table 7, we can see that each component of the proposed method can contribute positively to the final performance gain.

Table 7: Evaluation of each component on the RAPv1 dataset, the model is trained on the PETA dataset. SAT represents a single attribute token. MAT represents multiple attribute tokens. MAM represents multiple attribute tokens with the masking.

SAT	MAT	MAM	MIP	R@1	R@2
√				39.9	56.1
	√			41.3	61.6
	√	√		41.1	65.7
	√	√	√	42.2	68.6

A.3   Text-to-image Retrieval Examples

We also show examples of text-to-image retrieval on three datasets for models trained on the PETA dataset. Specifically, as shown in Figure 9, the first line is the results on the PETA dataset, the second line is the results on the RAPv1 dataset, and the third line is the results on the PA100K dataset. The second and third texts are unseen attributes. Each text corresponds to top-5 results. Figure 9 (a) shows the results of the CLIP method, and Figure 9 (b) shows the results of our proposed TEMS method for POAR task. For the PA100K and RAPv1 datasets, the TEMS model can also retrieve some correct images that belong to the unseen attributes, such as the results of "This person is customer.".

In addition, we also show examples using all the text descriptions to generate embeddings to search images on the PETA dataset.

(a)   (b)

Figure 9: Text search image visualization. (a) represents the top-5 results of the CLIP method. (b) represents top-5 the results of our TEMS method. The green (blue) dot indicates the given text description is consistent (inconsistent) with the ground truth of the image.

Results are shown in Figure 10. As can be seen from Figure 10 (a), (b), the results of the top-5 are all correct. The proposed model retrieved correct images based on textual descriptions, which is because the description includes different attributes from different groups. From these examples, we can see that using a paragraph description can also find different images of the same person, which may be helpful for person re-identification applications.

A.4   Image-to-text Retrieval Examples on Complex Scenes

Pedestrian attribute recognition in complex scenes can be effectively achieved through a two-step process involving person detection followed by attribute recognition for each individual. Here we use the WIDERAttribute dataset with 14 human attribute labels and 30 event class labels for testing. Dataset URL http://mmlab.ie.cuhk.edu.hk/projects/WIDERAttribute.

In Figure 11, we present the outcomes of testing our model on the WIDERAttribute dataset, which was trained using the PETA dataset. The test images were entirely new and did not overlap with the training set. Moreover, these images contain multiple pedestrians, often with occlusions, making the recognition task more complex. In some instances, while identifying specific pedestrians, there could be potential influences from other pedestrians' attribute features. However, overall, our model demonstrates commendable recognition ability even in such challenging scenarios.

1. This person is female.
2. This person has long hair.
3. This person is wearing casual in upper body.
   This person is wearing long sleeve in upper body.
4. This person is wearing casual in lower body.
   This person is wearing trousers in lower body.
5. The age of this person is between fifteen and thirty years old.
6. This person is wearing other in foot.
7. This person is carrying backpack.
8. This person is accessory hat .

(a)

1. This person is male.
2. This person has short hair.
3. This person is wearing casual in upper body.
   This person is wearing long sleeve in upper body.
4. This person is wearing casual in lower body.
   This person is wearing trousers in lower body.
5. The age of this person is over sixty.
6. This person is wearing other in foot.
7. This person is carrying other.
8. This person is accessory other .

(b)

Figure 10: Text search image visualization. The green dot indicates the given text description is consistent with the ground truth of the image.

1
This person is Female.
This person has ShortHair.
This person is wearing Stripe in upper body.
This person is wearing Other in lower body.
This person is wearing Casual.
This person is accessory FaceMask.

3
This person is Male.
This person has ShortHair.
This person is wearing Tshirt in upper body.
This person is wearing Other in lower body.
This person is wearing Formal.
This person is accessory Hat.

5
This person is Male.
This person has ShortHair.
This person is wearing Tshirt in upper body.
This person is wearing Other in lower body.
This person is wearing Formal.
This person is accessory Other.

7
This person is Female.
This person has LongHair.
This person is wearing Stripe in upper body.
This person is wearing Other in lower body.
This person is wearing Other.
This person is accessory Other.

8
This person is Male.
This person has ShortHair.
This person is wearing Tshirt in upper body.
This person is wearing Other in lower body.
This person is wearing Formal.
This person is accessory Hat.

9
This person is Female.
This person has ShortHair.
This person is wearing Logo in upper body.
This person is wearing LongPants in lower body.
This person is wearing Formal.
This person is accessory FaceMask.

2
This person is Male.
This person has ShortHair.
This person is wearing Tshirt in upper body.
This person is wearing Other in lower body.
This person is wearing Other.
This person is accessory Other.

4
This person is Male.
This person has ShortHair.
This person is wearing Tshirt in upper body.
This person is wearing Other in lower body.
This person is wearing Other.
This person is accessory Other.

6
This person is Female.
This person has LongHair.
This person is wearing Tshirt in upper body.
This person is wearing Jeans in lower body.
This person is wearing Casual.
This person is accessory FaceMask.

10
This person is Male.
This person has ShortHair.
This person is wearing Tshirt in upper body.
This person is wearing Other in lower body.
This person is wearing Casual.
This person is accessory Other.

Figure 11: Pedestrian attribute recognition in complex scenes.